Value, Morality & Social Reality

ESSAYS DEDICATED TO
DAN EGONSSON, BJÖRN PETERSSON
& TONI RØNNOW-RASMUSSEN

EDITED BY ANDRÉS G. GARCIA,
MATTIAS GUNNEMYR & JAKOB WERKMÄSTER

LIST OF CONTRIBUTORS

David Alm, Henrik Andersson, Olle Blomberg, Eric Brandstedt, Johan Brännmark, Krister Bykvist, Erik Carlsson, Stephen Darwall, Seyyed Mohsen Eslami, Cathrine V. Felix, Andrés G. Garcia, Mattias Gunnemyr, Frits Gåvertsson, Lena Halldenius, Jens Johansson, Benjamin Kiesewetter, Kirk Ludwig, Christian Munthe, Jonas Olson, Francesco Orsi, Herlinde Pauer-Studer, Ingmar Persson, Erik Persson, Wlodek Rabinowicz, Andrew Reisner, Caj Strandberg, András Szigeti, Matthew Talbert, Fabrice Teroni, Caroline Torpe Touborg, Tobias Hansson-Wahlberg, and Jakob Werkmäster.

PRACTICAL PHILOSOPHY
Department of Philosophy
Joint Faculties of Humanities and Theology
ISBN 978-91-89415-65-2

Value, Morality & Social Reality

Value, Morality & Social Reality

Essays dedicated to Dan Egonsson, Björn Petersson & Toni Rønnow-Rasmussen

Edited by
Andrés G. Garcia,
Mattias Gunnemyr, and
Jakob Werkmäster

LUND
UNIVERSITY

Value, Morality & Social Reality

Essays dedicated to Dan Egonsson, Björn Petersson & Toni Rønnow-Rasmussen

Published by the Department of Philosophy, Lund University.

Edited by: Andrés G. Garcia, Mattias Gunnemyr, and Jakob Werkmäster.

Cover image by Fabian Jones. Cover layout by Gunilla Albertén.

The printing of this book was made possible by Erik & Gurli Hultengrens fund for philosophy, Lund University.

ISBN: 978-91-89415-65-2 (print), 978-91-89415-66-9 (digital)

DOI: 10.37852/oblu.189

Suggested citation: Garcia, A. G., Gunnemyr, M. & Werkmäster, J. (2023) *Value, Morality & Social Reality: Essays dedicated to Dan Egonsson, Björn Petersson & Toni Rønnow-Rasmussen.* Lund: Department of Philosophy, Lund University.

Printed in Sweden by Media-Tryck, Lund University

Media-Tryck is a Nordic Swan Ecolabel certified provider of printed material. Read more about our environmental work at www.mediatryck.lu.se

MADE IN SWEDEN

Contents

Preface

Andrés G. Garcia, Mattias Gunnemyr, and Jakob Werkmäster

Dan Egonsson, Björn Petersson, and Toni Rønnow-Rasmussen have been pillars of the Swedish philosophical community for over three decades and made important contributions to its work in ethics. They have done so not only through their internationally renowned research about value, morality, and social reality but also through their teaching and efforts to guide the next generations of practical philosophers at Lund University. Being the same age, they began their studies at roughly the same time in the 1980s, and they have since then played central roles in shaping the research and teaching environment at the Department of Philosophy at Lund University. We would argue that their most impressive feats in recent years include their patient guidance of us, the editors of this festschrift, as we have gone from being their confused students to becoming their confused colleagues. Using the excuse that two of them are now in various stages of retirement and that all will be celebrating their 67th birthdays the upcoming summer, we have decided to put together this festschrift and arrange a symposium to express our gratitude and pay tribute to their work.

Their work has focused on a variety of themes from practical philosophy that are reflected in the title of the present festschrift. Within this anthology, we have collected thirty-two papers from as many philosophers, dealing with questions that lie in close proximity to those themes. For example, there are papers here on such varied issues as harm, aesthetics, final value, human dignity, social ontology, instrumental value, moral responsibility, and so on.

We owe thanks to all the contributors and reviewers that have helped us put together this festschrift: David Alm, Henrik Andersson, Olle Blomberg, Eric Brandstedt, Johan Brännmark, Krister Bykvist, Erik Carlsson, Stephen Darwall, Anton Emilsson, Seyyed Mohsen Eslami, Cathrine V. Felix, Frits Gåvertsson, Lena Halldenius, Jens Johansson, Benjamin Kiesewetter, Kirk Ludwig, Christian Munthe, Jonas Olson, Francesco Orsi, Herlinde Pauer-Studer, Ingmar Persson, Erik

Persson, Wlodek Rabinowicz, Andrew Reisner, Thomas Schmidt, Caj Strandberg, András Szigeti, Matthew Talbert, Fabrice Teroni, Caroline Torpe Touborg, Tobias Hansson-Wahlberg, and Robert Pál-Wallin.

Many philosophers who hoped to contribute and pay tribute to the work of Egonsson, Petersson, and Rønnow-Rasmussen were unfortunately unable to participate in the present anthology. We would like to encourage them and anyone else that feels indebted to them to express their gratitude personally. While Egonsson, Petersson, and Rønnow-Rasmussen are likely to recommend that people keep their appreciation to themselves, our advice is that people do what we have done here and ignore their aversion toward attention and collegial sentimentality. Egonsson, Petersson, and Rønnow-Rasmussen have made themselves the fitting targets of attention—whether they desire it or not.

VALUE, MORALITY AND SOCIAL REALITY

A SYMPOSIUM DEDICATED TO DAN EGONSSON, BJÖRN PETERSSON & TONI RØNNOW-RASMUSSEN

31 MARCH, 2023 B538, LUX, LUND

Dan Egonsson, Björn Petersson, and Toni Rønnow-Rasmussen have been pillars in the Swedish philosophical community for over three decades and have made important contributions to its work in moral philosophy. They have done so not only through their internationally renowned research about value, morality, and social reality but also through their teaching and efforts to guide the next generations of practical philosophers at Lund University. Using the excuse that two of them are now in various stages of retirement and that all are approaching their 67th birthdays, we have put together this symposium to express our gratitude and pay tribute to their work. But the symposium is just the cherry on the cake. The symposium is also a release party for a festschrift in their honor, with more than thirty contributions from as many scholars.

PROGRAM

10.45	Welcome
11.00	*Rock-bottom reasons* Cathrine V. Felix, Inland Norway University of Applied Science
11.30	*Preference, Information, and the Problem of Big Decisions* Johan Brännmark, Malmö University
12.00	Lunch (at Valvet)
13.30	*Causation, responsibility, and norms:* *Re-evaluating our norms in the face of climate change* Caroline Torpe Touborg, Umeå University
14.00	*Team reasoning, mode, and content* Olle Blomberg, Gothenburg University
14.30	Fika
15.00	*Do We Have Obligations to Collectives?* András Szigeti, Linköping University
15.30	*Goodness and Numbers* Wlodek Rabinowicz, Lund University
16.00	End of workshop
18.00	Dinner (at New Delhi).

If you want to attend, please register for the symposium by emailing mattias.gunnemyr@fil.lu.se no later than March 28th

The Animal Rights Debate Reconsidered

David Alm

Abstract. This paper concerns a certain kind of skepticism about moral rights (in the paper simply called "rights skepticism"), according to which the debate between different views about the nature of rights generally, or at least the debate over animal rights in particular, is misplaced and the participants are talking past one another. While I cannot show that skepticism about the animal rights debate is correct, I offer some reasons for endorsing it, in the form of significant differences between rights of the kind we can attribute to persons and rights we can attribute to animals. I then argue briefly that if skepticism about the animal rights debate is correct, we should not attribute rights to animals (where this is understood as a verbal claim). I then note that rights skepticism could lead us in the direction of rights *nihilism*, the claim that we can simply replace all talk of rights with talk of impersonal moral reasons and requirements, but go on to suggest that my defense of rights skepticism indirectly provides a response to the nihilist challenge.

One of the most contentious issues in the theory of moral rights is that of whether only persons have rights. The root of the difficulty is that the constituent concepts, (moral) *right* and *person*, are themselves contentious. Here I will be concerned only with the former, however, taking the latter notion more or less for granted. Considering rights, we find that (apparent) disputes over their nature appear so intractable that one might wonder whether they are genuine at all. Are the disputants even talking about the same thing? Conveniently, if misleadingly narrowly, let us

call a negative answer to that question 'rights skepticism'.[1] Perhaps it does not strictly speaking entail that the question of non-person rights is also merely verbal, but here I will simply grant that it does.

To illustrate how rights skepticism impacts the debate over non-person rights, let us briefly consider the two best-known views about the nature of rights, the so-called "benefit" and "choice" theories.[2] According to the former, the holder of a right corresponding to some duty is the *beneficiary* of the duty's being performed. Obviously a non-person could benefit from the performance of a duty no less than a person could, so there is no obstacle to a non-person's being a right holder on the benefit view.[3] On the choice theory, by contrast the right holder is the one who exercises *control* over the duty, who can waive it or insist on its performance. Since such exercises clearly require mental capacities usually taken as unique to persons, the choice theory implies that only persons can have rights.

I find myself attracted to rights skepticism, at any rate when it is applied to that particular debate about rights in which the opposing camps seem the furthest apart, which is precisely the one over the rights of non-persons. However, I cannot really *show* that rights skepticism is true, even in a restricted version dealing only with the debate over non-person rights. After all, to make that case, one would have to show that the term 'right' is ambiguous, meaning that it picks out two distinct moral phenomena, rather than just one (about which different philosophers have opposing views) — and I do not know how to do that. I simply do not have a sufficiently worked-out view about how, in general, to distinguish between moral phenomena. Failing that, though, I do wish at least to offer some support for the skeptical position, in the area of non-persons rights, and I will do that precisely by identifying the *differences* between those rights we can intelligibly attribute to non-persons and those that only persons could have. That will be the first, and main task, of this paper. As an ancillary to that task, I will also address the question of how we should use the word 'right', if rights skepticism holds. Finally, I will offer some reflections

[1] For examples of such skepticism, see van Duffel (2012) and Hayward (2013). The convenient label in the text is misleading because the rights skeptic, as understood here, does not (necessarily) deny that rights exist.

[2] The characterizations that follow are intended to be fairly generic and simple. More in-depth treatments can be found in many books on rights theory, such as Kramer, Simmonds & Steiner (1998). Note, however, that such textbook accounts typically concern rights quite generally, whereas I am concerned here solely with moral rights. I should also note that, though I describe the two theories in the text as "views about the nature of rights," I actually formulate them as ways of identifying the holder of a given right (in line with much of the literature). This is because a theory about the nature of rights *is* largely (if certainly not exclusively) a theory about what makes someone a right holder.

[3] To be sure, advocates of the benefit theory would need to say more about how to identify "the" beneficiary of a given duty: clearly it could not be just any creature who benefits, directly or indirectly, from the duty's being fulfilled. For present purposes, though, this complication matters little, for there is no apparent reason why this "specially privileged" beneficiary must be a person.

on the significance of the skeptical position. In particular, I will address the familiar question of why we should be talking about moral rights at all.

Before proceeding, it might be useful briefly to address the scope of the discussion to follow. Specifically, when I speak of "non-persons", which types of entity do I have in mind? There are mainly three kinds of entities to which moral rights are often attributed and that are not persons (at least arguably): animals, (small) children and collectives of persons. In each case, the question of whether entities of the type in question can have rights generates special difficulties. To simplify the discussion, I will accordingly restrict my attention in what follows to only one of these types of entity, namely animals. That way, we can avoid a pair of quite difficult issues: that of whether being a *potential* person matters to one's having rights, and that of whether collectives are agents (if not persons).

Turning to animals specifically, a pair of additional disclaimers are helpful. In the first place I will not be concerned with the empirical issue of what mental capacities animals possess. Indeed, and as I anticipated in the opening paragraph, I will not pause to specify in any detail what I myself take to be necessary for counting as a person. I will proceed, if only for convenience, as if animals are all non-persons (though I will add some relevant remarks later). Second, and relatedly, I am not suggesting that a proponent of animal rights must accept the benefit theory of rights, or more generally accept that any being capable of being "benefited" (or, probably not equivalently, of having "interests") is at least a possible right holder. To illustrate, likely the best-known defender of animal rights (Regan 2004) holds that all beings that are, in his term, "subjects-of-a-life" have a "right to respectful treatment," precisely in virtue of being such subjects (ibid., pp. 276ff). Further, being the subject of a life would seem to go beyond merely having "interests." Indeed, it involves capacities that plainly not all animals possess, such as "having a sense of the future, including [one's] own future," (ibid., p. 243).

We can now state rights skepticism, as applied specifically to the case of animal rights. While it could no doubt be understood in different ways, it will be convenient for present purposes to employ the following formulation: in one sense of the word 'right' animals have rights, and in another sense they do not, and there is no way of eliminating this ambiguity. We will never arrive at a clear, unambiguous answer to the question of whether animal rights exist. To repeat, I take it as implicit in this formulation that those philosophers who assert that animals have rights and those who deny this use the term 'right' to refer to distinct moral phenomena (at least to the extent that the term 'right' refers at all).

I also wish to add a caveat. Though the term 'right' itself is of course quite colloquial and is indeed used in a variety of contexts and senses, its strictly philosophical use is, inevitably, a good deal more technical and theory-laden. Further, the debate over who has moral rights is typically conducted in terms of that philosophical concept (unclear though it admittedly is). It could perhaps be charged, then, that the skeptic's ambiguity is an artifact of a philosophical debate, rather than something that inevitably grows out of our thoughts or language. I will make no

effort to assess this objection here, though I concede that it would eventually have to be confronted. While I would not wish to be taken as implying that current philosophical jargon is irreparably defective — nor that it is beyond reproach, for that matter — I note that we could probably in any case reframe what I say below about differences between person and non-person rights in terms of the arguably less technical notion of "being wronged" — though I will not stop to do so.

With these preliminaries, it is time to move on. The main task I have set myself in this paper, recall, is to outline the differences between (what I take to be) two distinct moral phenomena, both of which are typically called "rights," but only one of which can be correctly attributed to persons. As will soon become apparent, these (purported) differences are concerned largely, if in different ways, with the types of *reasons* we have for relating in various ways to persons and (non-person) animals. As a consequence, I should note, I am susceptible to two kinds of challenge. In the first place, critics could deny that the differences I identify are genuine. In other words, they could maintain that the reasons we have to relate to persons and animals are not different in the ways I claim they are. I cannot here deal fully with this kind of criticism, though I refer interested readers to Alm (2019), in which I characterize the rights of persons in greater detail. In the second place, someone could grant that the differences I mention do exist, but deny that they are sufficient to warrant our talking about "two distinct moral phenomena." This is the challenge I noticed earlier, and said I could not meet, because I do not know how to distinguish between what I have here called "moral phenomena." For what it is worth, though, I will at least assert that in my opinion the differences I identify below are significant enough that I would want to see rather strong arguments for not drawing the conclusion I prefer (that opponents and proponents of animal rights refer to distinct phenomena when they use the word 'right').

The second challenge just noted finds a counterpart also at the verbal level, with respect to the question of how we should use the word 'right'. In that context we may note that rights skepticism can lead us in either of two directions. On the one hand, we could rest content with the ambiguity, holding that there are two perfectly legitimate ways of talking about rights, and that in either usage the term refers to an important moral phenomenon. Those who take this line will no doubt concede that it is wise to mark the distinction linguistically in some way ('X rights' and 'Y rights'), but (they say) that is just for convenience. On the other hand, we could hold that the two phenomena picked out by the term 'right' are so different that using the same term to label them both indifferently would foster confusion rather than clarity. I stress that this second response is compatible with skepticism. It does not imply that it is a *mistake* to claim that animals have rights, and still less that such a claim amounts to an "abuse of language." It concedes to the skeptic that our actual language is not precise enough to allow us to say either of these things.

To be sure, there is no substantive difference between the two standpoints described in the preceding paragraph: they are concerned only with how best to use a certain word. Hence the choice between them is hardly of profound philosophical

significance. Yet words matter, too, and we do need to take a stand. I favor the second standpoint. It does not tell us which terms to use for the two phenomena, only that we should use different ones. However, I am also inclined to believe that we should restrict our application of the term 'right' to persons only, and therefore use some other term in speaking about animals. Indeed, perhaps it is sufficient, in that case, to speak merely of what it is right or wrong to do, in a purely impersonal sense. (We will return to this last point.) This preference is perhaps ultimately merely a matter of taste, but at any rate my preference is that, if a choice needs to be made, 'right' should be applied to the phenomenon that demands the most of the right holder.

One final caveat might be useful. My aim below is not to offer anything like complete descriptions of the two phenomena, but rather only to point at differences between them. What is more, while I offer a positive, if partial, characterization of rights properly so called — those that only a person could have — I provide only a negative characterization of rights *not* properly so called (according to me, anyway) — those that animals can also have. For present purposes, though, this blatant instance of invidious discrimination is not really much of a problem. Indeed, it does not really matter here whether the term 'right' as applied to animals refers to anything at all, whether there *is* any phenomenon there to contrast with rights properly so called. I have spoken of "two distinct phenomena," and will continue to do so, because I am inclined to believe that there *is* a genuine phenomenon picked out by talk of animal rights — perhaps something along the lines suggested by the benefit theory, or even Regan's view — though, again, I prefer not to use the word 'right' to refer to it.

I will now describe five differences of note between rights that can plausibly be attributed only to persons and rights (or "rights") that can plausibly be attributed (also) to non-persons. To repeat, I will not argue for these attributions, but will rather make do with describing the various features attributed and adding some hopefully elucidating remarks.[4] I also do not assert that the list is exhaustive. What matters, as noted earlier, is that they are significant enough to warrant talk of "distinct phenomena."

I start with what is probably the most obvious point of difference, that the rights of persons are associated with certain *powers* that non-persons, including animals, could not have (if only because they require the use of language). Most prominently, these include the powers of *waiver*, and its opposite, *demanding performance*. We have here also the power to demand *compensation* for harm. In some cases there is also the power to transfer a right to someone else. Admittedly, these powers could be exercised on an animal's behalf by a proxy, but it is unclear at best that this fact allows us to conclude that the animals themselves have rights.

Speaking of compensation specifically one could go further and assert that we are under no moral requirement to compensate animals for wrongful harm, without

[4] Again, see Alm (2019) for further discussion.

taking a stand on whether such harms are themselves right infringements. If so, there would seem to be a further difference. However, while I sympathize with this stronger claim I need not make it here: the position I have taken is strictly speaking compatible with there being an impersonal requirement to compensate animals for harm one has caused them, but (obviously) not a duty, corresponding to a right (properly so called).

To prevent misunderstanding, the point of adverting to the above difference between the rights of persons and those of non-persons, put in terms of "powers" and "control," is not to endorse the choice theory of rights, which (as I noted above) is defined precisely in terms of such powers.[5] I am not concerned here with how to identify the bearer of a given moral right, or the nature of such rights more generally. I wish merely to make the point that the ability to exercise certain powers is an essential component of the "distinct moral phenomenon" I am out to elucidate.

The second difference to which I wish to draw attention concerns reasons for *attitudes* (broadly speaking), rather than actions. First and foremost, I have in mind the reason to *resent* actions that wrong one. Again it seems clear that an animal could not have such a reason. I will not assert dogmatically that no animals are even *capable* of resentment (as opposed to mere anger) — though surely many are not. Whether a given animal has that capacity — or, perhaps more accurately, whether its observed behavior on a given occasion amounts to an exercise of that capacity — is a hard question to answer for several reasons, which I could not address here in any case. It is a further question what, if anything, is involved in being able to have *reason* to feel resentment — or any other emotion, for that matter — beyond merely having the capacity for feeling it. I will attempt no answer to that question, either. For present purposes it is enough to observe — or at least assert — that the capacity for having reasons, whether for emotions or for actions, requires mental capacities that few if any animals possess. This claim is in line with my earlier assertion that animals are not persons.[6]

We might also point to the related, if admittedly somewhat obscure, phenomenon of *forgiveness*. I do not have a firm view of the nature of forgiveness. Indeed, perhaps it is best understood as the exercise of a power, in which case it should have been addressed above, under the heading of the first difference. Or perhaps it should simply be understood in terms of emotions or other attitudes. Depending on how it is interpreted, it is perhaps possible for an animal to forgive a wrongful harm — though I would want to see the case made. In general, the "thinner" our notion of forgiveness, the more likely it is that an animal would be capable of it; but, by the same token, the less clear it is that such a capability is relevant to having rights.

---

[5] However, I have at least arguably committed myself to rejecting the benefit theory (at any rate as applied to persons' rights).

[6] In the case of emotions also, we could again appeal to proxies, noting that a person could feel a sort of vicarious resentment on an animal's behalf. As before, though, that fact (even if granted) does not obviously tell us anything about *the animal's* rights.

An additional phenomenon of some relevance here is that of *apologies*. While these are, obviously, actions rather than attitudes, it would seem that apologizing to some being makes sense only if it is capable of forgiveness. I note, though, that even if animals are not capable of forgiveness, a person could feel *guilt* over harming an animal, which might in turn manifest itself in a desire to say that one is sorry, or even to make amends somehow (cf. the remarks above about compensation). Such actions are not crazy, but they also differ in crucial ways from actions described in the same words which we perform vis à vis other persons (who clearly *are* capable of forgiveness).

A third difference is a bit more theoretical — and certainly controversial. I would maintain that duties — or at least duties corresponding to standard negative rights — are constituted by so-called *exclusionary* reasons.[7] Though I could not offer anything like a complete defense of that large claim here, I do need to say something about why it is supposed to be true. An exclusionary reason is a (second-order) reason not to do something for some reason. I take such reasons to be essential to rights (and duties) because we need it to explain the fact that a duty bearer is *bound by* another's demand or command. This phenomenon, I hold, cannot be explained simply by appeal to the idea of weighing (first-order) reasons for and against an action. It requires that the right holder — the one doing the demanding — rules out certain considerations, in favor of not acting as the right holder has demanded he act, to which the agent would otherwise have been free to appeal. What is more, I favor an account of those exclusionary reasons that help constitute duties that in turn refers to features characteristic of persons. That account is closely linked to the idea that in respecting the rights of another agent — which is at the same time also to respect his value — we treat him as "being in charge of" certain aspects of his life, or as being decisive over certain matters, in a way that only a being capable of autonomous choice could be (where the notion of "being in charge" must be understood in terms of exclusionary reasons).

Fourth, moral rights have certain normative consequences, generally accepted among rights theorists, that do not seem to apply to non-agents; nor could my preferred explanations of these consequences be extended to non-agents — though I could not show that here. Most prominently it is questionable whether so-called "deontological constraints," unclear though they admittedly are, that theorists frequently associate with rights, apply also to non-agents. One philosopher who has made this point is F. M. Kamm.[8] As such constraints are typically understood, they forbid killing one person in order to save the lives of several others. But, Kamm asks, is it equally a right violation to kill one dog in order to save several other dogs?

[7] I derive this notion from the work of Joseph Raz. See especially Raz (1986, 1990).

[8] See, e.g., (2007, p. 255).

In Kamm's view, and mine, the answer is no, and so there is a significant question about whether dogs, and other animals, have rights (of the kind that persons have).[9]

A fifth difference is doubtless the most controversial: even some philosophers who share my disinclination to speak of rights for non-persons would bristle at the suggestion that the feature I have in mind here is essential to rights. Yet it has sufficiently wide appeal to merit a mention (and I do accept it myself). I have in mind the idea that most or perhaps all of a person's rights are dependent in various ways on how he *acts*. It is reflected in the familiar idea that rights can be *forfeited* (though I prefer not to use that term myself.) As only agents can act, in the relevant sense, only agents can have forfeitable rights. I concede, however, that it could still be true that non-agents, and indeed agents also, could have non-forfeitable rights, so this fifth difference is perhaps less important than the others. There is a good deal more to say about this matter, but I will leave it aside here.

Having described the various differences holding between the rights of persons and the (purported) rights of animals, it is time to take a step back and reflect on what we have found. A first point concerns the differences just outlined. It should be immediately apparent that some of them are *practically* important, in that they seem arguably to matter to what we ought to *do* ("all things considered"), or at least to our reasons for acting. That is perhaps most obviously true about the fourth point, concerning deontological restrictions, but it is also, at least arguably if also less obviously, true of several of the others. Some other differences matter rather to how we ought to, or at least have reason to, *feel*. It should be plain, then, that they are not merely relevant to the merely verbal question of how to use the word 'right'. Nor would it be fair to say that though they matter to how we should *classify* our moral requirements, or our moral reasons for action, they do not matter to *which* requirements we have.

This last point is important because it also tells us something about the significance of rights skepticism. To see how, it is useful to contrast that doctrine with another type of skepticism in the moral domain. A notoriously intractable (apparent) disagreement in moral philosophy holds between *objectivists* and *subjectivists* about the moral 'ought'. According to the subjectivist, what a person (morally) ought to do in a given situation depends on what information he has (or, in some sense in need of definition, *should* have) in the situation. According to the objectivist, by contrast, what one ought to do is determined by the facts of the

[9] One might consider also cases of so-called "aggregation." Some philosophers maintain that we are not required to save the larger number when persons are involved, on the grounds (as Anscombe [1967] puts it) that no one is wronged if we save the smaller number instead. This view is admittedly quite controversial, and I am not inclined to defend it, but we may at least note that it might seem to hold no appeal at all when animals are involved instead of persons. In such cases, that is, perhaps even so-called "numbers skeptics" would maintain that we ought to save the larger number (insofar as we are under any requirement to act at all). At least one leading numbers skeptic (Taurek [1977], p. 306) takes this view when it comes to valuable objects, such as works of art, but he does not specifically address the case of animals.

situation, whether or not one knows (or even could know) about them.[10] A skeptic about this debate holds that it cannot be resolved and that the disputants are talking past one another.

For concreteness, imagine the following situation. Bob's leg is infected and will have to be amputated unless he is quickly given a powerful antibiotic. Betty has a bottle labeled "Powerful Antibiotic". However, she is unaware, and could not be expected to know, that the bottle has been tampered with and in fact contains a deadly poison. Ought she to administer the contents of the bottle to Bob? The objectivist says "no", while the subjectivist says "yes." The skeptic's diagnosis of the case is that in one sense of 'ought', the objectivist is right and in another sense of 'ought', the subjectivist is right — and that is all there is to say about the matter. That line is maddeningly unsatisfactory (which, unfortunately, does not show it mistaken). Whatever our views about the (apparent) dispute between objectivists and subjectivists, we are likely to respond in frustration with something like: "Well, but what *ought* Betty to do then (really)?" By the same token, we are unlikely to rest content with the suggestion that there is *nothing* she ought to do "really" (because the notion of what she "really" ought to do is vacuous or nonsensical).

Now again consider rights skepticism. Speaking of some particular animal, let us say, the rights skeptic proclaims that in one sense of 'right', the animal has rights, and in another sense of the word it does not, and that is all there is to say about the matter. The striking contrast is that *this* type of response, in the case of animal rights, is not as maddening as the parallel response in the case of Bob and Betty just described. What explains the difference? It is tempting to say here that a disagreement about what a person ("really") ought to do is "immediately practical" in a sense that a disagreement about whether some person, or animal, has a right is not. It is admittedly not easy to identify exactly what this notion of being "immediately practical" amounts to, and this is in any case not the place for an inquiry into that matter. For present purposes it might be enough to note instead that the claim that x has a right to some treatment could intelligibly be met with the question "OK, but what ought I to do (as far as x is concerned)?", whereas one obviously could not intelligibly respond in that way to the claim that one ("really") ought to φ. We can set aside the question of whether there is *anything* further to say once you have come to accept that you ("really") ought to φ, of whether there is anything for it but simply to φ. (In doing so, we are presumably also setting aside the question of whether the claim in question is "immediately practical.")

Explaining the contrast between the two cases in that way, however, raises a different issue that is harder to set aside. After all, if we express the difference between the two moral phenomena in the ways I have suggested, or something similar, why not simply make do with talking about what we *ought* ("really") to do (and perhaps feel)? Why bother with rights at all? This question raises the specter of what we could call *rights nihilism*, the view that there are no rights at all, or that

[10] The distinction goes back at least to Moore (1912).

talk of "moral rights" fails to refer. This position is plainly more radical than rights skepticism, and every rights theorist would sooner or later have to address it. It is worth asking, then, whether the above discussion of animal rights can help us respond to rights nihilism.

The obvious difficulty in answering that question is that it is not clear what would constitute a satisfactory response in the first place. The nihilist wishes to reduce, and indeed ultimately reduce *away*, rights to non-relational moral phenomena, concerned simply with the agent's reasons for action and for taking various attitudes. For such a reduction to be successful, the notion of a reason must not itself be relational. In other words, a reason for not harming another person (or whatever is at stake) must not be a reason "to" that other person. It is, however, not easy to say what it would mean for a reason to be relational, and hence whether a given attempt at reduction of the relational to the non-relational is successful.[11]

Though I will not here address the difficult question of "relational reasons," and am indeed not even able to provide necessary and sufficient conditions for a right's being "reduction proof," as it were, I do believe that I have identified above at least one feature of rights ("properly so called") that cannot plausibly be reduced to non-relational reason facts. I have in mind the significance of the right holder's control over the duty bearer's duty — the type of control I have said only a person could have.[12] The key point here, I hold, is that the right holder's demand or insistence on performance is not itself a (first-order) reason for action, but rather provides an exclusionary reason, as I have described. Hence, the relevant relational fact — that the right holder demands compliance of the duty bearer (or, alternatively, fails to waive his right) — is not the ground of a reason for action (in the way that, say, the relational fact that an action would harm another could serve as such a ground), but rather serves to make some other fact (itself most likely relational) decisive for what the duty bearer ought to do.

I concede that a demand could at least easily be made to *look* like a reason for action (if perhaps a peculiar one). After all, it would make perfect sense for an agent to *explain* or *motivate* his action by pointing precisely to the demand. (On being asked "Why didn't you take Bob's bicycle when you had the chance?" the response "Because he told me not to" would be no less intelligible than, say, "Because it would have broken his heart.") The demand, then, looks like something that could *rationalize* an action, and that is of course just what reasons are supposed to do. Here I cannot give this challenge the attention it probably deserves. That said, in response I would stress that the demand is not a consideration counting against the action to a certain extent, or with a certain weight, to be weighed against (or with) other considerations. Which would that weight be? Rather, either the right holder's say-so is decisive or it is not (he is "in charge" or he is not). If he is, then his demand

[11] Cf. Thompson (2004).

[12] I address this issue at greater length in Alm (2019, pp. 91ff).

settles the matter; if he is not, it makes no difference at all (though it might still be wrong not to do as he demands).

A final point, worth making fully explicit, is that the type of argument just made — whatever its merits — is not applicable to non-person rights (even if we grant that such exist). It presupposes that the holder of the purported right has the ability to control the duty bearer's duty in a way that only a person could. Does that mean that adherents of animal rights are unable to explain why it makes sense to attribute rights to animals, or indeed to speak of "rights" at all? No, it does not — though I have recommended against talking in that way. For all I know, however, it *does* mean that they would have to concede that such talk is simply reducible to talk of impersonal, non-relational requirements.[13]

References

Alm, David (2019) *Moral Rights and Their Grounds*. New York: Routledge.

Anscombe, G. E. M. (1967) "Who Is Wronged?" *Oxford Review* 5: 16–17.

Hayward, Tim (2013) "On Prepositional Duties" *Ethics* 123: 264–91.

Kamm, F. M. (2007) *Intricate Ethics*. Oxford: Oxford University Press.

Kramer, M, N. Simmonds & H. Steiner (1998) *A Debate over Rights*. Oxford: Oxford University Press.

Moore, G. E. (1912) *Ethics*. Oxford: Oxford University Press.

Raz, Joseph (1986) *The Morality of Freedom*. Oxford: Oxford University Press.

Raz, Joseph (1990) *Practical Reason and Norms*. Second Edition. Oxford: Oxford University Press.

Regan, Tom (2004) *The Case for Animal Rights*. Berkeley: University of California Press.

Taurek, John (1977) "Should the Numbers Count?" *Philosophy & Public Affairs* 6: 293–316.

Thompson, Michael (2004) "What Is It to Wrong Someone? A Puzzle about Justice" in R. J. Wallace, P. Pettit, S. Scheffler and M. Smith (Eds.), *Reasons and Value: Themes from the Moral Philosophy of Joseph Raz*. Oxford: Oxford University Press.

Van Duffel, Siegfried (2012) "The Nature of Rights Debate Rests on a Mistake" *Pacific Philosophical Quarterly* 93: 104–23.

[13] This paper is a significantly expanded, and I believe also improved, version of a section of Alm (2019, pp. 84-87).

Jumping the Hurdles of Moral Progress

Henrik Andersson

Abstract. In their work on moral progress, Dan Egonsson and Toni Rønnow-Rasmussen both express a worry concerning possible problems caused by value incommensurability. I show that this worry can be alleviated. This is done by explicating the structure of the concept of moral progress and the structure of value comparatives. With a better understanding of the relevant concepts, it becomes clear that value incommensurability does not pose a dead-end for work on moral progress, but is rather a natural part of the normative domain, and consequently, it is merely one more hurdle to pass in our pursuit of moral progress.

My life as a philosopher can be defined by the valuable impact of Dan Egonsson, Björn Petersson, and Toni Rønnow-Rasmussen. My first encounter with practical philosophy was when I attended a lecture by Dan. His animated lecture not only revealed his passion for philosophy, but also sparked my passion for philosophy and the passion of the other 60 students in the lecture hall.[1] I met Toni later on in my studies when I was to write an essay on John Harris' Survival Lottery. It was my first longer essay in philosophy and Toni's encouragement combined with his critical questions made me realize that I too could write philosophy and be part of the world that I, until then, only had watched from the outside. When Björn a few months later introduced me to Derek Parfit's *Reasons and Persons*, it was obvious that there was no way back. The discussions we had in the classroom on transitivity, rough comparability, and the Lexical View laid the ground for all my subsequent

[1] Later on in my career I had the burden of teaching the course that followed Dan's course. Needless to say, it was an impossible act to follow.

work in philosophy. It is clear that the three together with Wlodek Rabinowicz are fully responsible for my life in academia. Their influence continued in my doctoral studies. Toni was one of my supervisors and the value of his constant encouragement cannot be expressed in words, Björn's work as head of the department brought a sense of stability to the otherwise very unstable life of a doctoral student, and Dan's approach to life and philosophy was a valuable reminder of how to be grounded. A good example of their care for their students is how they developed a research program on moral progress. Their motivation to put in so much work in developing this was, as far as I can tell, mainly to help their doctoral students secure a job after their doctoral studies. With my background in value theory with a focus on value incommensurability, my role in the program was to write on the possible hurdles that potential incommensurabilites may cause in discussions on moral progress.

Unfortunately, the research program did not get funding, but some interesting work on the topic got published. Dan, for example, wrote an entry on "Moral Progress" for the International Encyclopedia of Ethics and Toni wrote an influential paper with the title "On Locating Value in Making Moral Progress". Both contain interesting discussions on the role of incommensurability. In this short paper, I will, as promised in the research application, provide my own thoughts on the role of incommensurability in moral progress.

Hurdles for Moral Progress

The concept of moral progress may at a first glance seem straightforward and easy to grasp; moral progress is the change to a morally better state, such as the abolition of slavery, the advancement of women's rights, or an individual's development as a moral agent. However, according to Egonsson (2018), it is unlikely that we will reach a consensus on how to define the concept. The concept is closely entangled with our metaethical views and our normative and axiological stances; this makes the likelihood of reaching a consensus low. Indeed, the fact that we cannot agree on what is the correct moral theory will sometimes make it difficult to agree on whether there has been moral progress. This may, however, not be the only problem for judgements about moral progress. Egonsson (2018: 9) goes on to state that if you believe that moral values can be incomparable and incommensurable, then "you will most likely dismiss the possibility of a comprehensive judgment about progress in morality".

I take the worry to be, roughly, that a definition of moral progress must be spelled out in terms of an improvement, a later state of affairs being better than a previous state of affairs.[2] If two states of affairs are incomparable or incommensurable, then

[2] This is a common view in the literature on the topic. See e.g., Macklin (1977).

they may not be related by an evaluative comparative relation such as "better than" and this would constitute a problem for work on moral progress.[3] Egonsson is not alone in finding incommensurability to be a threat to moral progress. It is, however, somewhat unclear how this threat should be understood.

Fortunately, Musschenga and Meyen (2017) gives some possible suggestions as to how we can interpret the potential threat of incommensurability. I will present them all since I believe that they are good characterizations of what, at first glance, may be believed to be the threats of incommensurability for moral progress. I will provide a critical discussion on each suggestion and, after clarifying the structure of the concept of moral progress, I end up with my own interpretation of the potential threat of incommensurability for judgements about moral progress.

Lack of a Positive Value Relation

One possibility Musschenga and Meyen discuss is that the state of affairs, before a change, is incomparable to a state of affairs after the change. With "incomparable" they mean that there holds no basic positive value comparative between the two states of affairs. That is, neither state is better than, worse than, equally as good as, or on a par with the other state.[4] More specifically, there holds no such relation between the two states of affairs with regard to the relevant dimension of the comparison. In the example they discuss, the relevant dimension of the comparison is "freedom". The thought is that a change can bring about both improvements and losses in terms of freedom but when we are to determine the overall change in freedom these improvements and losses will make it impossible to arrive at an evaluative judgment on the all things considered change in freedom. Consequently, it will be impossible to determine whether there has been moral progress with respect to freedom.

The example given by Musschenga and Meyen is, however, not convincing. Modernization can indeed bring about improvements and losses in terms of freedom but this does not establish that the two states of affairs will be incomparable with respect to freedom. Consider a person's development as a badminton player. First, the person starts out as a poor player, but then slowly

[3] The terms "incommensurable" and "incomparable" are in the philosophical literature used in many different ways. What most uses of the terms have in common is that they refer to some problems of comparison such as the lack of a common scale on which two alternatives can be placed or even the impossibility of attributing a comparative value relation to two alternatives. In this paper, I will not commit to any definition of the terms in question, but rather consider many different problems that may arise when we are to compare states of affairs. For more on possible definitions of "incomparable" and "incommensurable" see Andersson & Herlitz (2022) and Hsieh & Andersson (2021).

[4] See Chang (2002) for an argument for the possibility of parity. Strictly speaking this definition of incomparability leaves room for more than these four possibilities, but it is often assumed that these four value comparatives fully exhaust the space for basic positive value comparatives. See Rabinowicz (2008) for a discussion of the possibility of more than these four comparatives.

progresses and improves as a player. However, at some stage, the person has trained too much, gets tendonitis and thus becomes a worse badminton player. When the tendonitis has healed, her skills improve again. Some gains and some losses in the person's skills will take place over some time. These gains and losses will, however, at each moment, add up to a certain level of skill and there will be no problem to determine whether the person is a better or worse badminton player as compared to the first time the person played badminton. From this analogy, we learn that a series of improvements and deteriorations will not lead to a state that is incomparable to the first state.

There is, however, a closely related possibility that is, at least prima facie, more worrying. It could be that "freedom" is a complex and multidimensional concept and a change may bring about an improvement in one relevant dimension of freedom, but may also constitute a loss in another relevant dimension of freedom. Furthermore, it is possible that this concept does not provide us with a function to determine the overall amount of freedom.[5] The new state, that has come about due to the change, may thus not be better than the former state; there is no way for us to weigh these different dimensions and reach an all considered judgement about the freedom of the new state as compared to the former state.

Nonsubstitutability

The second interpretation provided is that a change may bring about an improvement in one value at the cost of a loss in another value. Furthermore, it is possible that one value does not substitute the other; they are *nonsubstitutable*. Musschenga and Meyen exemplifies:

> The value of human life, for instance, is thought to have a special status so that its loss cannot be compensated for by economic gains. It is often said that modern society has brought us more individual freedom, [but] also less community and solidary. If the loss of community and solidarity cannot be compensated by the gain in freedom, these values are incommensurable. (2017: 10)

On the nonsubstitutability interpretation, the loss of one value cannot be compensated by the gains of another value. While this may be true, I believe that the more pressing and fundamental issue is how nonsubstitutability can come about. I believe that the answer can be found in the previous interpretation. On that

[5] Or more specifically: does not provide us with a positive account of the relation. The function may tell us that the two states are incomparable. Imagine the view that a state is better than another if and only if it is better in all dimensions, two states are equally as good if and only if they are equally as good in all dimensions, and if none of these things are the case the states are incomparable. Such a function would help us determine the relation between all possible states of affairs, but it does not always give us a positive account since "incomparable" is understood as the lack of a positive account.

interpretation, there was no function that would help us determine the overall amount of freedom in a state of affairs. If there is no function that can determine how e.g., individual freedom, community and solidarity all contribute to the overall freedom it will result in nonsubstitutability.[6] In other words, this second interpretation seems to collapse into something similar to the interpretation I presented above.

No Common Scale

The third, and final possibility Musschenga and Meyen mention, is that there may be no common unit of measurement. More specifically: "[t]wo values, such as pleasure and fairness, are incommensurable if there is no cardinal scale of value according to which both can be measured" (2017:10). I take this to mean that if our evaluation of moral progress involves values that cannot be placed on the same cardinal scale, it will be difficult to arrive at an all things considered judgement concerning possible moral progress from one state of affairs to another.

This worry seems, however, exaggerated; it may be that the values must be placed on a cardinal scale in order to arrive at a judgment on how *much* moral progress there has been, but in order for us to arrive at a judgment on whether there has been moral progress at all, an ordinal scale suffices. That is, we need only to know that one state is better than another, we need not to know how much better it is.

The worry that there is no ordinal scale at which both states can be placed is similar to the worry I have expressed above. I argued that there might not be a complete function to determine how different values contribute to the overall value of e.g., freedom and for that reason we cannot determine how the two states relate with respect to freedom. That is, they cannot be placed on an ordinal scale that ranks the states with respect to freedom. However, the suggestion that there is no common scale may be even more radical since it could be understood to claim that there is no ordinal scale *whatsoever* on which two states can be placed. This radical view should not be attributed to Musschenga and Meyen, but it is nevertheless an interesting possibility that we will have reason to return to.

[6] Note that on some interpretation of "substitutability" nonsubstitutability need not necessarily rule out the possibility of a function that can determine the overall value in a state of affairs. Theoretically speaking, it could be the case that e.g., community and solidarity cannot be compensated by the gain of individual freedom, but nevertheless, freedom is more valuable than community and solidarity, and thus a state of affairs with more individual freedom at the cost of community and solidarity would count as a morally progressed state of affairs. To give another example, it may be clear to me that a day with my family is more valuable than a day in solitude, but the latter does not fully substitute the former, certain valuable moments of a time in solitude will not be compensated by the values gained of spending a day with my family. This is, however, not the common interpretation in the literature on nonsubstitutability and incommensurability.

The Structure of Moral Progress

With these worries more fully fleshed out, it is possible to consider whether they cause a serious problem for research on moral progress. To address this issue we first of all need to have a better understanding of the structure of the concept. Progress is a multifaceted notion and the term is compatible with many adjectives. We have clarified what area of progress we are interested in by adding the qualifier "moral". We have thus narrowed down the meaning somewhat, but the adjective "moral" is anything but specific; the fact that philosophers have struggled with the topic and produced ideas relating to the topic since at least 500 BCE shows its complexity and width. Fortunately, Egonsson (2018) gives us some guidance in understanding the formal features of moral progress. He suggests that:

> A statement of moral progress is about a change towards a morally better state and typically has the following structure:
>
> x has made (moral) progress regarding y in relation to z,
>
> Where x is the subject or maker of (moral) progress, y the matter of (moral) progress or what the progress consists in and z the dimension of comparison, that is, the relata between which the comparison is made. This dimension will concern different points in time but may also, according to some philosophers, be about different makers of progress.

This seems to be a promising start for understanding the concept. Interestingly, this specification of the concept mirrors the discussions in value theory, where scholars aim to explicate the structure of value comparisons. Indeed, most value theorists seem to accept that a statement such as "*x* is better than *y*" is underspecified. It is believed that all value comparisons must proceed in some certain respect, call this the *Requirement for Specification.* This idea has been popularized by Ruth Chang.[7] She writes:

> I will be assuming that all evaluative comparisons must proceed in some or other evaluative respect(s), what I call a 'covering consideration.' So, for example, Mozart cannot be better than Michelangelo *simpliciter* but only be better in some or other respect(s). … Without a covering consideration in terms of which a comparison proceeds, a comparison is incomplete; saying that Mozart is better than Michelangelo *simpliciter* does not tell us whether he is better with respect to chess, spelling, or creativity. Put another way, all (binary) value relations are strictly three-place: X is better than Y with respect to V. Since explicit reference to a covering consideration in every instance is cumbersome, we omit such reference, but an appropriate covering consideration is always implied. (Chang 2002: 666)

[7] Others have previously discussed ideas similar to this. For example, Peter Geach (1956) and Judith Jarvis Thomson (1997). How their ideas relate to the Requirement for Specification is, however, not fully clear. See Andersson (2016) for more on this.

After her seminal work on value relations most seemed to agree with her that there is a requirement to specify what we are to compare. However, as the requirement is characterized, it is rather unclear what sort of requirement it is. It is uncertain what strength it has and what meta-ethical commitments are entailed by it. This topic is not well explored; mostly it is just taken for granted that the comparisons must be specified. This is not the place to provide an in-depth examination of the structure of value relations but I will now briefly say what I take to be the nature of the requirement.

It seems reasonable that basic value comparatives have at least three relata. For that reason, there is a pragmatic requirement to specify what the relata are. Without the specification, we would not understand each other. Consider for example the following claim by Judith Jarvis Thomson:

> [A]ll goodness is goodness in a way, and […] if we do not know in what way a man means that a thing is good when he says of it 'That's good', then we simply do not know what he is saying of it. Perhaps he means that it is good to eat, or that it is good for use in making cheesecakes, or that it is good for Alfred. If he tells us, 'No, no, I meant that it is just plain a good thing,' then we can at best suppose he is a philosopher making a joke. The same is true of betterness: it, too, is always betterness in a way. (Thomson 1997: 276.)[8]

In short: it will be difficult for us to fully understand what comparison you have in mind if you do not specify the covering consideration.[9] The requirement can consequently be seen as an instantiation of H. P. Grice's cooperative principle (Grice 1989). That is, we need to be as informative as is required in the given context. Or as Grice argues you should "make your conversational contribution such as it is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged" (1989:26).

This is also true when we are to address the worries of incommensurability and moral progress. It is only by understanding the structure of value comparisons that we can understand how we can solve the possible tension caused by incommensurability. With the structure of moral progress and value comparatives clarified we can now return to the challenge posed by incommensurability.

[8] As expressed in the previous footnote it is, however, not clear how Thomson's claim relates to the Requirement for Specification. Thomson, for example, states that things can be "better for Alfred", but the relata "for someone" does not seem to be what people have in mind when they use the term "covering consideration". See Andersson (2016: fn 25) for more on this.

[9] Furthermore, some hold that it is a logical feature of "at least as good" that it is transitive. However, it is fully plausible that A is at least as good as B with respect to V, B is at least as good as C with respect to V and yet C is at least as good as A with respect to V'. If we did not specify the covering consideration the previous example would seem to show that the transitivity of "at least as good as" was violated.

Jumping the Hurdles

While Musschenga and Meyen present three interpretations of why incommensurability can be a hurdle for moral progress, I believe that these three interpretations collapse into two different possibilities. The first possibility is that when we are to consider the moral progress in terms of some specific covering consideration such as "freedom", the changes involve certain losses in freedom, but also certain gains. Furthermore, the concept "freedom" does not provide us with a function that determinately specifies how these gains and losses should be aggregated or weighed in order to reach a conclusion on whether the change has amounted to an all things considered improvement in freedom or not. The covering consideration is underspecified in the sense that it does not provide a complete ranking of the possible state of affairs. That is, when asked whether there has been moral progress from the state of affairs A to the state of affairs B with respect to freedom, the change from A to B may involve many different dimensions of freedom, such as freedom of speech and freedom of religion, and consequently, it could be impossible to determine whether there has been an improvement.

The other and more radical possibility is that there is no aspect whatsoever to compare A to B with, i.e., there might not always be a specification such as "freedom" available. Or more generally, there is no ordinal scale at which both A and B can be placed. This could pose a problem for discussions on moral progress. The framework presented previously, however, gives us the tools to address these worries.

No Complete Ranking

Concerning the first worry, we can always follow Grice's suggestion: Since the context does not provide us with the relevant covering consideration, we must specify further in order to find a concept that gives us a function that determines whether there has been moral progress. There is always the possibility to consider a more precise concept in order to reach a judgement on whether there has been moral progress with respect to this more precise concept. That is, if we cannot determinately tell whether there has been moral progress with respect to freedom, we can specify the concept further and perhaps determine whether there has been moral progress with respect to e.g., personal freedom. To this, it could be objected that we are not interested in whether there has been moral progress with respect to personal freedom, but in whether there has been moral progress with respect to freedom simpliciter.[10] This wish is understandable, but incommensurability does not

[10] An insightful comment from Jakob Werkmäster prompts me to remind the reader that, following Egonsson's definition, moral progress does not consist in a random change from one state to another, but that there must be an intentional maker of this change for it to count as moral progress.

pose a threat in these cases. If it is the case that A and B are incommensurable with respect to freedom, i.e., if it is not the case that A or B is determinately better than the other with respect to freedom, nor the case that they are determinately equal in this respect, then there clearly has not been moral progress from A to B. Or at least it is indeterminate whether there has been moral progress. As Egonsson (2018) has pointed out, moral progress involves a change to something better. When there is incommensurability, we cannot determinately judge that a betterness relation obtains between two states, and consequently we cannot determinately say there has been moral progress.

In this context, incommensurability poses no problem. Compare this to discussions on incommensurability in axiology. In those discussions, the possible threat incommensurability poses to rational choice is often highlighted. The threat is clear: if rational choice is grounded on positive value comparatives such as "at least as good as", then how can rational choice be possible in the face of incommensurability?[11] The same problem does not occur in discussions on moral progress. The fact that there is incommensurability could even act as a motivator; if we have not clearly morally progressed, then this should motivate us to reach a state of affairs that is determinately better than a previous state of affairs. That is, incommensurability can encourage us to progress as individuals and as a collective.[12]

No Common Scale

More, however, needs to be said about the second worry. First, it is surprising that no covering consideration whatsoever is applicable. It should be possible to compare two states of affairs in some evaluative respect. That there is *no* common scale that allows a comparison seems to be a radical claim. This possibility seems especially unlikely in discussions on moral progress. Consider, for instance, the moral progress of a society. How can it not be possible to compare state A with state B with respect to e.g., the emancipation of women, labor rights, or non-discrimination against children? It seems likely that we can always make some comparison of states A and B, and thus it seems radical to argue that no covering consideration whatsoever is applicable.

This possibility is, however, considered in detail by Toni Rønnow-Rasmussen when he discusses the possibility that there might not be a "basic value perspective" that can be used when choosing between two alternatives:

> For instance, consider the following scenario: If you spend all of your salary on an expensive hobby of yours, it will be good for you. Call this option A. But you might

[11] See Chang (2016) for more on this view.

[12] At least incomparability between a previous state and the present state does not hinder further moral progress. See Kitcher (2011: 242-245) for more on this.

> instead send some of your money to charity; this would be a good thing to do, although not necessarily good for you. Call this B. Now, suppose you are not sure what to do, and you cannot do both. So you would like to be able to not merely haphazardly have to decide between A and B; you would like to settle this issue in an unbiased way by comparing them to each other. By "unbiased", I mean that you don't want to settle in advance which of the two options (A or B) should be ruled out. (2016:143)

With this, Rønnow-Rasmussen does not argue for what I claim to be the radical claim that no covering consideration whatsoever is applicable. His claim is more balanced as he believes we in fact can compare A with B:

> [W]e can for instance say that B is better than A with respect to your making moral progress. But, then again, A is better than B with respect to your enjoyment. The first comparison seems to be one with respect to something that is good, period, namely making moral progress, while the latter is done in terms of giving you pleasure, i.e., something that is good-for you. But if you want to compare these options in an unbiased way precisely because you cannot make up your mind about whether to realize what is good or what is good-for you, it is not clear how you should proceed. (2016:143)

The problem is not that there is no covering consideration available, but that it is not clear to you which covering consideration you should apply. He continues:

> Whatever covering value you introduce here will exemplify either one of three possible values: either it will be a good, period or something that is good-for you or it will be a complex consisting of something that is both good, period and good for you. As to the two former possibilities, they seem not to be consistent with an unbiased comparison. So although A and B certainly in one sense are comparable, given the premises I am discussing here, it is not quite clear that A and B are in fact comparable in all cases; the comparison will in some cases be biased and so in those cases, it cannot be represented in an unbiased way. And if that is the case, we are not in the position to make the kind of comparison that we set out to do. This suggests that, just as we might want to say that x and y are incommensurable in the sense that we cannot represent these values by a cardinal unit of measure, there is another sense of incommensurable in which we cannot represent these values by an unbiased covering value (i.e., a value that does not bias the comparison)."(2016:144).

Rønnow-Rasmussen's example involves certain value concepts that are normatively irreducible in the sense that they cannot be understood in terms of each other. Furthermore, there is no more fundamental normative concept that both are reducible to. This means that these concepts cannot even be placed on the same ordinal scale.[13]

[13] This is a much-discussed idea. While Rønnow-Rasmussen places the discussion in the realm of moral progress, similar examples have been presented by others, most notable perhaps by Henry Sidgwick

Rønnow-Rasmussen's worry can, however, be met by the reasoning developed above. That is, we must determine whether we are interested in moral progress, period, or moral progress for a specific person, only with such a specification can we arrive at an answer on whether there has been moral progress. However, Rønnow-Rasmussen might want to dig his heels in and argue that he wants to know whether there has been moral progress in an unbiased way. Producing a satisfactory answer to this will be more challenging.[14]

Remember, we want to know how two states of affairs relate to each other and it is possible that one state of affairs is better with respect to "good for" and the other with respect to "good, period". Let us also assume that we do not wish to compare these states of affairs in terms of "good for" or "good, period". If these are irreducible to a common denominator there will be no answer as to how they relate. At least they are not reducible to a common denominator that is of relevance for the purpose of determining whether there has been moral progress.[15] Where does this leave us? Above I argued that if there had been moral progress, we would be able to tell if a betterness relation holds between the two states under consideration, and if no such relation holds, then there has not been any moral progress. This line of argument does, however, not seem satisfactory as a response to Rønnow-Rasmussen. His worry is that there is no way to determine whether there has been moral progress in an unbiased understanding of "better than" and "moral progress". There is no unbiased value comparative to plug into our definition of "moral progress" and consequently, we cannot provide an unbiased definition of moral

in his discussion about the Dualism of Practical Reason. Roughly, Sidgwick (1874:507- 509) argued that what is morally right can sometimes conflict with what is prudential right. Our duty and our self-interest are derived from different basic principles and are thus normatively irreducible. This gives rise to a dualism of practical reason that can in some occasions give rise to conflicting requirements. Another good exposition of this can be found in Nagel (2012:133) "This great division between personal and impersonal, or between agent-centered and outcome-centered, or subjective and objective reasons, is so basic that it renders implausible any reductive unification of ethics –let alone of practical reasoning in general. The formal differences among these types of reasons correspond to deep differences in their sources. We appreciate the force of impersonal reasons when we detach from our personal situation and our special relations to others […] The two motives come from two different points of view, both important, but fundamentally irreducible to a common basis."

[14] Interestingly, the challenge is similar to the challenge posed by some intertemporal comparisons. An example of an intertemporal comparison could be if we are to compare an object that yet does not exist with something that exists now, or to compare the existing object with an object that used to exist but is now destroyed. What could make comparisons such as this problematic is that it is not clear what perspective the agent ought to adopt. The agent exists here and now. Does that mean that the agent is to decide, given the current circumstances, the evaluative norms of present society, the epistemic situation of the time, and so on, on what to be the best? Taking such a perspective is to be biased towards the present. Sometimes this might be the correct approach but in other circumstances the comparison might be such that she ought to be impartial. Rønnow-Rasmussen's worry is not an intertemporal problem *per se* but it shows a structural similarity: what perspective ought we to take?

[15] This would in fact constitute an example of "non-comparability" see Andersson "The possibility of Incomparability" for more on these examples.

progress. To respond to this worry by stating that there has been no moral progress in the unbiased understanding of the term, is to misconstrue the worry.

With this it is clear that none of the above suggested strategies for dealing with incommensurability in the domain of moral progress will amount to a satisfactory answer to Rønnow-Rasmussen's worry. The question of relevance now is whether his worry actually constitutes a problem for philosophical work on the concept of moral progress. The answer is that it does not. What it at most shows, is that moral progress mirrors structures found on more fundamental levels. If "good, period" and "good for" lack a common denominator then, naturally, moral progress with respect to good, period and moral progress with respect to good for, also lack a common denominator; that is, they are two irreducible normative concepts. All discussions of moral progress must thus be either in terms of "good for" or in terms of "good, period". And if one holds that there is only one true basic normative concept, "good, period" or "good for" and it is unclear how one is reduced to the other, then this is not a problem for discussions on moral progress per se, but a much more troubling problem for normative theory in general.

It should also be noted that the idea that there is a single criterion of moral progress, as utilitarians may argue, can be questioned. Adherents of other normative theories should not find the view that several criteria are of importance when evaluating moral progress to be surprising.[16] For example, those who agree with Martha Nussbaum (2011) that there are ten central capabilities that ought to be reached for a person to live a good life, would find this to be a natural claim. These capabilities are understood to be irreducible and thus it is implicitly assumed that there is no overall concept of moral progress, but moral progress can only be measured with reference to one of these capabilities.[17]

Conclusion

In this paper, I have presented alleged problems incommensurability causes for moral progress. By providing an explication of the concepts involved it was shown why these prima facie problems are no real challenges for a discourse on moral progress. In some cases, the instantiation of value incommensurability shows that there has been no moral progress and in other cases, a specification of what dimension of moral progress we have in mind must be given. Furthermore, this discussion shows that the dimensions of moral progress mirror the distinctions that

[16] See Sauer et al (2021) for a good overview of the merits of a multicriteria approach to moral progress and Kitcher (2021) for a recent discussion on "the many modes of moral progress".

[17] Of course, this does not rule out the possibility that these sometime may overlap. There could sometimes be a "happy coincidence" such that according to all central values or capabilities there has been moral progress. I am grateful to Anders Herlitz for pointing out this possibility to me.

can be made in value theory between different kinds of value such as "good, period" and "good for". Features that belong to the evaluative landscape are not to be treated as dead ends when developing a coherent understanding of moral progress, rather they are natural hurdles that need to be passed. Just as Björn, Dan and Toni showed me how to tackle the hurdles of academia and progress as a philosopher, I have now shown how we can deal with conceptual hurdles of moral progress.[18]

References

Andersson, Henrik, (MS), "The Possibility of Incomparability".

Andersson, Henrik, (2016), "Vagueness and Goodness Simpliciter", *Ratio* 29(4): 378-394.

Andersson, Henrik & Herlitz, Anders (2022), "Introduction", in Andersson, H. & Herlitz, A. (eds) *Value Incommensurability: Ethics, Risk and Decision-making*, Routledge.

Chang, Ruth, (2002), "The Possibility of Parity", *Ethics*, 112(4): 659-688.

Egonsson, Dan, (2013), "Moral progress", in *International Encyclopedia of Ethics*.

Geach, Peter, (1956), "Good and Evil", *Analysis* 17(2): 33-42.

Grice, H. Paul, (1989), *Studies in the Way of Words*, Harvard University Press.

Hsieh, Nien-hê and Henrik Andersson, (2021) "Incommensurable Values", *in Zalta, E. N. (ed) The Stanford Encyclopedia of Philosophy.*

Kitcher, Philip, (2011), *The Ethical Project*, Harvard University Press.

Kitcher, Philip, (2021), *Moral Progress*, Oxford University Press.

Musschenga, Albert W. and Greben Meynen, (2017), "Moral Progress: An Introduction", *Ethical Theory and Moral Practice* 20(1): 3-15.

Macklin, Ruth. (1977), "Moral progress", *Ethics*, 87(4), 370-382.

Nagel, Thomas, (2012), *Mortal questions*, Cambridge University Press.

Nussbaum, Martha. C. (2011). *Creating capabilities: The human development approach*, The Belknap Press of Harvard University Press.

Rabinowicz, Wlodek. (2008), "Value relations", *Theoria*, 74(1): 18-49.

Rønnow-Rasmussen, Toni, (2016), "On Locating Value in Making Moral Progress", *Ethical Theory and Moral Practice.*

Sauer, H., Blunden, C., Eriksen, C., & Rehren, P. (2021), "Moral progress: Recent developments", *Philosophy Compass,* 16(10).

Sidgwick, Henry (1981) [1907], *The Methods of Ethics* (7th ed.), Indianapolis: Hackett Publishing Company.

Thomson, Judith Jarvis, (1997) "The right and the good", *The Journal of Philosophy* 94(6) 273-298.

[18] I wish to thank Mattias Gunnemyr, Anders Herlitz, Wlodek Rabinowicz and Jakob Werkmäster for their helpful comments on an earlier draft of this chapter. My work on this chapter was funded by the Swedish Research Council grant number 2018-06698.

Team Reasoning, Mode, and Content

Olle Blomberg

Abstract. A "we-intention" is the kind of intention that an individual acts on when participating in joint intentional action. In discussions about what characterises such a we-intention, one fault line concerns whether the "we-ness" is a feature of a we-intention's mode or content. According to Björn Petersson, it is an agent-perspectival feature of its mode. Petersson argues that content accounts are incompatible with theories of so-called "group identification" and "team reasoning". Insofar as such group identification and team reasoning are commonplace in many joint action situations, such an incompatibility would be a serious problem for content accounts. I here argue, however, that Petersson's incompatibility thesis should be rejected.

1. Introduction

Recently, Björn Petersson and I wrote a paper together. The paper is an expression of *our* collective view on collective moral obligation. While this view could be completely discontinuous with our respective personal views on the subject matter, I believe there is much continuity and agreement between our collective view and our personal views. However, I will here focus on an assumption that we make in our co-authored paper regarding which at least our levels of credence differ.

In the paper, we argue that, for it to make sense to ascribe a moral obligation to a group, each member must have a context-specific capacity to group-identify—to view their situation from their collective perspective—and at least have a general capacity to deliberate about what they together ought to do (to "team reason"). This assumes that, *in some sense*, "an individual agent can have attitudes that are held

from her group's viewpoint" (Blomberg & Petersson, 2023: 11). But how should this idea be understood?

A salient fault line in discussions about shared agency is how the attitudes required of participants who intentionally act together should best be characterised, where some argue that the collective nature of the participants' intentions is a feature of the intentions' contents while others argue that it is (also) a feature of their mode (Schweikard & Schmid, 2021). Petersson (2015; 2017) has articulated and defended the view that intentions are mental states that allows for perspectival variation, so that an individual agent can, as philosophers in this area sometimes put it, intend in the "I-mode" or in the "we-mode". According to Petersson (2017), the contrasting content approach to we-intention is incompatible with, and cannot make sense of, the capacities for group identification and team reasoning that I and Petersson argue are necessary for making sense of the idea of collective moral obligation.

In this paper, I critically examine Petersson's arguments for what I will refer to as his *Incompatibility Thesis* concerning the content approach to we-intentions. I argue that his arguments for it fail: The content approach is compatible with the group-identification and team-reasoning framework that I and Petersson draw on.

My aim here is not to argue against Petersson's mode account. I find his mode account both coherent and appealing, and it is arguably congenial for characterising the attitudes of team reasoners. I am thus not retracting from our assumption "that a version of the perspectival understanding of group identification is the most promising candidate for capturing the notion that would fulfil the role assigned to it in the 'team reasoning' framework (Petersson, 2017)." (Blomberg & Petersson, 2023: 12 fn. 16)[1] However, given that there is room in the content approach for distinguishing between explicit and implicit content, Petersson's mode account does not have a substantial advantage over content accounts of we-intentions. Because of this, my level of credence in our assumption is lower than Petersson's.

In the next section, I briefly introduce Petersson's mode account of "we-intentions" and the view of mode and content of intentional states which underpins it. I contrast the account with Michael Bratman's (2014) influential content account and what is arguably the mainstream view of content that underpins it. In section 3, I briefly introduce what team reasoning is and how it is or can be related to questions about shared agency and we-intentions. In section 4, I go on to present and critically discuss Petersson's argument(s) as well as potential additional arguments and considerations that could be advanced in favour of the *Incompatibility Thesis*. I conclude that the arguments do not succeed and that the *Incompatibility Thesis* should be rejected. Nevertheless, I suggest in the Conclusion (section 5) that the mode approach is nevertheless congenial for the team-reasoning framework.

[1] Within the team-reasoning framework, group identification is a mechanism for *agency transformation* (see section 3).

2. Petersson's Account of We-Intention

Accounts of shared agency are motivated by the ubiquity and importance of joint action in our lives. When we do things together, we do not merely perform actions in parallel or in strategic interaction. What is missing does not seem to lie in the agents' behaviour, but in the participants' attitudes. After all, judging from their behaviour, two snowboarders riding down a slope in close proximity may either be strangers who simply try to keep an appropriate and safe distance from each other, or friends snowboarding together. What makes the attitudes they would have if they were friends snowboarding together distinct from the attitudes they would have if they were strangers interacting strategically? On one type of account, it is their contents. Put in the jargon of the field, a "we-intention" is an ordinary intention with a distinct type of content. A we-intention is here the attitude that explains and rationalises an individual's participation in a joint intentional activity. On Bratman's (2014) content account, for example, each friend's we-intention would at its core be, roughly, an intention that they snowboard down the slope by way of this very intention and the other's intention that they snowboard down the slope and by way of their meshing sub-plans for snowboarding down the slope.[2]

On a different type of account, having a we-intention is not (mainly) a matter of having an intention that is about "us" or our joint activity; rather, it is a matter of having an intention that is "ussy", of having the intention in a collective way or mode (Schmid, 2014: 12). But what could this mean? Petersson's perspectival account of we-intention provides one possible answer to this question.

It is common to think that the content of an intentional state with a direction of fit—a perception, intention or belief for example—determines that state's conditions of satisfaction, such as the conditions under which the perception is veridical, the intention successfully executed, or the belief true (Ludwig, 2016: ch. 7). But according to François Recanati (2007), whose work Petersson draws on, the content of an intentional state does not determine the conditions of satisfaction on its own. Petersson relies on an "informal characterisation" and "common sense notion of 'content'" (2017: 211), where the content of an intentional state is what the state is directed at or about: "The thing that is believed, perceived, desired, intended, etc." (ibid.) On this notion of content, the conditions of satisfaction are partly determined by aspects of the intentional state's attitudinal mode. For example, if your intentional state is a perception, then for it to be veridical, what you perceive must then and there cause your perceptual experience. This self-referentiality of the perceptual experience and the reference to the subject of experience that it involves

[2] Bratman only provides jointly sufficient conditions for an interpersonal pattern of intentions and beliefs that could fulfil the functional role that he identifies "shared intention" with. However, necessary conditions for each participant's we-intention can plausibly be extracted from Bratman's account of shared intention (see Ludwig, 2016: 249-250). While I disagree with Bratman on some details, I believe that he has provided a powerful and illuminating reductive account of we-intention.

are, according to Recanati, not part of the content of your perceptual experience. Instead, it is part of the mode of the intentional state of perception.

By contrast to the attitudinal mode of perception, there is no self-referential and subject-referential condition in the attitudinal mode of belief. Your belief that your cat is on your hallway mat can be true even if the fact that the cat is on the mat does not play any role in the genealogy of your belief. Nevertheless, given the common sense notion of 'content', the truth conditions of your belief are not exhausted by its content. Your belief comes with a tacit perspective that determines how the content of your belief should be evaluated, such as that it is you in particular who has this belief and that it concerns whether the cat is presently on the mat that is now lying in your hallway. That my cat was on my hallway mat yesterday morning before I sold both cat and mat to you at noon is irrelevant to whether or not your belief today is true. Here, the truth conditions are determined partly by the circumstances of the agent and the way that the belief state is embedded in the agent's psychology and body. We can think of what the attitudinal mode contributes to the conditions of satisfaction as reflecting facts about the functional role of the attitude within the cognitive architecture of the agent as well as about the circumstances in which the agent is embedded (cf. Roth, 2000).

An intention is arguably self-referential in a way similar to a perception: a subject's intention is not successfully executed unless the intended action is (non-deviantly) caused by that very intention of the subject (see Roth, 2000). Petersson's innovative development of Recanati's view is the idea "that the subject of intention is a perspectival feature of intending rather [than] an element in its content" (2017: 212). Like in the case of perception, the self-reference to the subject's intention is, Petersson (2017: 212-213) argues, part of the attitudinal mode of intention rather than part of its content. The *subject of intention* is here a property of the attitude: it is the agent perspective in which the intention is held, and it is distinct from the *ontological subject* who has the intention. This thus allows for the possibility that participants who engage in joint intentional action each have an intention directed at the joint action which is held from their we-perspective. This we-perspective then partly determines how the content of such an intention should be evaluated, that is, what the intention's success conditions are. According to Petersson then, a state of intention has an agent-perspectival mode (2015: 30; 2017: 213-214).[3]

Petersson notes that classifying his perspectival account of we-intention as a mode account is misleading insofar as it suggests that we-intending is a distinct attitudinal mode along with perceiving, remembering or I-intending: "It would be less misleading to say that my approach assigns an additional, agent perspectival, feature to some kinds of attitudes, besides mode and content, and that some kinds of attitudes permit perspectival variation, not only when it comes to temporal and

[3] Petersson (2017: 214 n. 17) suggests that this framework could also be applied to perceptual states, allowing for a we-mode account of joint attention. I critically discuss such accounts in (Blomberg, 2018).

spatial perspectives, but also of agent perspectives." (2017: 213) With this qualification in mind, I interchangeably refer to Petersson's account as a perspectival account and as a mode account of we-intention.

Content and mode accounts need not differ regarding what the conditions of satisfaction for we-intentions are. Petersson's extension of Recanati's framework is itself compatible with the conditions of satisfaction for we-intention that is implied by Bratman's theory of "shared intention" for example (where a shared intention is an interpersonal pattern of we-intentions and beliefs that functions to coordinate participants' joint intentional action).[4] However, the accounts differ regarding to what extent the conditions of satisfaction are determined by the we-intentions' "contents" and to what extent they are implicitly determined by the functional role of the we-intention and the circumstances in which the participant (the ontological subject) is embedded. This is not an uninteresting difference. A mode account will arguably be less conceptually and cognitively demanding than content accounts such as Bratman's, which, for example, require participants to have higher-order intentions. One could also argue that it is better in line with our typical experiences as participants in joint activity. When engaged in joint activity, one is usually not focused on one's own and other's intentions. Except when things go wrong, one's intentions seem to rather simply be directed at the activity itself.

However, here things get slippery. Proponents of the content approach sometimes point out that when it comes to an account of we-intention, "the complex content of the intentions […] may be only tacit or implicit" (Bratman, 2014: 104). Furthermore, like Petersson, Michael Schmitz (2017) argues that agent perspective is part of the mode of an intention, but Schmitz takes the mode to be *part of* the intention's content. Similarly, Recanati (2007: 55) himself distinguishes between "the strict content of an intentional state . . . and its 'overall' or 'complete' content which includes the aspects of content determined by the mode."

It would be unfortunate if the dispute between proponents of mode and content accounts were merely verbal. In light of this, it is especially interesting that Petersson has argued that there is an important functional difference between perspectival (we-mode) we-intentions and I-mode we-intentions (intentions 'that we J'). Before examining Petersson's arguments for this in section 4, I need to briefly introduce the notions of group identification and team reasoning.

3. Team Reasoning and We-Intention

According to theories of team reasoning, individuals sometimes reason practically directly about the question "What should we do?", where this question is not

[4] Petersson (2007) has reservations regarding the conditions of satisfaction implied by Bratman's account, but these reservations are independent of the choice between a mode and a content approach.

equivalent to each individual asking the question "What should I do in light of my expectations about what you will do?" Team reasoning involves two stages. In the first stage, each individual (each team member) considers what it is best for the team to do. Given that there is one unique answer, this yields a judgement regarding what the team ought to do (a joint action). In the second stage, each then considers what he or she should do as part of that joint action. One reason for thinking that human beings sometimes do engage in such team reasoning is that it is arguably the best explanation of how individuals find the obvious and uniquely rational solution in a so-called Hi-Lo game.

To illustrate a Hi-Lo game, I will consider "the footballers' problem" (Sugden, 2003): Two players on the football team are trying to make a pass play. In the heat of the game, they cannot communicate. The pass play can be made to the left or to the right of the receiving player. A pass to the left would be preferable. A pass play to the left will be brought about if player 1 passes the ball to the left while player 2 runs to the left to receive it. Least preferable is a failure of coordination. The "game" can be represented as follows, where numbers represent utility for each player:

	Player 2	
Player 1	Left	Right
Left	10, 10	0, 0
Right	0, 0	5, 5

Orthodox game theory provides no determinate rational solution to this game. Each player is supposed to choose the best response to whatever she believes the other player will do. The theory tells each player: If the other plays left, then play left; if the other plays right, then play right. But whether the other plays left or right depends on what the other thinks that the player himself or herself will do. There is no factor that can rationally tip the players' expectations about whether the other will go left or right. But, for each player, passing/running left is intuitively the rational thing to do! Indeed, without viewing the situation through the theoretical lens of game theory, it is hard to see that there is a problem at all here.

The problem is not that players are acting egoistically, acting so that their own private preferences are satisfied. Indeed, it is natural to think that each player on a football team evaluates the outcomes in light of her "team preferences", that is, in light of her team's standard of success. (A player's private preferences need not always be aligned with her team preferences—what best furthers a football player's individual career goals need not always be what best furthers the shared goal of her team. Nevertheless, it is typically assumed that players (participants) share a single shared team utility function that is common knowledge between them, and that the team utility is equal to the average of their expected private utilities.) However, that the players have and act such that the shared team utility is maximised does not

solve a Hi-Lo problem. If the players are restricted to best-reply reasoning, who ask "What should I do for us?", then they arguably cannot rationally solve this coordination problem.

What is needed for the football players to overcome the problem is an *agency transformation*: the unit of agency presupposed in each player's practical reasoning must be changed from herself considered as a private individual ("I"), to themselves considered collectively as a team ("we"). Each player first selects the outcome that is best for the team (the first stage of team reasoning). The team reasoning footballers' problem could thus be represented like this:

	Player 2	
Player 1	Left	Right
Left	10	0
Right	0	5

Each then (in the second stage of team reasoning) intends to do his or her own part of the action profile—in this case pass/run left—that is likely to bring about the outcome that is best for the team.

According to Michael Bacharach (2006), whose account Petersson draws on, team reasoning is the result of "group identification" and the framing of a decision problem as a problem facing the group or team. As a result of identifying with the group, a team reasoner frames the coordination problem as a problem for himself or herself and the other agents considered as a team. The notion of group identification is taken from the social identity approach in social psychology (for a review, see Hogg, Abrams and Brewer 2017). In Bacharach's and Petersson's view, an agent does not voluntarily choose whether or not to identify with a group. Rather, group identification is rather arationally triggered by circumstances and situational cues that make group identity salient. One such cue, speculates Bacharach (2006), is the strong interdependence that exists between individuals' interests in a social dilemma such as the Prisoner's Dilemma or Hi-Lo.

Like Petersson (2017), I will simply take Bacharach's account for granted. I take it that we do group identify and often tacitly or explicitly do engage in team reasoning. Given that situations involving joint action will often be situations that resemble Hi-Lo in that there is a need for coordination and a scope of mutual advantage, joint action will often involve group identification and team reasoning. If this is right, then an account of we-intention should at the very least be compatible with a theory of team reasoning.

It is not obvious where in the team-reasoning framework that one should locate the we-intention. Petersson identifies the conclusion regarding what we should do as the we-intention. In his view, a participant's we-intention is thus the output of the first stage of her team reasoning. But Natalie Gold and Robert Sugden (2007: 126,

128) instead identify the we-intention with the intention to do one's own part in the optimal profile—that is, with the output of the second stage of team reasoning. This suggests that a we-intention is simply an ordinary individual intention with a distinct etiology. Since Petersson takes the intention to do one's part to be an ordinary individual (I-mode) intention, there is no substantive disagreement here.[5] Further others identify the we-intention with a commitment that is found upstream of the team reasoning, so that the we-intention establishes the group identification (Bacharach, 2006: 199 n. 7; Hakli, Miller, & Tuomela, 2010; cf. Bratman, 2014: 181-182 n. 19). Similarly, while Gold and Sugden (2007) identify the we-intention with the intention to do one's part, they suggest that a Bratmanian shared intention can "set the framework within which" (136), and provide "the background circumstances in which" participants' team reasoning can occur (117). But if Petersson's *Incompatibility Thesis* is true, then whatever this suggestion comes to, a shared intention could not directly and rationally prompt the participants to engage in team reasoning.

4. Assessing the Case for the *Incompatibility Thesis*

Petersson's (2017) arguments for the thesis that content accounts of we-intention are not compatible with Bacharach's team-reasoning framework is set out in a critical discussion of Bratman's (2014) and Raimo Tuomela's (2007; 2013) accounts of we-intention.[6] In the next four subsections, I will consider different strands of Petersson's discussion.

4.1 The Mode-Mirroring Assumption

What appears to be Petersson's core argument for the *Incompatibility Thesis* is expressed briefly in the following key passage:

> Unlike 'we intend to J' phrases of the form 'I intend that we J' are not commonly used in ordinary language, but if there is a question to which such a phrase is the answer, this seems to be a question I may ask myself when I am about to form my intention concerning us. Like in the intention that is expected to result from this deliberative process, 'we' figures in the content of the question but it is asked from my perspective rather than ours. In this framework, my intention that we do this rather than that would presumably result from what I want for us, my caring for how well we do. So, it seems reasonable to assume that the Bratmanian co-operator would not

[5] Petersson explained this in an email sent to me on April 26, 2022.

[6] Tuomela's general approach is not exclusively content-based, but Petersson argues that Tuomela's definition of a we-intention implies that participants having ordinary intentions and beliefs with certain contents appear to be sufficient for having a we-intention.

> be asking what we should do in Bacharach's distinct sense. Being a Bratmanian co-operator would not help in the potential Hi-Lo game.
>
> […] I believe […] that the explanation of her [the Bratmanian co-operator's] failure is that the group merely figures in the content of her attitudes, and that this content is conceived from an individual perspective. No *agency transformation* occurs. (Petersson, 2017: 205, emphasis in original)

An intention of the form 'I intend that we J' is here referring to a we-intention according to a simple content account. What Petersson is saying here is that if his intention 'that we J' is formed as the result of his conscious deliberation (rather than acquired spontaneously in response to his situation), then the deliberation must be prompted by "a question I may ask myself when I am about to form my intention concerning us." (ibid.) Further, Petersson suggests that the *only* question that he may ask himself when forming such an intention 'that we J' is a question of the form "What should I do?" But why could he not rationally form an intention 'that we J' as a result of deliberation prompted by the question "What should we do?" We need some reason for thinking that he could not.

The assumption that a Bratmanian co-operator would deliberate and act in accordance with what "I want for us" may suggest that Petersson thinks that the problem is that the Bratmanian co-operator cannot act in accordance with the team's preferences. On this reading, "from my perspective" would thus mean something like "given my standard of success". However, Petersson is not excluding the possibility that the Bratmanian co-operator is guided by the group's standard of success. Petersson thus accepts that a Bratmanian co-operator could ask "What should I do for us?", where 'us' merely figures in the content of the question, and adopts the standard of success of the co-operator's team in working out the answer. I may want what is best for us, where this is different from what I want us to do: When snowboarding with a group of friends, I may personally want us to take the shortest and steepest route down to base camp together, but I may nevertheless reason and act based on a stronger desire that we do what is best for the group, where this may be finding a route that takes everyone's skill level, time constraints and scenic preferences into account. This is not in tension with the content approach.

While the Bratmanian co-operator can make decisions in light of what is best for the team or group, she cannot, according to Petersson, make decisions from her team's *agential perspective*. There is no room for a "genuine agency transformation" in the content approach (Petersson, 2017: 214, 216). In critically discussing Tuomela's view, Petersson writes:

> My suggestion is that to treat the switch from I-mode to we-mode as an agency transformation of the required sort, we need a *perspectival* condition on the we-mode, a condition requiring a collectivistic feature of the way in which an intentional state (in the head of an individual) is held, rather than just requiring certain kinds of contents in her goals and beliefs. (Petersson, 2017: 210)

This is suggestive but what exactly is the problem for the content approach supposed to be? In the following, I offer a diagnosis of why Petersson thinks that there is a problem here.

Suppose that I raise the question "What should we do?" aloud out in the open between us. There is uptake of this question on your part, and we start to deliberate about what we ought to do. This deliberation unfolds in a conversational mode, with you and I talking to each other. In this case, the posing of the question, the establishment of the audience's uptake, and the deliberation that follows are all part of an interpersonal activity carried out by *us* (see Clark, 1996). The subject who deliberates, "we", is here identical to the agent of the intended action that the deliberation results in—also "we". This is a recognisable phenomenon of shared deliberation that the content approach can readily make sense of (Bratman, 2014: ch. 7). Indeed, on the content approach, one might think that this is the only way in which the question "What should we do?" can be asked as a practical deliberative question. That is, one might think that the question must be publicly asked and addressed to ourselves (that is, to us) in this way. Without shared deliberation that is carried out through social interaction between us, it may seem that "I", the deliberating agent, could not really pose the practical question about what "we", the agents whose collective options are supposed to be under consideration, should do (unless I have decision-making authority over the others). At most, one might think, I could ask "What should I do concerning us?" This question would prompt best-reply reasoning, not team reasoning.

This line of thought builds on the assumption that the subject who deliberates cannot be distinct from the agents of the joint action that is intended as a result of the deliberation: only *I* can practically deliberate about *my* actions; only *we* can practically deliberate about *our* joint actions. Christopher Woodard (2011) calls this *the mirroring assumption*. According to the mirroring assumption, "the unit of action always matches, and is determined by, the unit of agency" (263). In the footballers' problem, the unit of action consisting of both players' component actions could, according to the mirroring assumption, only be the focus of deliberation and reasons for action for the unit of agency consisting of both players, where the players would have to make up something like a joint agent or group agent.[7] Bacharach's team-reasoning framework is clearly inconsistent with the mirroring assumption (Woodard, 2011). When team reasoning, each player is a separate unit of agency who deliberates about and responds to reasons that concern an extended unit of action that includes the contributions of both players.

Petersson is thus clearly not making the mirroring assumption. On his view, the ontological subject who asks and deliberates about what we should do is an individual. However, Petersson is arguably making an analogous assumption regarding "the subject of intention" and the agents of the joint action that is intended as a result of team reasoning. (Recall that the subject of intention is distinct from

[7] Petersson uses the term "unit of activity" rather than "unit of action" (2015: 32; 2017: 200).

the ontological subject, that is, from the agent having the intention.) He is making what I will call *the mode-mirroring assumption*: The agential perspective of an intention matches, and is determined by, the agential perspective of the practical reasoning that results in the intention being formed. Given this assumption, if agents are restricted to having intentions in the I-mode—ordinary individual intentions 'that we J'—then they can only deliberate effectively from the I-perspective. They could thus only ask "What should I do for us?", not "What should we do?"

If we reject the mirroring assumption, then we should arguably also reject the mode-mirroring assumption. As far as I can see, the plausibility of the mode-mirroring assumption piggybacks entirely on the plausibility of the mirroring assumption. I suspect it seems plausible due to the same (attractive but mistaken) line of thought that I sketched above to illustrate the mirroring assumption. Given that the very idea of team reasoning depends on rejecting the mirroring assumption, those of us who believe that team reasoning is a valid form of practical reasoning should arguably reject both the mirroring assumption and the mode-mirroring assumption. At any rate, Petersson does not provide any argument for why the mode-mirroring assumption should be accepted. Hence, it is unclear why I could not rationally form an intention 'that we J' as a result of deliberating about what we should do (rather than about what I should do for us).

4.2 Prediction and Deliberation

Those who argue for the possibility of intentions 'that we J' have argued that an agent can intend 'that we J' because the agent can make reasonable assumptions or conditional predictions about what others will intend and do if she manifests that she intends that they, she and the others, J (Bratman, 2014: 73-75; Ludwig, 2016: 208-210). For example, shared background assumptions may enable such intentions to be formed, or an agent can often reliably predict that another agent will do their part of their joint J-ing when the other recognises the agent's intention that they J. These assumptions or predictions about others' behaviour or participation is comparable to assumptions or predictions about other non-agential parts of nature: e.g. just as my rational intention to light a match depends on my assumption that there will be oxygen in the room, so my rational intention that we dance depends on my assumption that there is a reasonably good chance that you will accept my invitation to dance. This involves a stance toward other agents that is different from the stance towards others presupposed by team reasoning. Team reasoning involves the idea that an agent can directly deliberate about what "we" should do, where one takes a deliberative stance not only toward one's own actions, but also toward the actions of the other team members. In light of this, one might think that the content approach to we-intentions and the team-reasoning framework are incompatible.

However, it is not clear that there is any real tension between intentions 'that we J' and the deliberative stance toward others involved in team reasoning. Proponents of the content approach such as Bratman and Ludwig offer arguments for how best-

reply reasoners can rationally intend 'that we J' because they are not making the assumption that team reasoning is possible. But once team reasoning is on the table, it arguably provides a new route to rationally forming intentions 'that we J'.[8] (I examine this possibility further in the next section.) Unless one has reasons to reject the very possibility of intentions 'that we J', we have no reason yet to think that only a perspectival (we-mode) account of we-intentions can make sense of the sort of agency transformation that is part of the team-reasoning framework.[9]

4.3. Agency transformation and voluntary control

Even if the content approach to we-intentions is compatible with a team-reasoning framework, this does not mean that a content account such as Bratman's provides an explanation or account of "genuine agency transformation". One could thus criticise content accounts for being incomplete. Petersson follows such a softer line of criticism against Tuomela's definition of a "we-mode joint intention". About this definition, Petersson raises the warranted complaint that

> there is no explicit condition in this definition [of we-intending] *preventing* the Tuomelian we-mode reasoner from framing the situation as a Bacharachian team benefactor rather than as a team reasoner, i.e. in terms of what I should do for us, rather than in terms of what we should do. (Petersson, 2017: 208, my emphasis)

This complaint could certainly be raised with respect to Bratman's account as well. So, perhaps we should read Petersson as articulating a challenge to the content approach: it must make sense of the agency transformation required for team reasoning.

But why would not Petersson's complaint also be warranted against his own perspectival account? I have argued that it is not clear why team reasoning could not give rise to intentions 'that we J'. But I think it is equally unclear why ordinary individual reasoning could not give rise to Petersson's perspectival we-intentions. Take Bratman's example of the members of an audience at a wonderful concert acquiring intentions that we applaud. Bratman describes the circumstances in which these intentions are formed as follows: "each of these intentions in favour of the group's applause is formed on the assumption that the others also so intend, an assumption grounded in common knowledge of the kind of person who attends such concerts" (2014: 73). Given my reasonable assumptions about what the others will do, and given my desire of bringing about what is best for the whole audience, I ask

[8] When each has formed an intention 'that we J', there is rational pressure for each to intend to do their own part of their J-ing (Bratman, 2014: 64; Ludwig, 2016: 103-104). The transition from intending 'that we J' to intending to do 'my part of our J-ing' corresponds to the second stage of team reasoning.

[9] For scepticism about intentions 'that we J', see e.g. (Schmid, 2008). For responses, see (Bratman, 2014: 60-64; Ludwig, 2016: 102-104). Petersson is not sceptical about such intentions.

myself, "What should I do?", and arrive at the answer: I should clap as part of our applauding. On the basis of this answer, I could arguably then permissibly form the we-intention in favour the group's applause given that I desire that the whole audience applauds and my expectation that they will do so partly by way of the support of my we-intention in favour of the audience applauding. If there are perspectival we-intentions, and if Bratman's account of how intentions 'that we J' can arise is successful, then I do not see why this account could not also describe a route by which perspectival we-intentions could be rationally formed. This account does not require the participants to be team reasoners; they can be mere "team benefactors". It is thus not clear why Petersson's perspectival account of we-intention is just as incomplete as the content account when it comes to making sense of the agency transformation that is part of team reasoning.

If we can both intend that we applaud in the I-mode and intend to applaud in the we-mode, my suggestion that ordinary I-reasoning could lead one to rationally form either sort of we-intention would require that the agent had the capacity for voluntary shifts of agent perspective, a perspective that is supposed to be a feature of the mode. After all, practical reasoning can be a consciously controlled activity. If it is rationally permissible for me to either form an intention in the I-mode or an intention in the we-mode from the same premises, then it would be odd if I could not choose which we-intention to form.

Petersson would resist this. His view is that while we have much voluntary control over the *contents* of our intentions, the matter is different when it comes to control over the *mode* and agential perspective of our attitudes:

> [W]e do not seem to have the same capacity for voluntary shifts of attitudinal modes or perspective. I do not deliberately switch from fearing that p to believing or hoping that p. […] The same appears to for perspectival differences within an attitudinal category. It seems that I cannot directly choose how distant in time the object of a certain episodic memory should appear to me, for instance. (Petersson, 2017: 215)

If this is right, then we could not choose whether to form a perspectival (we-mode) we-intention or an I-mode we-intention. While this would not establish that group-identification and taking a deliberative perspective toward the group's joint action always leads to a perspectival we-intention, nor that ordinary best-reply reasoning always leads to an I-mode we-intention, such a match between modes of reasoning and the agential perspectives of the intention does seem natural. Perhaps this is just a fact of how our psychology works. There need be no deeper explanation of how different modes of reasoning and different modes of we-intending match up.

However, Petersson is mistaken about what is outside the agent's direct voluntary control. True, I could not directly transform the intentional state of fearing that p to a state of believing that p. But the agential perspective is supposed to be a perspectival feature that can vary within one and same attitudinal mode, in this case within the attitudinal mode of intention. Even if it is true that I cannot directly shift

the temporal perspective of my episodic memories, this is arguably not due to the perspective being part of the mode rather than the content, but rather due to the attitudinal mode being that of remembering. Consider that I also do not have voluntary control over the *content* of my memories. When it comes to the attitudinal mode of intention though, I *do* have voluntary control of what I intend, the intention's content. In light of this, I could arguably also have voluntary control of the agential perspective of my intention. Nothing follows regarding the agent's voluntary control over the agential perspective from the discovery that the collectivistic feature of we-intentions is part of their mode rather than their content.

I nevertheless find it to be a plausible hypothesis that group identification, and the switch between I-reasoning and team reasoning, is not under an agent's direct voluntary control. But this issue is distinct from whether the agential perspective and the contents of our intentions are under our voluntary control. Proponents of content accounts can, just as proponents of mode accounts, appeal to the socio-psychological theory of group identification and theories of team reasoning in order to make sense of agency transformation and the extent to which it is under our control. I-reasoning and team reasoning could both, it seems, result in either (I-mode) intentions 'that we J' or in we-mode we-intentions.

5. Conclusion

I have critically discussed Petersson's arguments for the thesis that a content approach to we-intention is incompatible with Bacharach's (2006) team-reasoning framework. The allure of his core argument can, I believe, be explained away once we see that if we accept the validity of team reasoning, then what I have called the mode-mirroring assumption, and not only the mirroring assumption, should be rejected. Furthermore, a proponent of the content approach can make use of the same ideas from social psychology and team reasoning theory to make sense of agency transformation as a proponent of the mode approach. To sum up, Petersson's arguments for the *Incompatibility Thesis* are unsuccessful.

My arguments do not bear directly on the plausibility of Petersson's own positive account of we-intention. In my view, his account is the clearest and best-developed version of a mode account available. I also think that there is an allusive fit between his account and Bacharach's theory of team reasoning. Petersson's account and Bacharach's theory both make salient the possibility of joint deliberation and joint intentional action in which agents are not explicitly thinking about and acting with respect to their group, but where the collective feature of the deliberation and action enters the experience of the participants in a more implicit way. In the case of Bacharach's theory, the collective feature is part of the players' framing of the decision situation; in the case of Petersson's account, the collective feature is implicit in the functional role of the we-intention. While such joint deliberation and

joint intentional action are not incompatible with the content approach, they are not the kinds of cases that this approach puts centre stage.

If an agent can intend from an I-perspective or a we-perspective, then it seems intuitively plausible that the mode of her practical reasoning—I-reasoning or team reasoning—should "colour" the perspective of the intention that this reasoning concludes in. Nevertheless, the explanation of agency transformation consists at its core of the group identification mechanism and the patterns of inference mandated by team reasoning (see Pacherie, 2013: 1834). Arguably, the introduction of a switch to a we-perspective in the intention's mode is an appealing embellishment to the theory rather than a required component of it.[10]

References

Bacharach, Michael (2006) *Beyond individual choice: Teams and frames in game theory* (N. Gold & R. Sugden, eds.). Princeton, N.J.: Princeton University Press.

Blomberg, Olle (2018) "We-Experiences, Common Knowledge, and the Mode Approach to Collective Intentionality". *Journal of Social Philosophy,* 49(1): 183–203.

Blomberg, Olle, & Björn Petersson (2023) "Team reasoning and Collective Moral Obligation". *Social Theory and Practice.* Advance online publication. https://doi.org/10.5840/soctheorpract2023120177

Bratman, Michael E. (2014) *Shared agency: A planning theory of acting together*. Oxford: Oxford University Press.

Clark, Herbert H. (1996) *Using Language*. Cambridge, UK: Cambridge University Press.

Gold, Natalie, & Robert Sugden (2007) "Collective intentions and team agency". *Journal of Philosophy*, 104(3):109–37.

Hakli, Raul, Kaarlo Miller, & Raimo Tuomela (2010) "Two Kinds of We-Reasoning". *Economics and Philosophy* 26(3): 291–320.

Hogg, Michael A., Dominic Abrams, & Marilynn B. Brewer (2017) "Social Identity: The Role of Self in Group Processes and Intergroup Relations". *Group Processes & Intergroup Relations*, 20(5): 570–81.

Ludwig, Kirk (2016) *From individual to plural agency – Collective action: Volume 1.* Oxford, UK: Oxford University Press.

Pacherie, Elisabeth (2013) "Intentional joint agency: Shared intention lite". *Synthese*, 190(10): 1817–39.

Petersson, Björn (2007) "Collectivity and circularity". *Journal of Philosophy*, 104(3): 138–56.

[10] Thanks to Mattias Gunnemyr, Kirk Ludwig and Abe Roth for helpful feedback on a draft of this chapter. The research was funded by the Lund Gothenburg Responsibility Project (PI: Paul Russell), which is in turn funded by the Swedish Research Council.

Petersson, Björn (2015) "Bratman, Searle, and simplicity. A comment on Bratman: Shared Agency, A Planning Theory of Acting Together". *Journal of Social Ontology*, 1(1): 27–37.

Petersson, Björn (2017) "Team Reasoning and Collective Intentionality". *Review of Philosophy and Psychology*, 8(2): 199–218.

Recanati, François (2007) "Content, mode, and self-reference" in S. L. Tsohatzidis (Ed.) *John Searle's Philosophy of Language: Force, Meaning, and Mind* (49–63). Cambridge, UK: Cambridge University Press.

Roth, Abraham Sesshu (2000) 'The Self-Referentiality of Intentions'. *Philosophical Studies* 97(1): 11–51.

Schmid, Hans Bernhard (2008) "Plural action". *Philosophy of the Social Sciences*, 38(1): 25–54.

Schmid, Hans Bernhard (2014) "Plural self-awareness". *Phenomenology and the Cognitive Sciences*, 13(7): 7–24.

Schmitz, Michael (2017) "What Is a Mode Account of Collective Intentionality?" in G. Preyer, & G. Peter (Eds.) *Social Ontology and Collective Intentionality: Critical Essays on the Philosophy of Raimo Tuomela with His Responses* (37–70). Cham, Switzerland: Springer.

Schweikard, David P. & Hans Bernhard Schmid, "Collective intentionality", in E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy* (Fall 2021 Edition), URL = <https://plato.stanford.edu/archives/fall2021/entries/collective-intentionality/>.

Sugden, Robert (2003) "The logic of team reasoning". *Philosophical Explorations*, 6(3): 165–181.

Tuomela, Raimo (2007) *The Philosophy of Sociality: The Shared Point of View.* Oxford, UK: Oxford University Press.

Tuomela, Raimo (2013) *Social Ontology: Collective Intentionality and Group Agents*. Oxford, UK: Oxford University Press.

Woodard, Christopher (2011) "Rationality and the Unit of Action". *Review of Philosophy and Psychology*, 2(2): 261–77.

The Assurance Problem for Transfers Between Generations and the Necessity of Economic Growth

Eric Brandstedt

Abstract. Population ageing is a fact of all advanced economies. Fewer people are born all the while current members live longer. The support which old people have come to depend on, for example through elderly care and pensions, thus becomes increasingly expensive. This accentuates an assurance problem. Although it has been and still is the case that the young are willing to support the currently old, this support is not unconditional. In return they trust that coming generations will support them one day. Historically pro-old welfare state institutions (e.g., pension systems) have offered individuals this assurance: their claim on future generation to support them has been credible simply by positive economic and demographic development. Economic growth has been a blessing for the cooperation between generations necessary to realise old age support. This paper describes this assurance problem in simple game theoretical terms, argues that it has been neglected in historically prominent justifications of pro-old welfare state institutions, and discusses what can be done to preserve trust in times of population ageing and weak economic growth.

Introduction

Population ageing is a growing problem in all advanced economies. Each new birth cohort is smaller than the previous one due to falling birth rates and because people live longer. The elderly dependency ratio increases, which means that more people

are in need of support and fewer are in a position to support them (see e.g., Harper, 2016; Bongaarts, 2004). This is a challenge to the social arrangements found in all welfare states through which some goods are transferred from those who work to the senior citizens who no longer can or will provide for themselves but yet have extensive needs. I will refer to such arrangements as pro-old welfare state institutions (Birnbaum et al., 2017) and focus mainly on unfunded, pay-as-you-go public pension systems, such as Social Security in the US, National Insurance in the UK and the National Public Pension System in Sweden.

I shall argue that population ageing reveals a problem with a previously considered unproblematic assumption behind the most common justifications of these pension systems, that is, that they depend on economic growth. The justification in question comes in prudential terms and is supplemented with only a weak sense of fairness. It goes: despite the fact that pro-old welfare state institutions seem to transfer vast sums of money between those currently in the workforce to the old, this is not an altruistic gift to the old, but rather a kind of loan or investment that is later paid back with interests by the next generation of workers. Pay-as-you-go pension systems are cooperative schemes between generations: Generation 2 (those currently working) pays for the old age support of Generation 1 and in return Generation 3 pays for their support, and so on. The goods in question are transferred upstream, from the young to the old, and indirectly reciprocated if and when the next generation makes their contributions. Thus, no one has to sacrifice anything in supporting these systems. Contributors are, as it were, investing in their own retirement, in the form of an institutionalised claim on future generations. In return they get a promissory note that they will be reciprocated by those who will work when they themselves are retired.

The problem with this is that the promise of a future return on contributions made today becomes less credible with population ageing because the upfront investment costs rise drastically. In the following section, I will elaborate on how population ageing erodes the trust young contributors need to support pension systems and pro-old welfare state institutions more generally. I call this *the assurance problem for transfers between generations* and argue that the only credible solution to it is further economic growth. In section three, I show that this fact has been insufficiently appreciated in the literature for the simple reason that it has seemed so obvious that economic growth will continue. But now, with population ageing and other threats to these prospects, this assumption must be scrutinised. In section four, I argue that the problem cannot be dealt with by merely switching to another justification – e.g., intergenerational justice or altruism – and so the conclusion is that if we want to maintain pro-old welfare state institutions, there is no alternative but to support economic growth (more concretely this could, for example, be done by increasing immigration or raising the retirement age).

2. The Assurance Problem for Transfers Between Generations

As we grow older, the risks of disease, injury and frailty increase (but note that on average, the number of healthy and able-bodied years also increase with population ageing). At a certain age individuals will no longer be able or willing to participate in productive work and thus need income support to maintain a decent standard of living. Before the establishment of the welfare state, old people depended on the good will of their younger relatives, which was a precarious dependence (Stuifbergen and van Delden 2011). Some had no relatives, others no one who cared for them. Another way of addressing this predicament is for individuals to save for their own retirement. An individual could, in theory, plan for their old age by putting away some of the surplus of their productive years to spend it when it is better needed in their old days. However, because no one knows how long they will live or how great needs they will come to have, individuals will in practice have difficulties in determining how much they should save. Furthermore, most of us are not that prudent but rather subject to various biases, and so likely to end up regretting our actual savings.

Pro-old welfare state institutions offer an effective, efficient, and fair solution to this problem. Collectively financed and organised pensions, health and elderly care pool risks and utilise economy of scale to provide an efficient insurance against these age-related social needs. Although such institutions have been designed differently in different places, they are typically justified in prudential terms. The argument is that *if* we were prudent, we would want to save some of the surplus we make in our productive years for when we are old. Rather than saving the money ourselves, which again is insecure, we can enter a social insurance agreement in which society as a whole is a risk-pool. To be actuarially fair, such a deal requires low premiums on the young with lesser needs and higher premiums on the old with greater needs – contrary to their differentiated abilities to pay. But the fact that all age allows for a neat solution to this mismatch: individuals can even out their contributions by paying more than their actuarially fair share during their productive years, as a saving for the higher premiums of their olden days which they otherwise could not afford. This enables mutually beneficial insurance solutions to the risks of old age.

Pro-old welfare state institutions, however, depend on cooperation between generations over time. This is easiest to see in the case of funded pension systems, as the contributions to them are later returned by coming generations. Whether the contributions come in the form of taxes or social levies, they give rise to legitimate expectations on a later return, often of a return with interest. For example, even if all transfers in pay-as-you-go pension systems are synchronic, it is not the case that the young simply give to the needy old (although a few particularly altruistic contributors may perhaps view it that way). Rather, they pay premiums to qualify

for later benefits. Contributors get promissory notes saying that they will be provided for when they are at a certain age. This is why these institutions depend on the cooperation between generations over time. Even when all transfers are synchronic, they depend on a continued inflow of contributions over time because this is how the transfers are justified: individuals contribute because they expect to get something in return.

This cooperation can be modelled in game theoretical terms as a version of a Hi-Lo game (Binmore, 2011; Heath, 2013). Consider the following scenario proposed by Ken Binmore (2011: 87):

> [I]magine a world in which only a mother and a daughter are alive at any time. Each player lives for two periods. The first period is her youth, and the second her old age. In her youth, a player bakes two (large) loaves of bread. She then gives birth to a daughter, and immediately grows old. Old players are too feeble to work, and so produce nothing. One equilibrium requires each player to consume both her loaves of bread in her youth. Everyone will then have to endure a miserable old age, but everyone will be optimizing given the choices of the others. All players would prefer to consume one loaf in their youth and one loaf in their old age. But this 'fair' outcome can only be achieved if the daughters all give one of their two loaves to their mothers, because bread perishes if not consumed when baked.

There are two equilibria here: one according to which there is no cooperation between the players and everyone ends up with more goods than they need in their youth and less than they need in their old days; and another according to which the players cooperate to produce the outcome in which the goods are continuously distributed between the players and so between each players' life periods. This explains the possibility of cooperation between generations needed to sustain pro-old welfare state institutions (in a steady state economy). In the world described by Binmore all that is needed to arrive at the fair outcome, as he calls it, is that previous players have given bread to their mothers, i.e., that the cooperative behaviour is under way. If this is a fact, then each new daughter can do no better for herself than to also give a loaf to her mother. Doing so is the only way in which she can expect bread when she is old. A nonconformist daughter risks being severely punished by her own daughter. Her fate is in the hands of her own future child; the cooperation is upstream.

As we see in Binmore's analysis, the cooperation required to sustain is not a solution to a Prisoner's Dilemma, that is, the situation is not such that each player is rationally required to defect. This is important because it suggests that the public goods served by pro-old welfare state institutions can be maintained over time and that free riding is not a major concern. It also suggests that the cooperation in question does not presuppose sentimental feelings, filial obligations or altruism. Even egoistic individuals can even out their consumption over the different life segments of a typical life and thereby protect themselves against age-related risks.

The analysis does not, however, show that there are no threats against the stability of this neat solution.

Whether or not a young contributor can later enjoy a certain level of old age benefits provided by the welfare state depends on what others do, in particular future workers, employers, politicians and officials. The value of the 'investment' depends on future generations making good on the claim thereby imposed on them. Thus, the young must trust that they – the future agents – will do their part in this transgenerational project if they – themselves – are to get their due. This is an instance of what Amartya Sen (1967) calls an 'Assurance Problem' (see also e.g., Runge, 1984; Kogelmann and Stich, 2016).[1] This is a problem of coordinating expectations in situations of interdependent choices, i.e., when what is best to do depend on what others do. If the young can trust that others will reciprocate, then it is best for them as well as for society at large to contribute. If they cannot so trust, they better not contribute and end up a sucker.

Considering the fact that pro-old welfare state institutions exist in most developed countries, one might then think that this problem has been dealt with. But this is not so. The reason is population ageing and other threats to economic growth. In the light of these dim prospects, these institutions cannot offer credible promises that everyone will continue to cooperate. As noted by Runge (1984: 171): 'If contributions to public goods depend on institutions' capacity to predict behaviour, then these institutions must be continually maintained in the face of normal degradation'. The specific assurance problem for transfers between generations highlighted by population ageing is overcoming the reasonable worry an individual might have today about how pro-old welfare state institutions are not maintained. Population ageing requires of each new generation that its members transfer an increasing share of the goods they produce during their productive years to maintain current benefit levels. Each generation gets a worse deal than the previous one and the rules determining contributions and benefits will become increasingly contested. The expectations for future benefits may soon exceed future generations' willingness to pay.

In Binmore's game, trust is easy to establish: each player must just expect that all other players (including future players) are rational and act on their own self-interest. In reality, however, it is more difficult. The payoff function of pro-old welfare state institutions is not only determined by the rationality of other agents, but also by exogenous factors, such as socioeconomic and demographic changes. As birth rates fall and life expectancy increases, each new player enters a less favourable cooperative system than their predecessors. Each new generation must transfer a greater share of what they produce during their productive years. This makes the promissory note they get in return riskier. It is as if each new daughter

[1] Note that the Assurance Problem has implications similar to a repeated PD (with the fear of retaliation), which is why Heath (2013) uses a repeated PD to analyse the structure of intergenerational cooperation.

had to produce an increasingly large loaf to feed their increasingly hungry mother. Even if their daughter *can* bake bigger breads, the upfront costs still increase which may make other options more tempting. There are always opportunity costs on savings for the future. Furthermore, if for whatever reason the granddaughter defects, that will primarily affect her compliant mother and the cost will be all the more burdensome on her because they will also be unfair: the mother would thus be punished both by the hard work she put in to baking the large bread and by not having any bread in her old age.

The rationality at play here involves more than the three directly adjacent generations, that is, the mother, daughter and granddaughter. The risk of defection ripples down backwards from one generation to the next. If some future generation finds that the deal is not to their advantage, that they would be better off eating their bread than passing it on, then it is not in the interests of their predecessors to cooperate either as they could not then count on their cooperation being reciprocated, and then this is true of their predecessors too, and so on. The assurance needed is that the cooperation will be maintained indefinitely (Heath, 2013).

That is not to say that the benefit ratio must stay put throughout the different iterations of the game. In Binmore's game it does: the difference between the contributions of the daughter (a loaf of bread) and the benefits she later receives from her granddaughter (a loaf of bread) is the same (=1) for each generation. In reality, however, benefit ratios often change from one generation to the next. Depending on economic productivity and demographic changes, a generation may get a handsome return on their contributions, whereas another ends up net-contributors over the course of their lives.

Considering the fact that existing pro-old welfare state institutions have produced varying benefit ratios and still continued to operate one might think that this is not a problem. Perhaps individuals will accept even a negative benefit ratio, i.e., that they are net-contributors, because they would still benefit from the possibilities of having protection against age-related risks (even if say, they end up taking not full advantage of the concrete benefits) and, if we are not comparing to alternative means of saving, something is of course better than nothing. But there are alternatives. If the benefit ratio is too low in the publicly funded pension system, an individual could do better for themselves by individual savings. To the extent that pro-old welfare state institutions rest on the rationale of indirect reciprocity as described above, each player must reasonably expect to get a fair return. The point is that if at some future time, the benefit ratio falls below a fair level, then individuals can no longer expect future generations to cooperate and so they may judge that it is in their best interest to find another solution to the risks of old age. If a player comes to suspect that the others (including future generations) will not do their part, they have a reason to defect to avoid the worst-case outcome of contributing to a system bound to crash.

3. Cheap Assurance in Good Times

Economists and philosophers have not seriously considered this assurance problem. The explanation for this oversight, I submit, is that they have had a positive outlook and not seriously entertained the prospect that economic growth could come to an end (cf. Forrester, 2019: ch. 6). They have assumed that the economy will continue to grow and that each new cohort will be larger than the previous one. As a result, they have not paid attention to this problem, but been assured by the growth prospect. In this section, I will substantiate this point by critically discussing two prominent accounts of pro-old welfare state institutions: one by the economist Paul Samuelson and the other by the philosopher Norman Daniels.

Paul Samuelson (1958), which is the original source of the standard justification of pension systems, assumed a growing economy: young contributors could generally expect benefits higher than their contributions because if there is population growth, each new generation contains more productive workers and so produces more things. If one expects this demographic trend to continue indefinitely, it is in the interests of all generations to contribute. As long as there is an inflow of new productive workers, everyone is made better off by agreeing to transfer goods from the young to the old. But he did, however, recognise something like the assurance problem. He noted that even as it is the self-interest of both the young and the old in a society to agree to support the elderly, this optimal distribution of goods between age groups cannot be guaranteed by 'cold and selfish competitive markets' (1958: 473). This is because individuals are not unconditional co-operators. They are only willing to transfer goods to the old if they are given some assurance that future generations will do so too. To overcome this problem, he argued, one should change the rules of the game: 'Let mankind enter into a Hobbes-Rousseau social contract in which the young are assured of their retirement subsistence if they will today support the aged, such support to be guaranteed by a draft on the yet-unborn' (1958: 479-80). The contract he envisioned was simply money and the institutions needed to maintain their value.

Money allows the young to store value they produce during their productive years and exchange it for consumption goods when they are old. That is, conditional on everyone accepting the value of money at a plausible exchange rate. The problem with this solution, though, is that this condition need not obtain: there is inflation. The money the young store during their productive years may lose its value, leaving them with insufficient protection when they need to exchange it for consumption goods. Whether or not their savings are maintained and protected against inflation is partly in the hands of future generations. The value of money, whether collected in pension funds or cookie jars, depends partly on the productivity of workers yet unborn when the savings decision is made.

This creates another assurance problem. The present generation investing their produce in money need some kind of assurance that the value of this intangible

goods will remain intact over time and allow them to exchange it for more tangible goods, such as food and housing, when they so need. Money in itself is an insufficient assurance to produce the social optimum. This also points to an important similarity between unfunded, pay-as-you-go pension systems and funded, individual or collective, savings schemes (see also Heath, 2013: 60ff). Both of them essentially are claims on future generations and depend on future generations making good on these claims by engaging in productive work.

Money alone does not provide a solution to the assurance problem. Although the function of money to store value over time is remarkable, this is a function which can only be effectively discharged in societies whose economies are well-maintained, which means that various institutions, such as a well-functioning government and a central bank, need to be in place.

This brings us to Norman Daniels's (1988) justification of pro-old welfare state institutions, which is contractualist and focused on the design of just institutions. The key idea is that a fair design of institutions such as pension systems is in everyone's self-interest due to the fact that everyone ages. There is no conflict between generations: the care for the elderly that the young provide is in their own interest. When they in turn are old and in need of assistance, they will be grateful that such services are available at an affordable price. In other words, if we were to think prudently about it, we would organise society such that a decent standard of living is maintained throughout the different stages of our lives.

The problem with this so-called 'prudential lifespan account' is that it assumes static background conditions and thereby fails to account for socioeconomic changes (for a more general critique see McKerlie 2013). This is not an accidental consequence, but a feature of this account (Daniels, 1988: 51f). If the prudent deliberator, for instance, knew that she was young, she might bias the savings plan towards the interests of the young, and vice versa, if she knew she was old, she might choose substantial transfers from the young to the old. Furthermore, the prudential choice must be binding on the entire lifespan of the agent. The alternative would allow the prudential agent to buy into a scheme with low contributions when she was young and then switch to one with high benefits when she grew old. The choice of the prudential deliberator must be set in stone and binding on all individuals in society, regardless of age and previous contributions. Thus, the prudential deliberator faces the choice behind a veil of ignorance, where she has no knowledge of her age, and furthermore faces the task of allocating an already fixed budget, a fair lifetime share. In reality, however, society is subject to socioeconomic change, as the economy and the population either grow or shrink.

Daniels recognises that socioeconomic and demographic changes may lead to a birth cohort problem, which is different from the age group problem he addresses with the justification laid out above. But he sees this as a practical problem with no bearing on the justification of the institutions in question. He writes: 'On my approach, at least in the case of health care and income support, the solution to the age-group problem is basic and the solution to the birth-cohort problem is

secondary, though both are important. Solving the birth-cohort problem requires "fine-tuning" the institutions which solve the more basic problem' (Daniels, 1988: 136). The fine-tuning he has in mind is marginal adjustments of the benefit levels of health care and pensions as the economy grows or shrinks. Inequalities between birth cohorts are unfortunate in that they may lead to a discontent which undermines the support for such institutions. Thus, 'approximate equality in benefit ratios should be a practical target of public policy' (Daniels, 1988: 128).

Daniels' reasoning glosses over the difficulties involved in dealing with the assurance problem and in particular the seemingly inevitable decay which these institutions presently are subject to. If benefit ratios of health care and pension systems slowly decline, each new generation is offered a worse deal than that of the previous generation. At some point on this slope, individuals will judge that it is not in their long-term interest to contribute to such institutions. However, even before that point is reached, they may reasonably judge it unfair that they pay more and benefit less than their predecessors did. A birth cohort which enjoys the benefits of systems, such as health care and pensions, without paying enough to maintain these systems is effectively a free-rider on the cooperation between generations which such systems rely on. This problem goes to the roots of the justification of these institutions.

Daniels argues that the conflicts over resource distribution between the young and the old in society are overblown. Once we reckon with the fact that we all age and adopt an age-neutral point of view in justifying age-related welfare state institutions, we will see that they are in everyone's interest. Again, this is only true against the background of a growing economy. A growing economy will likely lead to one kind of change in benefit ratios, namely the positive one that each generation turns out better off than its predecessors – and this inequality between birth cohorts does not strike many as unfair (indeed, the opposite: many believe that this is what intergenerational justice demands). The problem arises when considering the prospect of a shrinking economy. Even a credible possibility that this might be in the cards risks the cooperative project between generations which allows for a prudent allocation of consumption goods over the course of a lifetime. This is a central problem welfare states face these days and a clearly formulated response to it is lacking in the normative-political literature.

4. Are There Alternative Justifications?

Considering the facts that pension systems presuppose economic growth and that there are constraints imposed by population ageing, one might seek to anchor these systems in something more solid than that offered by the standard justification. I will here discuss some alternative justifications and reforms of these institutions, but argue that they all fall short of successfully dealing with the problem outlined.

A Signalling Device

The first suggestion is a signalling device by which co-operators can signal their willingness to cooperate. This kind of solution is often suggested for dealing with assurance problems in general and so might work in this instance too. A classic example (see e.g. Kogelmann and Stich, 2016) is a blood oath through which two agents signal their commitment to a joint project. The idea in short is that instead of merely saying that they will do their part of the project (cheap talk), they confirm their commitment by jointly taking on an upfront cost. Having done that, each of them knows that they are serious about undertaking the project and that they have a reason to want to get the project going. These days, the equivalent of a blood oath may be something like a down payment. The question, then, is if one could find a similarly credible commitment device in the case of pro-old welfare state institutions. Is there some way in which individuals could assure one another that they will continually contribute come what may?

It does not seem so. There is no obvious way in which future, yet unborn, contributors could signal their commitment to the project, and that is the kind of assurance requested. Present contributors could, of course, do so by transferring vast sums to the presently old, much like the daughter signals her cooperative intention by giving bread to her mother, but their upfront commitment will not be reciprocated now. One possibility, though, might be for the state itself to signal the commitment of future generations on their behalf. That is, something like a contract or constitution on future generations which compels them to follow through on their commitments. Let us therefore evaluate one solution which might be seen as an instance of this.

Musgrave's Rule

Musgrave's Rule states that the benefit ratio should be fixed at some level, such that each individual, no matter their birth cohort, enjoys the same (or roughly the same) benefit ratio (Musgrave, 1981: 109). If this is applied, then perhaps each contributor could trust that their cooperation would be meaningful because they would be guaranteed a fair return.

Underlying Musgrave's rule is a principle of fair risk-sharing between generations (Musgrave, 1981: 104). Different institutional designs have different risk profiles, as is seen in considering the two main kinds of pension systems: defined benefit schemes (DB), in which the benefits are determined and contributions varies in accordance with what is required to realise the benefits given socioeconomic change, and defined contribution schemes (DC), in which the opposite is true, contributions are determined and the effects of socioeconomic change only affect benefit levels. If we follow Musgrave in distinguishing only two variables in this context, productivity development and demographic change, then these two systems impose different risk profiles: roughly speaking, DB schemes

place all risks on the contributors whereas DC schemes place all risks on the beneficiaries. If productivity decreases or the elderly dependency ratio increases in a DB scheme, this imposes a greater burden on those in the labour force as they will have to make bigger contributions to social security, whereas in a DC scheme, contributions remain the same and the benefits instead decrease. Musgrave thought that neither of these risk distributions were fair, and argued instead for sharing the risks, as well as the windfalls, between contributors and beneficiaries. That is because they would produce differences in lifetime expectations of individuals belonging to different birth cohorts. In either a DB or DC scheme, an individual may be penalised by unfortunate socioeconomic changes. She may end up with a much lower benefit ratio than her older or younger relatives.

If Musgrave's Rule can be implemented in some feasible institutional design, it would indeed provide additional assurance that the inter-generational cooperation is worthwhile. A fixed benefit ratio would make these systems stable over time. It would also lessen certain risks which otherwise could make an individual reluctant to sign up to the intergenerational cooperation. But the rule is still relatively vague and so may fail for that reason. In particular, it says nothing about what is an appropriate ratio of benefits to contributions. Even if it gives assurance against suddenly falling benefit levels by pre-determining benefits, it could not fully guarantee the continuation of the cooperative project. Population ageing might lead to absolute burdens so extensive that some future generation still choose to opt out. Again, for the benefit ratio to be maintained in an ageing society, contributions must increase exponentially. Fewer contributors share the payment burden and increasingly numerous beneficiaries demand the benefit levels they were promised. This will strain the willingness to contribute: the system may seem like a pyramid scheme too risky to invest in.

Intergenerational Justice and Altruism

Another possibility is to argue that the justification of these institutions does not stand or fall by population ageing. One might, for example, suggest that they are matters of intergenerational justice. The argument might be that each generation is bound to do their fair share in the ongoing scheme of cooperation between generations. Or alternatively that the reason why a generation should contribute is an altruistic one: it is because of the needs of the elderly and nothing else. However the population ages and the strains of contributing to these systems increases, it is still a fact that there are old people with great needs. This, one might argue, is reason enough and it is simply unfortunate – but not unjust – that those in the workforce will have to work a bit harder to achieve this result.

Both of these proposals seem problematic though. The empirical trends we have considered risk making each new generation slightly worse off than their predecessors and now the argument is that they are still required to make sacrifices for their better off predecessors. This regressive transfer seems to be a perversion of

justice. Another problem is that they moralise pro-old welfare state institutions. If the demands imposed by institutions of intergenerational cooperation are grounded in some moral or perfectionist ideal which some citizens may reject (if not now, then at some later point), then contributing to these institutions by complying is all the more risky and costly. It is a great strength of the standard justification that the reason to contribute to these institutions is not a controversial moral ideal, such as filial obligations, rewards in the afterlife, or the present government dictating that this must be done. According to the liberal conception of legitimacy, which is prevalent in most of the states having these kind of institutions, political power is legitimate if and only if it is exercised in a way in which all those subjected to it can accept in the light of an understanding of themselves as free and equal members of society (Rawls, 2005; cf. Song, 2012). If every citizen sees the institutional order to which they are subject to as harmonising with their own view of themselves rather than as something alien, imposed on them by others, this will secure stable support for it over time.

Yet another problem is that future generations may at some point reasonably refuse to contribute as the contractual conditions have become just unacceptable (see also Vidlund et al., 2017). If population ageing continues and is not offset by e.g., productivity growth, the demands of complying with this principle will gradually become so burdensome that the young have to sacrifice resources they need to live a decent life to provide for their much more numerous predecessors. Again, it is not a Prisoner's Dilemma in which the young have reason to think that future generations will not cooperate irrespective of what they do. But they do have reason to worry that coming generations may reasonably refuse to comply with the cooperative enterprise because it is not in their interest as the expected benefit ratio is too low. The high equilibrium of cooperation between age groups risk being upset by unaddressed population ageing or even credible prospects thereof.

Ignoring the Problem or the Race to the Bottom Solution

Another possible response to the assurance problem is to lean on the built-in stickiness of these institutions and argue that this problem is merely theoretical. In reality, it is not the case that individuals of some cohort can opt out of contributing. Individuals are born into authoritative institutions and compelled to contribute whether they want it or not. In other words, the game theoretical model of the assurance problem is a misleading idealisation. In the real world, it is very difficult to mobilise enough political support to change these institutions. Consider, for instance, a scenario under which current workers have to pay hefty sums to fulfil existing pledges to the elderly and someone proposes to lower the benefit levels to reduce the payment obligations. Now, this may be clearly in the interests of the youngest workers (say, those under the age of 30), but not so for workers closer to their own retirement and for those who will retire within 10 years, it is perhaps to their disadvantage. Thus, it would be difficult to mobilise political support for

change and the more likely trajectory would be a race to the bottom with an increasingly worse benefit ratio for each generation.

Existing institutions create path dependencies and make certain otherwise irrational actions rational. This may be the most important explanation for why pro-old welfare state institutions so far persist despite ageing population structures. Then, of course, it is another matter whether this is just or right. One could judge that such a race to the bottom, where each new birth cohort is offered a worse deal than their predecessors due to population ageing is unfair and that future agents *ought* to opt out. It is furthermore unlikely that each new generation would let such unfairness pass. A dissatisfaction with pro-old welfare state institutions has been growing since the 90s and has already led to some reforms to avoid them being quashed under the pressure of the ageing population structure. Most likely, extensive reforms of existing institutions are required and have already been implemented in some countries.

Pro-growth Policies

As pro-old welfare state institutions presuppose economic growth for stable persistence, there is no better way of addressing the assurance problem than policies which aim to increase economic activity either directly by productivity growth or indirectly by growth of the working population.

Governments should foster an economic climate conducive to economic growth. This means promoting innovation, research and development, required infrastructure investments, controlling debt and inflation, and to not push economic externalities on the future. If there is sustainable economic growth each generation turns out better off than their predecessors and the assurance problem is dealt with. Perhaps there is not even the need for the economy to grow, but just not to contract, that is, at least a steady-state phase of development. Under this condition, new contributors will not get any interest on their, as it were, investments, but still benefit from an efficient way of allocating goods between their different life stages. This, of course, presupposes a constant population: if there is population growth, there must be economic growth.

Fortunately, if there is population growth, there is usually economic growth too because more individuals often mean more producers, innovators and consumers. Another way of seeing this is to think about these systems as pyramid schemes which demands a continual inflow of new contributors. Thus, an indirect way of securing pro-old welfare state institutions is through increasing the number of workers or hours worked. This could come either through increasing the fertility rate such that more people are born into the state, through increasing immigration, or through raising the retirement age. With the current trends in most advanced economies of decreasing birth rates, the first possibility seems less promising. One could, of course, imagine a government implementing various pro-natalist measures to increase birth rates (countries have done this for various reasons, although none

with any great success, see e.g., Togman, 2019), such as child benefits or tax breaks. A better alternative, although not without its problem either, is to increase immigration. An additional reason for this is that population ageing also increases the need for more staff in elderly care and health care. Finally, probably the best option is to offset some of the effects of population ageing by raising the age at which individuals retire. Doing so generates both more contributions through increasing the number of hours worked and lessens the benefit burden by shortening the time during which pension benefits are paid out. This, of course, presupposes that people are able and willing to work longer into old age.

5. Conclusion

Population ageing accentuates a difficult assurance problem for pro-old welfare state institutions. Everyone wants to secure a good standard of living for their old age, but they depend on others to cooperate in a transgenerational project necessary to realise this. Up until recently, this problem has been insignificant because individuals have been able to rely on rosy economic growth as an assurance that they will enjoy an even greater support when they are old. But not anymore. Population ageing means that each new birth cohort is required to make bigger contributions to see to it that existing pledges are met and get more uncertain and less credible promises that they will be fairly reciprocated.

Pro-old welfare state institutions are public goods, which allow individuals to even out their consumption over the course of their lives, as well as warranting them social protection against the risks of old age. They have been very effective at this by drawing on the economy of scale. They have also had the advantage of not depending on some controversial moral ground, which is why they have existed in different kinds of welfare states (liberal, conservative, and social democratic) and persisted regardless of political shifts. It is, however, a public good that essentially depends on overcoming this assurance problem. I have argued that if we want these institutions to continue to persist, there is no alternative but to promote economic growth.[2]

[2] It can seem petty and unkind to argue as I have done here and present it to you, Toni, Björn and Dan, as a pension gift. Let me therefore assure you that I sincerely wish you a good pension and that I wholeheartedly believe that you deserve one. The department of philosophy in Lund, where I started as a young student in 2005, has been and still is dear to me. I grew up academically in the milieu which the three of you strongly contributed to creating. This milieu was friendly, appreciative and encouraging but also straightforward, critical and questioning. It forced me to become an analyst and to hone my arguments. When I reflect back on my time as a student in practical philosophy in Lund, I think particularly warmly of Toni who was my first and perhaps most important mentor when I was a student. I can sincerely say that I wouldn't have pursued an academic career if it wasn't for your encouragement then.

References

Barry B (1997) Sustainability and intergenerational justice. *Theoria: A Journal of Social and Political Theory* 89: 43–64.

Binmore K (2011) *Natural Justice*. Oxford: Oxford University Press.

Birnbaum S, Ferrarini T, Nelson K and Palme J (2017) *The Generational Welfare Contract: Justice, Institutions, and Outcomes*. Cheltenham: Edward Elgar.

Bongaarts J (2004) Population ageing and the rising cost of public pensions. *Population and Development Review* 30(1): 1–23.

Daniels N (1988) *Am I My Parents' Keeper? An Essay on Justice Between the Young and the Old*. Oxford: Oxford University Press.

Forrester K (2019) *In the Shadow of Justice: Postwar Liberalism and the Remaking of Political Philosophy*. Princeton: Princeton University Press.

Harper S (2016) *How Population Change Will Transform Our World*. Oxford: Oxford University Press.

Heath J (2013) The structure of intergenerational cooperation. *Philosophy & Public Affairs* 41(1): 31–66.

Kogelmann B and Stich SGW (2016) When public reason fails us: Convergence discourse as blood oath. *American Political Science Review* 110(4): 717–730.

McKerlie D (2013) *Justice Between the Young and the Old*. Oxford: Oxford University Press.

Musgrave RA (1981) Reappraisal of financing social security. In: *Public Finance in a Democratic Society. Vol. II: Fiscal Doctrine, Growth and Institutions*. New York: New York University Press, pp. 103–122.

Ostrom E (1990) *Governing the Commons: The Evolution of Institutions for Collective Action*. New York: Cambridge University Press.

Rawls J (2005 [1993]) *Political Liberalism, exp. edn*. New York: Columbia University Press.

Rothstein B and Teorell J (2008) What is quality of government: A theory of impartial political institutions. *Governance* 21(2): 165–190.

Rothstein B (2012) Good governance. In: Levi-Faur D (ed) The Oxford Handbook of Governance. Oxford: Oxford University Press.

Runge CF (1984) Institutions and the free rider: The assurance problem in collective action. *The Journal of Politics* 46(1): 154–181.

Samuelson P (1958) An exact consumption-loan model of interest with or without the social contrivance of money. *The Journal of Political Economy* 66(6): 467–482.

Sen A (1967) Isolation, assurance and the social rate of discount. *The Quarterly Journal of Economics* 81(1): 112–124.

Song E (2012) Rawls's Liberal Principle of Legitimacy. *The Philosophical Forum*: 1–21.

Stuifbergen MC and van Delden JJM (2011) Filial obligations to elderly parents: A duty to care? *Medicine, Health Care and Philosophy* 14(1): 63–71.

Togman R (2019) *Nationalising Sex: Fertility, Fear, and Power*. New York: Oxford University Press.

Vidlund M, Väänänen N, Mielonen A and Kuitto K (2017) Pension system design and intergenerational redistribution: Applying musgrave's rule in a comparative setting. *Review of Sociology* 27(4): 40–60.

Preference, Information, and the Problem of Big Decisions

Johan Brännmark

Abstract. Many of the examples considered by philosophers when discussing preferences concern choices between relatively specific and simple objects, *e.g.*, me having a preference for an apple over an orange at t_1. Such preferences seem to have a straightforward relation to what it is rational for me to choose, and possibly also to what would be good for me. Some authors, like Dan Egonsson and Edna Ullmann-Margalit, have however worried about whether our standard way of thinking about preferences and rational choice will work when applied to bigger life decisions. In this paper, it will be argued that there really are deep problems with the idea of there being *best* options for how to lead one's life, but also that this should not be taken as grounds for thinking that there is something wrong with the idea that preferences matter. Instead, that there sometimes is no *best* option to choose is just a characteristic of what it is like to lead a human life.

Preferences matter. At least some of them. Depending on whether we are subjectivists or objectivists about the good or human well-being, we might differ on how much and in which ways they matter, but whenever a person has a clear-cut and reasonably well thought-through preference, it is something that we would typically take seriously when thinking about what we might do to benefit that person. Even many (perhaps even most) would-be paternalists will justify going against our current inclinations by saying that one is giving people what they *would* prefer or choose themselves if "they had complete information, unlimited cognitive abilities, and no lack of self-control" (Thaler & Sunstein, 2003: 1162).

Not all who emphasize the notion of *preference* go as far as Thaler & Sunstein, but most if not all place some kind of *information requirement* on the preferences that count (Egonsson, 2007). For the subjectivist, there is a delicate balance to be struck here: on the one hand, avoiding that misguided preferences count, on the other hand avoiding that the requirements are pushed up to superhuman levels, since we might then end up with *my* good being determined by the preferences of an ideal being that is alien to the person I actually am (*cf*. Rosati, 1995: 311). While striking this kind of balance is a perennial problem for subjectivists, the focus in this paper will be on a more specific problem (although it has some bearing on more general matters as well).

The decisions we make differ wildly in terms of where on the scale of complexity that the objects of choice under consideration are located. Sometimes we make small decisions, like whether one is going to eat an apple or an orange at a certain point in time, sometimes much bigger ones, like whether one is going to become a parent, which career to choose, moving to live in another country, etc. Preferentialists often assume that the notion of preference is just as applicable to all choices. But some theorists, like Ullmann-Margalit (2006) and Egonsson (2007), have worried about *big decisions* posing a special problem for preferentialism, or rational-choice theory in general. In the present paper, it will be argued that these worries should be taken seriously, and that we should think of the notion of preference as primarily being applicable to smaller decisions, not the big ones. This then has consequences for how we should think about prospective and retrospective judgments about such big decisions.

The Good and the Best

When it comes to matters of well-being, or the person's own good, philosophers typically identify *desire-fulfilment theories* (*e.g*., Heathwood, 2016) as one of the main approaches. In terms of the relevant attitudinal states that matter for such theories, there are two key notions that tend to be appealed to, *desires* and *preferences*. Many philosophers use these loosely and interchangeably, but there is an important difference between the two: desires are *monadic* states, while preferences are *dyadic* – the latter are essentially comparative. This means that while desires can be useful for understanding what is *good* for us, just knowing the direction in which our desires run will not give an answer as to which of two options is the *better* or the *best* one. Since we typically cannot get everything we want, even if we start out with considering what we desire, we ultimately tend to end up with questions about what we prefer or should prefer.[1]

[1] One can certainly talk about *strengths* of relevant desires (or pro-attitudes in general) and possibly compare these in order to determine what is best for us, but it is far from clear that we have a good grasp of what desire strength would mean, which would not ultimately turn on what we prefer; see Barrett (2019: 234-37) for a discussion of some of the options.

If we look to adjacent areas of inquiry, like decision theory and economics, the notion of preference is the central notion, but then there is also a live question about how it should be interpreted, where some favor a *mentalistic* interpretation, *i.e.*, preferences as motivational states in agents, capable of explaining why an agent did what she did, whereas others favor a behavioral interpretation, the *revealed-preference* account, where preferences just describe choice behavior. A lot of the modeling done by decision theorists and economists does not necessarily hinge on taking either stance, but we can still wonder both about which interpretation that is typically being assumed and which one is most reasonable. Hausman (2012) argues for the mentalistic interpretation on both points, but there are those who disagree, *e.g.*, Angner (2018) raising doubts about the first one, and Thoma (2021) about the second. Philosophers of well-being presumably tend towards some version of the mentalistic interpretation, however, since they will want an account of the good as rooted in motivational or evaluative mental states of the agent. To the extent that a mentalistic account of preferences makes sense, they can however potentially lean on well-developed accounts of rational choice that have been advanced within decision theory and economics.

What is a preference, then? Hausman (2012: 35) argues that preferences should be understood as *total subjective comparative evaluations*, and Bradley (2017: 47) puts forward a similar account: "a preference for α over β is best viewed as an all-things-considered comparative judgement that α is better than β that is instantiated in a disposition to choose the former over the latter when both are available". This is a type of account that should be in line with how many philosophers talk about preferences. According to it, if we prefer A over B there is no room for further evaluation between having this preference and deciding which of A and B that is the best. It is of course perfectly possible to speak loosely in terms of preferences also with respect to partial rankings, say, *e.g.*, one's preferences over wines simply by taste, but the notion of preference proposed by Hausman and Bradley involves all-relevant-things-considered rankings. In terms of what determines what is *best* for us, this would seem to be the relevant notion.

As a *descriptive* model, there are probably few people (at least by now) who think that rational choice unqualifiedly describes how human beings function all the time. But as an idealization, it is still possible that this kind of modeling allows us to make sense of much of the basic dynamics of human decision-making. Economics is probably the clearest example of a discipline where such models are used, and where the idea arguably is that findings about what idealized people would choose, and what would happen because of those choices, tell us something about the dynamics of choice in the real world.[2] This type of model might however also be understood in a different way, namely as articulating an *ideal* of rationality, *i.e.*, as a normative

[2] Sugden (2009: 7) notes that economists are not always explicit about how they view the relation between their models and the real world, but contends that "[i]ntuitively, they believe that their models support conjecture about the real world".

theory. The idea then is that both we and these theoretical constructs belong to the general kind *decision-makers* and that the constructs represent the perfection of being a decision-maker. The fact that we often fail to live up to the tenets of rational choice is as such no objection to it as a normative model – rather, it is a prerequisite.

It should be noted that if we opt for an account of preferences along the lines of Hausman and Bradley, then technically speaking, as actual human beings we often do not have preferences in that strict sense, since we will not have considered all relevant things. But even for everyday decision-making a distinction between partial and overall rankings would seem to make sense, so while the strict sense of preference is one that perhaps only characterizes a fully informed person, we can arguably be said to often at least approximate the forming of such preferences. And this account of rational choice could potentially then still serve as a kind of regulative ideal in our deliberations, as something that we could strive to emulate, and where good deliberation would be about informing ourselves and reflecting on that information in order to arrive at a sense of which options that we prefer, and in which order.

Thinking About Big Decisions

While it might occasionally be difficult to compare even a literal apple to a literal orange, many of the everyday choices that we face are relatively straightforward, and because we have previous experience with the alternatives, or at least something in their vicinity, we know what we like and often we also know what we prefer and when among the things that we like. But sometimes we also face very different choices, ones where the alternatives are highly complex and where pursuing one alternative will shape the course of one's life. A big decision. Here is an example from Sumner (1996: 129):

> Suppose that I find myself at a career crossroads when I am in college. On the one hand I am a star pitcher on the baseball team, courted by scouts who assure me that I have an excellent chance of making it to the major leagues. On the other hand I also have a brilliant record in philosophy, with the prospect of a career in university teaching. Up to now these two career paths have been compatible but now the former would lead me to the minor leagues while the latter would take me to graduate school. Because I realize that choosing either option will effectively foreclose the other, I investigate both as thoroughly as I can before deciding in favour of the long-range security of a teaching career. I go to graduate school, earn my doctorate, and land a job in a good philosophy department. There I find the demands of teaching and writing to be pretty well as I anticipated. Indeed, as the years pass everything goes more or less as expected except for the growing realization that this life is just not for me. My dissatisfaction at first manifests itself only in a free-floating irritability, but after a while it deepens into apathy and depression.

Everyone faces big decisions at some point in their lives, albeit perhaps not the exact choice between philosophy and baseball. The person in Sumner's example seems to have deliberated in accordance with rational choice as a regulative ideal, informing himself, reflecting, and forming a preference. But he ultimately still ends up with a sense of disappointment. A question that naturally invites itself is this: would he have been better off had he chosen differently?

Before we address such issues, we should first say something about the possible outcomes here. One possibility is that through careful philosophical work we can arrive at a conception of what grounds an option being better than another that will help us answer any such questions, or at the very least will make us confident in there typically being answers to them, even though epistemic limitations might often prevent us from arriving at those answers. But there is also another possibility. That careful reflection makes us realize that there often is no *best* option. The things that could ground an option being better than another might not always obtain. Indeed, there might be reason for thinking that they often will not, or even sometimes cannot. Of course, as a working hypothesis, this might sound partly defeatist, but we are not at the start of our collective inquiry here. Sometimes it might not be reasonable simply to keep on working under what might be called *the myth of the hidden* (Brännmark, 2021), assuming that there must be some theory X that will ultimately provide all the answers we initially want, and thinking that if a particular approach does not provide all those answers, there just has to be something wrong with it. Maybe we have already gotten all that there is to get.

Many theorists push more fundamental worries to the side in order to focus on more specific issues, but when it comes to the matter of there possibly being deeper problems with the idea of best options for big decisions, there are some philosophers who have pressed such fundamental worries. One example is Edna Ullmann-Margalit. Already in her early work with Sidney Morgenbesser (1977), she pointed to the limitations of standard rational-choice theory. The focus then was on very small choices, not just choices between apples and oranges, but between apples and apples. There are many situations where we just have to *pick* something, where there is no reason to prefer one thing over the other, but where we still have to actually move in order to get something. And we do. But the problem pointed to by Ullmann-Margalit and Morgenbesser was not practical, it was theoretical: that a very common type of action is one for which rational-choice theory does not have an adequate account. Friends of rational-choice theory might perhaps shrug this off by pointing out that in cases of picking we do not need rational guidance – it is basically a coin toss. In Ullmann-Margalit's later work, she develops her worries further, however. She suggests (2006: 157) that rational-choice theory might be understood as analogous to classical Newtonian physics, which holds well for a middle range of objects, but not for the extreme micro and macro ends. Similarly, rational-choice models hold well for a range of middle-sized, ordinary decisions, but not when it comes to very small decisions, like picking, or very big ones. It is this latter class to which Sumner-style examples belong.

More precisely, Ullmann-Margalit is looking at decisions with the following four characteristics: (i) They are *transformative* or 'core affecting', changing one's life projects, making one into a different person than one would otherwise be or become. (ii) They are *irrevocable*, and not in the trivial sense in which everything one does is irrevocable, but in that ordinary reversals are not possible. (iii) They are taken *in full awareness*, knowing that (a) one must make a genuine choice between viable alternatives, and (b) that the decision is transformative and irrevocable. (iv) The option not taken casts a *lingering shadow*; a consequence of full awareness is living with the option taken not just in isolation but in awareness of how it involved rejecting some other option. This kind of lingering shadow need not be about regret or disappointment, but it means that there is an added weightiness to how one decided.

Now, since there are four dimensions to how Ullmann-Margalit characterizes big decisions, it is possible that these might come apart, where some choices will have some of these but not all.[3] The first two features are also the ones that she highlights in her discussion, and which, she argues, are similar to how our reasons run out in cases of picking; they run out here as well. At least if you are a subjectivist about what is good for us, rationality operates within a certain frame of reference set by our beliefs and desires. But in making big decisions, we are in a way choosing such a framework, and "[i]f reasons are forever from within a system or a framework (Wittgenstein: from within a 'language game'), the choice of the framework itself cannot be justified by appeal to reasons" (Ullmann-Margalit, 2006: 171). These are cases of *opting* for something, rather than making a choice that can be fully determined by reasons. For some of these choice situations, there might be ways of just partly committing to an alternative, trying it out while keeping a backdoor open, but such strategies are typically only partly available, and in some cases not fully committing to an option might mean that one will be living a lesser version of the life in question. She also notes that the "evidence seems to suggest that people are in fact more casual and cavalier in the way they handle their big decisions than in the way they handle their ordinary decisions" (Ullmann-Margalit, 2006: 165), that people often *drift* into certain life paths rather than consciously *opting* for them, perhaps partly because we find opting situations difficult to deal with.

Another philosopher who has raised a worry about big decisions is Dan Egonsson (2007). In looking at the type of information requirement that preferentialist accounts of the good often come with, Egonsson notes that while these standardly are framed in *quantitative* terms (having all the relevant information), there is arguably also a *qualitative* dimension, one that is not captured by a set of

[3] Especially the *transformative* aspect has been the focus of some discussion in recent years, with Paul (2014) being a seminal work. Paul distinguishes between experiences being *personally* or *epistemically* transformative, where becoming a parent would exemplify both, but something like tasting durian fruit for the first time would just be epistemically transformative.

propositional attitudes such as beliefs.[4] For instance, Mary the fruit scientist might know everything about apples and everything about oranges, including how other people have described what it is like to eat them and how they taste. But until Mary has eaten both apples and oranges herself, can she really have a fully informed preference for one over the other? Something would seem to be missing.

This far, Egonsson's point mainly pertains to questions about how relevant certain theoretical constructs are to us as human beings (since these constructs are abstractions, they tend to be little more than bundles of propositional attitudes). But even if the qualitative dimension is important, for many of our everyday choices we already have the relevant experiences, or at least similar-enough experiences for being able to vividly imagine what it would be like to have one or the other of two options. Our real-life preferences can accordingly often have the relevant qualitative foundation. Egonsson worries, however, about choices like *becoming a philosopher*. If we take something like Sumner's example, the person could be understood as having been successful in vividly imagining the different *components* of the life. Maybe I can, already as a student, imagine what it is like to write a paper or to give a lecture, even though I have only done lesser versions of these up until that point. But leading a certain life involves doing things over and over again, and there are then cumulative effects that will shape how one's experience of these different components will evolve over time. Even if we can imagine what certain elements are like, and even if we might even imagine sequences of events in a certain order, such imaginings will inevitably be severely compressed. Something will still escape us, namely "the quality that is a result of experiencing every single element in the time sequence in a certain order and *tempo*" (Egonsson, 2007: 37). And unlike with learning about how different fruits taste, sampling will not work. Similar to Ullmann-Margalit, Egonsson identifies a problem having to do with scale*:* how a certain model of forming reasonable preferences might work well for many everyday smaller decisions, but not as well for highly complex macro decisions about things like which path one's life should take.

In addition to these worries, there is also another feature of big decisions that should be noted. Even to the extent that we form something like a preference for A over B, if it is a big decision, both A and B will inevitably be what might be called *skeletal* objects of choice. While the problem in cases of *picking* is that two objects are basically indistinguishable, so that there is nothing to set one option apart from the other, the problem here is rather that the options are not just neat packages where it is more-or-less determinate what one will get. There is no one way of being a philosopher (and no one way of being a baseball player either). Even if one has a more specific idea of, say, *being a philosopher* in mind, like Egonsson's (2007: 28) example of being a Wittgenstein-like philosopher, this is still a skeletal conception of a life as a philosopher, which can then be filled out in many different ways. Similarly with a type of choice that is often taken as a paradigmatic example of a

[4] Egonsson is influenced here by the vividness requirement put forward by Brandt (1979: 111-12).

big decision: whether to become a parent or not. Even setting aside the difference, especially given the social expectations, of becoming a father or becoming a mother, children do not come out of a single mold. You never become a parent *simplicter*, you become a parent of a specific child or, eventually, several specific children. Your experience will be very much colored by the quirks and traits that make any child into a specific human being. And while it seems unlikely that you can become a parent without certain shifts taking place in your life, these will still depend in exact character on what kind of work you have, what your other social relations are like, and so on. In short, while parents will certainly share some broad types of experiences that non-parents will lack, one should not assume that there is such a thing as *the* experience of being a parent. An important part of this variability is that the kind of skeletal objects which feature in the relevant preferences will entangle with other parts of our lives. If I choose a life in academia then that will surely influence what my parenthood will look like, and if I choose to become a parent that will shape my career in academia. It is not like choosing an apple over an orange at t_1 and then just having that apple.

Now, the mere fact that the exact outcomes of our choices are complex and involve elements of chance is not as such a problem for traditional rational-choice models or expected-utility theory. They are built precisely to handle that. There is accordingly an obvious strategy that one might pursue when conceptualizing big decisions: to think of them in terms of choices over complex lotteries. However, we would then run into a version of Ullmann-Margalit's point about scale: the fact that a certain solution works for certain kinds of choices does not mean that it works for all. To begin with, the lotteries in question would have to be very, very, very complex, because the different exact permutations that options like "having a career in philosophy" or "being a parent" will have are really multifarious, especially since these are options that interact with options from other choice situations that we face. While there are many mid-sized choices that are also uncertain in some respects, for such choices we might still be capable of assigning meaningful subjective probabilities to different possible outcomes (even if these are just estimates), but when we are considering possible life paths there will always be many unknown unknowns, if for no other reason than that such life paths unfold over several decades and will thus be entangled with large societal developments as well. It is simply impossible to know all the possible component and sub-component outcomes to which the relevant subjective probabilities would have to be assigned. There is radical uncertainty here. Philosophers often focus on examples which are designed so that we know all that matters about what will happen in the different options that we face, but as Jacobson (2013: 121) points out this often means that one "ignores the commonplace uncertainty under which we make decisions." For smaller decisions, this simplification can perhaps be warranted, but not for big decisions.

As already pointed out, there is a worry about how much idealization that will have to be involved in conceiving of which preferences that would be reasonable to

form. In potentially conceiving of big decisions in terms of choices over complex lotteries, where the agent would fully know and understand these lotteries, we would need to move up to not just superhuman but god-like cognitive and precognitive capacities. This looks like a move that would be made on pain of irrelevance. Another possibility here, in order to find a role for rational-choice models to play, might be to bracket some of the complexities involved in big decisions and just focus on their main salient features. To some extent this is probably what we actually tend to do in real life, but it is less clear if this is viable as a form of rational choice – it would seem to involve an arbitrariness in demarcating which considerations enter into our decision-making, an exercise in pretending that we are making a choice of a certain kind, when it is really a choice of a very different kind.

Prospective and Retrospective Judgments

The worries stated above about the limited applicability of rational-choice thinking should not be taken to mean that we can make no reasonable judgments whatsoever about different paths our lives can take. At the very least, some possibilities can be just obviously bad. If someone risks a life of being held captive and tortured for decades, then the finer points of not being able to imagine such a life because of how it would transform the person, how the tortures will be experienced in a drawn-out way, or because there are many different more precise ways in which the details can be filled out, pose no problem for being able to conclude that such a life would just be bad. But in the big decisions that matter, our interest lies not with the options that we already know are bad, but rather the options which have something going for them and where we want to know which option is the *best* one. This is where rational-choice thinking runs into problems with handling big decisions.

Now, it is often pointed out that in thinking about what is good or best for us, and which actions or events that might benefit us, there are two main perspectives that one can take, that of the *agent* and that of the *spectator*. But there is also another distinction between perspectives, one that is orthogonal to the first one, namely between the *prospective* and the *retrospective* – before and after a specific choice. Philosophers have often focused on prospective judgments, at least when it comes to agents.[5] When it comes to spectators, things are different, partly because one of the main tasks that we have as spectators is to react to what people have done. But with respect to agents, one might think that what really matters is getting the prospective judgments right – doing the right thing. With respect to such judgments, the problem of big decisions does not mean that we should toss the idea of being

[5] One important exception to this is the notion of *regret*, where there is a relatively extensive literature (*e.g.*, Williams, 1981; Bagnoli, 2000; Jacobson, 2013), but often this is a discussion about *moral* rather than *prudential* choices.

informed to the side, staying misinformed can still be unreasonable. The problem is rather that there is a limit to how far informing ourselves and deliberating on that information can take us. At a certain point, all that remains is something like a leap of faith (*cf*. Ullmann-Margalit, 2006: 172).

Yet even if it is true that prospective judgments are *more* important than retrospective ones, this does not mean that the latter are unimportant. For moral choices this might be obvious – it is difficult to see how we would be able to develop as moral agents without considering our past actions and learning from them. To a significant extent, the role of retrospective judgments in such cases is however largely prospective – it is oriented towards future choices of the same kind. One characteristic of big decisions, however, is that typically they do not involve learning experiences of this kind. If I choose to become a philosopher, in the sense of having a whole career in the discipline, it is not as if I then become better equipped to make that kind of choice the next time I am faced with it. Similarly, if one becomes a parent, then one is (barring tragic outcomes) a parent for the rest of one's life. Still, most of us do occasionally think retrospectively about such choices, wondering whether we made the right decision.

If we consider Sumner's example, it features the agent in both the prospective and the retrospective situation. Prospectively, he is facing a big decision, trying to make it in an informed way, thinking through both options thoroughly. But it is ultimately a situation where, even when being informed about the options, there is no knowing how either option will play out more precisely and how they will be experienced by him, especially over time. Retrospectively, he might find that he made a mistake, perhaps a faultless one, but still a mistake. How should we think about such retrospective judgments? As already pointed out, the notion of preference is essentially comparative, so at first sight it seems well-suited for guiding such judgments, maybe one was wrong in preferring one option over the other? But as already indicated, it is far from clear whether this type of model is applicable here. The person in Sumner's example cannot know what his life as a professional baseball player would have been like if he had gone down that path instead. Indeed, that life path could have played out in multiple ways. It is also quite possible that there are other versions of life as a philosopher that could have been his given various circumstantial factors playing out in certain ways, and where his feelings would be different. There could also have been other aspects of his life that turned out differently, and where these would be entangled with his professional life in ways that made him feel differently.

In looking at the paths our lives take, one aspect of the problem of big decisions is precisely that things are not determined once and for all by those big decisions, but that a number of small choices and events gradually put flesh on the basic skeletal object we opted for, and where we might never know where various such (in one sense) small variations will ultimately take us. Let us look at another example. Say that one decides to study philosophy at university. Even if one has certain ideas in place providing some direction, like wanting to study at a place

dominated by analytic philosophy, there are invariably more fine-grained details about exactly where and when one studies, who one's teachers and classmates are, and which key texts that one reads in one's formative years, details that all contribute to shaping the trajectory that one's philosophical life will take. For instance, with regard to myself, on my very first philosophy course I had a great teacher in the history of moral philosophy, who also happened to use as a main text for that course a book with a fairly Hegelian take on this history (a somewhat unusual choice given that it was predominately an analytic-philosophy department). And to some extent, if I look back at my own trajectory, it has been one of doing philosophy largely in the analytic vein, but with persistent Hegelian tendencies, and usually with an eye to the historical tradition of which one's own work forms a (very) small part. It is of course impossible to know what would have happened if that had not been my first course. Obviously, it spoke to me in a way that probably required some latent tendencies already to be in place, but we can have many such tendencies, and depending on which concrete environments that we end up in, different ones might come to dominate. So maybe if I had studied philosophy somewhere else, my trajectory would have been quite different. Better or worse? There is no retrospective standpoint from which that judgment can be made in a determinate way – depending on which more specific shape one's path in philosophy (or in any other career, for that matter) takes, one's standards of what is worthwhile, interesting, and valuable will be different. One's attempts at retrospective judgments are inevitably made *within* a framework.

Or take another example, this time from literature – both in the sense that it comes from a novel and that it is about the life of a professor of literature. The novel is *Stoner* (Williams, 1965), which takes us through the relatively unremarkable life of William Stoner, and in large parts it is about a life spent in academia.[6] One thing that this story captures well is how chance plays a role in putting flesh on the skeletal life paths that we opt for. On one occasion, Stoner fails a student that is a protégé of one of his colleagues. While he was aware of how this would upset his colleague, he felt that this was simply something he had to do. However, then it turns out that this is not just something that eventually blows over, but that his colleague will hold a grudge for years to come, even when he becomes head of department. In one way, it is just a petty grudge, but it becomes something that has a big impact on how Stoner's life unfolds. In this type of case, it might seem straightforward that Stoner's life would have been better if the whole incident with this student had not happened. On one reading, it is simply an incident of a kind that takes place within a given framework, and which can then be assessed based on the values and goals of that framework. This is the case with many of the small decisions and events that contribute to putting flesh on the bones of the skeletal life paths that we have opted for. Nevertheless, some choices that are small in one sense might at the same time constitute possible important forks in the road, even while staying on the same basic

[6] For a nuanced and philosophically informed reading of *Stoner*, see Gåvertsson (2020).

life path. While it seems clear that it would have been better for Stoner if he had not faced the decision of failing that student, it is less clear that he made the wrong choice when faced by it. He could certainly have acted differently, but maybe that would ultimately have meant a life led with less integrity. Which would be better? A life with more integrity or a life with more external accomplishments? As spectators (here: readers) we might not be able to tell. What we do know is that towards the end of the novel, Stoner does engage in retrospective thinking and he is, on the whole, content with his life. There might not be an answer as to whether it was the *best* life he could have led, but retrospectively it can still reasonably be understood as good enough for a human life.

What these examples point to is a kind of mixed subjectivist view. Preferences can still matter, when there are preferences, in the sense of informed all-relevant-things-considered comparative evaluations, to be had. For many choices and situations, we might however instead think in terms of *desires* (or other monadic pro-attitudes) rather than *preferences* as the central attitudes in terms of which we understand what is good for us. Of course, as already mentioned, it is relatively common among philosophers to think of desires as relevant motivational/evaluative attitudes, but then that often involves a kind of pure desire view. What is suggested here is instead a more complex position, where preferences are the relevant attitudes with respect to those choices for which preferences make sense, *i.e.*, for choices between different middle-sized objects (so to speak). For many such decisions, there are accordingly really options that are the *best* ones. We might not always know which, but in such cases there can be something like the right answer to the question about what to choose, and we can meaningfully try to deliberate, informing ourselves and imagining what the options would be like, under the regulative ideal of rational choice.

For big decisions, however, there will typically not be any *best* option, because there is no rational way of preferring one option to the other. There can however still be *largely good* and *largely bad* trajectories that our lives can take, both in terms of the basic skeletal options and in terms of how the relevant life paths play out more concretely. Even if we cannot say what would be *best*, as subjectivists about the good, we can at least say things like these: If a person is dissatisfied with her life, then there is a problem. If a person ends up reasonably satisfied, there is arguably no problem: one has at least ended up on the good side of things. Within a certain life path, there might be changes that can be made to how we lead it that would be improvements. But at a certain point, some possible changes will become so drastic that they would have transformed the entire framework of evaluation, and all that might then remain to be said is that such an alternative life path would simply have been different from the one we actually took.

References

Angner, Erik (2018) "What preferences really are." *Philosophy of Science*, 85(4): 660-681.

Bagnoli, Carla (2000) "Value in the guise of regret." *Philosophical Explorations*. 3(2): 169–87.

Barrett, Jacob (2019) "Interpersonal comparisons with preferences and desires." *Politics, Philosophy & Economics*, 18(3): 219-41

Bradley, Richard (2017) *Decision theory with a human face*. Cambridge: Cambridge University Press.

Brandt, Richard (1979) *A theory of the good and the right*. Oxford: Clarendon Press.

Bränmark, Johan (2021) "Means paternalism and the problem of indeterminacy." *Moral Philosophy and Politics*, online first.

Egonsson, Dan (2007) *Preference and Information*. Aldershot: Ashgate.

Gåvertsson, Frits (2020) "Platonic perfectionism in John Williams' *Stoner*." *SATS*, 21(1): 39-60.

Hausman, Daniel (2012) *Preference, value, choice, and welfare*. Cambridge: Cambridge University Press.

Heathwood, Chris (2016) "Desire-fulfillment theory" in G. Fletcher (Ed.), *The Routledge handbook of the philosophy of well-being* (135-147). London: Routledge.

Jacobson, Daniel (2013) "Regret, agency, and error" in D. Shoemaker (Ed.) A*gency and Responsibility*, vol. 1 (95-125). Oxford: Oxford University Press,

Paul, Laurie (2014) *Transformative experience*. New York: Oxford University Press.

Rosati, Connie (1995) "Persons, perspectives, and full information accounts of the good." *Ethics*, 105(2): 296-325.

Sugden, Robert (2009) "Credible worlds, capacities and mechanisms." *Erkenntnis*, 70(1): 3-27.

Sumner, Wayne (1996) *Welfare, happiness, and ethics*. Oxford: Clarendon Press.

Thaler, Richard, & Cass Sunstein (2003) "Libertarian paternalism is not an oxymoron." *The University of Chicago Law Review*, 70(4): 1159-1202.

Thoma, Johanna (2021) "Folk psychology and the interpretation of decision theory." *Ergo*, 7.

Ullmann-Margalit, Edna & Sidney Morgenbesser (1977) "Picking and choosing." *Social Research*, 44(4): 757-785.

Ullmann-Margalit, Edna (2006) "Big decisions: Opting, converting, drifting." *Royal Institute of Philosophy Supplement*, 58: 157-172.

Williams, Bernard (1981) "Moral luck" in *Moral Luck* (20–39). Cambridge: Cambridge University Press.

Williams, John (1965) *Stoner*. New York: The Viking Press

‘They Smiled at the Good and Frowned at the Bad’

The Fitting Attitude Analysis Reconsidered

Krister Bykvist

1. Introduction

‘They smiled at the good and frowned at the bad’, the famous children’s story tells us about Madeleine and her friends. We all agree that these smiles and frowns are in some sense fitting. It seems fitting to favour the good and disfavour the bad. This intuition (call it the fundamental fittingness intuition) almost feels like a truism. Indeed, it has become increasingly popular to try to turn this intuition into an explicit definition of value in terms of fitting attitudes. Furthermore, it is often assumed that to say that an attitude is fitting is to say something *deontic* rather than something evaluative. The resulting account, the fitting attitude analysis of value (the FA-analysis, for short), thus promises a reduction of the evaluative to the deontic.

However, what seems to be such a simple and compelling intuition has been shown to be fraught with difficulties when understood as an explicit definition of value. The friends of the FA-analysis have provided many ingenious replies, but we are far from reaching a consensus about whether these replies are successful.[1]

My aim in this paper is not to add some new objections to the FA-analysis or some new replies to old objections. I think it is time to take a step back and reconsider the reasons that led people to accept the account in the first place. Are these reasons always compelling? Do they univocally speak in favour of a reductive definition of value in terms of the deontic? I shall argue that the answer to each

[1] I count Toni Rønnow-Rasmussen, to whom this paper is dedicated, as one of the leading proponents of the FA-analysis and someone who has given some of the most ingenious replies.

question is 'No'.[2] The most important considerations that led people to adopt the FA-analysis can in fact be taken into account by a *value primitivist* who thinks goodness is not analyzable in terms of the deontic.[3] If this is correct, then, in light of all the objections to the FA-analysis, one should seriously ask oneself whether this analysis is worth the price. Why bother with patching up the FA-analysis, if value primitivism captures the most important intuitions that motivated the FA-analysis in the first place and, in addition, avoids its problems?

The paper is structured as follows. In Sections 2, I shall give a fuller characterization of the FA-analysis. In Section 3, I shall briefly present the most important objections to the FA-analysis. In Section 4, I shall discuss the main considerations that are taken to support this analysis. I shall show that these considerations are either not very compelling or when they do support the FA-analysis they also support a value primitivist account of value. The best version of the value primitivist account of value shares all the important virtues with the FA-analysis without falling prey to the objections that haunt this analysis. Section 5 concludes.

2. What Is the Fitting Attitude Analysis of Value?

Common to any complete FA-account is that it embraces the fitting attitude biconditional schemas for value (the FA-schema, for short):

FA-schemas:

x is good iff it is fitting for S to favour x.

x is bad iff it is fitting for S to disfavour x

x is neutral iff it is fitting for S to be neutral towards x.

x is better than y iff it is fitting for S to prefer x to y.[4]

x is equally as good as y iff it is fitting for S to be indifferent between x and y.

[2] Reisner (2009) also provides some interesting arguments for a negative answer to these questions, but he focuses mainly on the question whether the FA-analyst unwarrantedly relies on certain value intuitions.

[3] Value primitivism is thus just the denial of the reduction of the evaluative to the *deontic*. It does not exclude the reduction of some evaluative notions to others.

[4] Different notions of preference can be employed here. The main distinction is between the notion of a two-place attitude, such as a disposition to choose one item over another when presented with a choice, and the notion of a comparison of strength between two monadic attitudes: one item is more strongly favoured or less strongly disfavoured than another.

FA-schemas are supposed to be necessarily true and the right hand side in each conditional is more fundamental, either because the concept of the relevant value is analyzed in terms of the concept of fitting attitude, or because the property of value is analyzed in terms of the property of fitting attitude. I will mainly focus on the conceptual version of the FA-analysis, since this is the most popular version, but most of my points can be applied to its ontological cousin as well.

One important dimension along which FA-analyses differ is scope. The most ambitious version of the FA-analysis would analyze *all* evaluative notions in terms of fitting attitudes. A less ambitious form would analyze some but not all evaluative notions in this way. My main target in this paper is this more ambitious version, since it is only this version that can be said to offer a *wholesale* reduction of the evaluative to the deontic.

FA-analyses may also differ as to which deontic notion is being invoked in the analysans of value. Most defenders of the FA-analysis agree that it is not enough to say that the good is what it is fitting to favour and leave it at that, because the term 'fitting' is very ambiguous. For simplicity, I shall use the term 'fitting' in the following as a place-holder for some candidate deontic notion when the discussion does not require a choice of a specific deontic notion.

Another factor that is crucial for a complete FA-analysis is the choice of *subjects* of fitting attitudes: When x is good, for *whom* is it fitting to favour x? Plausibly, this depends on which value we are talking about. For example, when something is good for me but not you (part of my well-being but not yours) it is perhaps fitting for me but not you to favour it. But when something has intrinsic value or final value it is fitting for *everyone* to favour it (if they were in the right epistemic situation in relation to what is assigned value). Indeed, that intrinsic value is objective seems to be the standard view in the FA-camp, and I shall assume this in the following.[5]

3. The Problems with the Fitting Attitude Analysis

3.1 Two Circularity Threats

(a) Favouring with evaluative content?

An old objection to the FA-analysis is that in many cases the fitting attitude will have evaluative content.[6] For example, if you are admirable, the fitting response is admiration. But admiration of you can't just be some brute liking of you or the features that make you admirable. After all, I can just happen to like you and the

[5] One further complication that I sidestep here is that it might be fitting for a subject to not just favour something, but to favour something *for certain specific kinds of reasons*. For example, when it is fitting to admire someone it seems fitting to admire the person *for her strengths and achievements*.

[6] See, for instance, Ross (1939) pp. 276, 278 and Rabinowicz and Rønnow-Rasmussen (2004), p. 395.

particular achievements that explain your admirability. To admire you I also seem to need to judge your achievement as something *good*. But if this is conceded, defining admirability in terms of judging something as good would give us a circular account. One kind of goodness, admirability, is defined in terms of judging as good.

The value primitivist can simply side-step this issue and just accept that some goods merit a response in terms of evaluative attitudes. There is no need to adopt circular definitions or appeal to value experiences that don't require the possession of value concepts.

(b) Evaluative notion of fittingness?

It is crucial for a reductive FA-analysis that the notion 'fitting' can be understood in purely deontic terms. However, it is striking that very few defenders of the FA-analysis invoke paradigmatic deontic notions such as *obligation* and *duty*. Of course, it is perfectly understandable why they avoid such notions, for the following analyses are pretty hopeless:

x is good $=_{df.}$ S has an obligation to favour x.

x is good $=_{df.}$ S has a duty to favour x.

Obligations and duties are naturally thought of as being within one's voluntary control. But it is obvious that there are goods that one cannot favour at will. Furthermore, one can have an obligation or duty to care about something because one promised to care about it or because one has a role-duty to care about it, e.g., a parent's duty to care a lot about their delinquent off-spring. The object itself may lack value. We'll come back to this so-called 'wrong kind of reason' objection in Section 3.3.

One alternative is to link the relevant notion of fittingness to what *ought to be*. Roughly, the idea is that the good is defined as what it is most fitting to favour, and most fitting to favour is equated with what ought to be.[7] Even if I cannot muster a favouring for a good, it may still be true that it ought to be that I favour this good

One obvious problem here is that the notion of ought-to-be seems in many respects closer to the evaluative than to the deontic. Indeed, it is often called the *ideal* ought, the ought of *desirability*, or, simply, the *evaluative* ought. The ought-to-be notion seems to tick many of the boxes on the intuitive test list for the evaluative: it can be applied to situations and not just actions, failing to comply with an ought-to-be does not entail blameworthiness, and ought-to-be judgements do not settle the question of what is advisable to do, nor do they close deliberation about what to do. Indeed, it is popular to equate what ought to be with what is best (or a necessary condition for being best). If this is true, then an FA-analysis that invoked the ought-to-be would become circular: the good is defined as what it is *best* to favour!

[7] A proposal of this kind is suggested in Zimmerman (2001).

Among contemporary defender of the FA-analysis it is popular to invoke the notion of *reason* instead of the notion of duty, obligation, or ought:

> x is good $=_{df.}$ S has *reason* to favour x.

This account leads directly to the 'wrong kind of reason' objection, for we can have reason to favour what is not good. For example, all sides agree that it would not do to say

> x is good $=_{df.}$ S has *moral reason* to favour x.

since we can have moral reason to favour x because we promised to favour, we have a role-duty to favour, or because favouring has beneficial consequences. I'll talk more about this in Section 3.3. This problem aside, it is not clear that this account will provide a reduction of the evaluative to the deontic. For what do we mean by 'reason'? According to one very popular account, reasons are considerations that *speak in favour* (of actions or attitudes). On this account, if there is a reason for me to favour x, then there are some considerations that speak in favour of me favouring x. The expression 'speaking in favour' is of course very slippery but it should be noted that it lends itself easily to an evaluative reading:

> x speaks in favour of A-ing = rates A-ing a plus= x makes it to some extent *good* to A;

> x speaks against A-ing = x rates A-ing a minus = x makes it to some extent *bad* to A.

But, obviously, this reading could not be accepted by an FA-analysist who wants to reduce the evaluative to the deontic.

According to another popular account, reasons are *starting points in reasoning*. But not the starting point in any kind of reasoning, the starting point in *good* reasoning, presumably. But then we again get a circular FA-analysis, since it would in fact define goodness in part in terms of good reasoning.

Of course, I have not shown that reason-talk or ought-talk *must* be understood evaluatively. The point is just that the FA-defender needs to come up with positive arguments against evaluative readings of ought to be or reasons.

3.2 Embracing a Circular Analysis?

As a reply to this circularity threat, Rabinowicz and Rønnow-Rasmussen have recently suggested that it is best to abandon any reductionist ambitions and instead embrace a circular FA-analysis of value along the lines of:

> *Circular FA:* x is good $=_{df.}$ x has features that make x good and that provide reason for S to favour x.

They argue that this analysis is acceptable despite being circular, for circular analyses of this kind 'still allow us to exhibit structural connections between central concepts (value, reason, pro-attitude). Thereby, they can provide relevant information to those who have the concepts but are not clear about their mutual relationships.'[8]

It is true that this analysis gives us *some* information; it is not utterly trivial like an analysis of the form 'x is F $=_{df.}$ x is F', for 'x is good' will entail 'the features that make x good would provide reason for S to favour x'.

However, even though Circular FA is not utterly trivial, it is clear that much of what makes a philosophical analysis significant is lost. We can no longer say that the defining concepts are *prior* to the target concept, for one cannot grasp the defining concepts (which contain the concept of goodness) without first grasping the target concept of goodness. Nor can we say that we can determine the *extension* of 'x is good' by determining the extension of 'x has features that make x good and which provide S reason to favour x', for in order to determine the extension of the latter expression we need to know the extension of 'x is good'. In Humberstone's words, we have a case of *inferential circularity*.[9] This should make us question whether Circular FA should be seen as a philosophical analysis rather than just an informative conceptual truth.

Another consideration that speaks in favour of treating Circular FA as only an informative conceptual truth is that the only really illuminating conceptual truth is the *left-to-right* conditional:

> If x is good, then x has features that make x good and that provide reason for S to favour x.

The right-to-left conditional:

> If x has features that make x good and that provide reason for S to favour x, then x is good.

is not illuminating, since this is just an instance of the perfectly *general* conceptual truth that, for any x, F, and G, if x has features that make x have F and that provide reason to G x, then x has F. In contrast, the left-to-right implication is not an instance of a perfectly general conceptual truth.

Now, since the illuminating link is only in the left-to-right direction, one can wonder why we should insist that the *whole* biconditional provides an illuminating philosophical analysis. Indeed, one may wonder, quite generally, why a biconditional that is only illuminating in one direction should qualify as a philosophical analysis at all. It seems more plausible to say there is an interesting

[8] Rabinowicz and Rønnow-Rasmussen (2006), p. 120.

[9] Humberstone (1997). That Circular FA is inferentially circular is conceded in ibid.

analytical link between the concept of value and the concept of reason to favour and leave it at that.

It should be emphasized that a value primitivist can accept that there is such a link. It is a mistake to think that a primitive concept must be an analytically isolated concept; a primitive concept can have interesting analytical links to other concepts. More precisely, that 'x is F' analytically entails 'x is G' does not show that there must be some non-trivial condition H such that:

x is F $=_{df.}$ x is G and x is H

To take a non-moral example, it is an illuminating conceptual truth that

If x is crimson, then x is red.

But from this it does not follow that crimson is not a primitive notion and that there must be some illuminating analysis of crimson in terms of red.

To take a more philosophically interesting case, it is an illuminating conceptual truth that

If one knows that p, then one believes that p.

But from this it does not follow that knowledge is not a primitive notion and that there must be some illuminating analysis of knowledge in terms of belief. Indeed, primitivists about knowledge, such as Timothy Williamson, would deny that there is such an analysis to be found; they would happily accept that knowledge conceptually entails belief but insist that knowledge is a primitive concept.[10]

Similarly, it is perfectly consistent to be a primitivist about value and still accept that it is a conceptual truth that if x is good, then x has features that make x good which also give S reason to favour x. And she could of course accept similar conceptual truths about badness, neutrality, and betterness: (i) it is a conceptual truth that if x is bad, the x has features that make x bad which also give S reason to disfavor x, (ii) it is a conceptual truth that if x is neutral, then x has features that make x neutral which also give S reason to be neutral towards x, (iii) it is a conceptual truth that if x is better than y, then x and y have features that make x better than y which also give S reason to prefer x to y.

It is this kind of value primitivism I shall argue has all the benefits of the FA-account but none of its drawbacks. However, the conceptual links listed above are only approximately correct and thus in need of some qualifications, some of which I will add in the next section.

[10] See, for instance, Williamson (2000).

3.3 The 'Wrong Kind of Reason' Objection

Even if the FA-analyst succeeds in finding a suitable non-evaluative reading of ought to be or reason, she needs to find a good reply to the 'wrong kind of reason' objection, which we have already touched upon. Roughly put, the objection is that in many cases we have reason to favour something even though it is not good, (or reason to disfavor something even though it is not bad.)

The simplest counterexamples are cases in which the attitude of favouring what is not good *has some beneficial effects*, as in the famous 'Saucer of mud' example, in which a malicious demon will impose a severe punishment unless we desire a saucer of mud.[11]

Other examples involve intrinsically good attitudes whose objects lack intrinsic value. It seems perfectly fine to say that we have reason to take pleasure in neutral things, such as innocent silly games, because the pleasure-taking attitude itself is intrinsically good, even though the objects of this attitude lack intrinsic value.[12]

Another group of cases involve *moral* reasons to favour. For example, one can have strong moral reasons to protect and cherish one's children even though they are unworthy and bad, because parents have a moral duty to do so.[13] Or, we can imagine a case in which I have promised to protect something that lacks value.

Intuitively, the reasons for which we have reason to favour or disfavour things in these examples are of the 'wrong kind', hence the label the 'wrong kind of reason' problem. The reasons are of the wrong kind in that they seem not to have anything to do with the goodness or badness of the object. The challenge is to find a plausible characterization of the 'right kind' of reasons.

This problem has spawned an industry aimed at coming up with proposals. But it is fair to say that all are highly controversial and none of them has gained wide acceptance.[14]

3.4 The Problems with Solitary Goods

In nutshell, the problem of solitary goods is that there are good states of affairs that it is never fitting to favour, because it is logically impossible or unreasonable to favour them, if we understand favouring in any of the following ways:[15]

---

[11] Crisp (2000).

[12] See Rabinowicz and Rønnow-Rasmussen (2004), p. 404.

[13] See, for instance, D'Arms and Jacobson (2000), p. 69.

[14] For some proposed solutions, see, for example, Parfit (2001), Skorupski (2007), Rabinowicz and Rønnow-Rasmussen (2004), p. 420, Schroeder (2010), and Zimmerman (2011), p. 8.

[15] I discuss this problem much more thoroughly in Bykvist (2009).

> *Factive favourings*. This category comprises all favourings that are factive in the sense that if you favour p, then p is true. Examples are bringing about that p, successfully pursuing p, being glad that p, and taking pleasure in the fact that p.
>
> *Choice-dispositional favourings*. These favourings consist in dispositions to choose. Examples are desires, preferences, and intentions.
>
> *Doxastic favourings*. If you favour p, then you believe that p. Examples are being pleased that p, and being glad that p.

Here is how an FA-analysis who adopts factive attitudes gets into trouble. Suppose that hedonism is true. Then this is a good state of affairs, call it *Solitary Good*:

> there being happy egrets but no (past, present, or future) factive attitudes, agents, or believers.[16]

It cannot be fitting to take a factive attitude towards Solitary Good, since it is impossible to take a factive attitude towards this state of affairs, and it cannot be fitting to do what is logically impossible,

It is possible to have a choice-dispositional attitude towards Solitary Good, but it can hardly be fitting to take such an attitude towards this state of affairs since it is logically impossible to act on this attitude and bring about the state of affair.

It is possible to take a doxastic attitude towards Solitary Good. But it can hardly be fitting to take such an attitude towards the state of affairs since, necessarily, if a person believes that there are happy egrets but no believers, his belief is false. It can hardly be fitting to undermine oneself in this way.

3.5 The 'Distance' Problem

How strongly we should feel about something seems to depend on how 'close' we are to this thing.[17] For example, how strongly we should feel about something seems to depend not just on its value but also on *modal* matters. It does not seem to be fitting to have more intense emotional feelings towards better states of affairs that are only *remote possibilities*. For instance, it does not seem fitting to have a more intense emotional positive feeling towards my daughter's not being abducted by aliens and taken to an intergalactic torture chamber than towards her not suffering the pain of a serious car accident, which could more easily have happened.

[16] One could argue that this state of affairs is impossible, since egrets are likely to be believers. However, for the argument to work, we only need to assume that there are no believers of propositions about happiness, egrets, past, present, future, factive attitudes, agents, and believers. Even if egrets should count as believers, they are not able to entertain propositions of this kind.

[17] For a discussion of the distance problem, see Bykvist (2009) and Oddie (2005).

How strongly we should react emotionally seems also to depend on *temporal* matters. It seems fitting that the extreme horror we once felt towards some terrible massacre softens with time. Other things being equal, it is not fitting to feel the same intense emotion towards past sufferings that occurred thousands of years back in the past as we do towards some current suffering of the same severity.

The final example has to do with *partiality*. It seems fitting for a parent to feel more strongly about the happiness of their own child than the happiness of a stranger's child, even though these instances of happiness are equally valuable.[18]

In all these cases, the degree to which it is fitting to positively respond to a state of affairs does not correspond to the degree to which it is good. How strongly one should favour an objectively valuable object depends on the 'distance' between oneself and the object.[19] It is, therefore, all too crude to say that it is always fitting to feel more strongly about a better state of affairs or to be emotionally indifferent between states of affairs of the same value.

Value primitivists can happily accept solitary goods without imposing any *ad hoc* restrictions on the link between goodness and fitting attitudes. Freed from the idea that goodness must be defined in terms of fitting attitude, they can endorse a *qualified* link between goodness and fitting attitudes:

> If x is good, then if it is possible to favour x without undermining oneself and also possible to successfully pursue x, then it is fitting to favour x.[20]

The qualification of the link between goodness and fitting attitude is not *ad hoc*, since it is not fitting to favour (in a factive sense) what is impossible to favour, to favour (in a doxastic sense) what will be self-undermining to favour, or to favour (in a choice-dispositional sense) what cannot be successfully pursued.

Note that the FA-analysists cannot just add this qualification to the right-hand side of their definition of goodness. For if they do and define goodness thus

> x is good $=_{df.}$ if it is possible to favour x without undermining oneself and also possible to successfully pursue x, then it is fitting to favour x.

we get the absurd result that all states of affairs that are either (a) not possible to favour, (b) not possible to favour without undermining oneself, or (c) not possible

[18] For a proposed solution to this, see Zimmerman (2011). For a very thorough criticism of the main solutions to the partiality problem, see Sylvan (2021).

[19] This talk about 'distance' should not be taken too literally. I am not assuming that there is a metric of distance. Nor am I assuming that one should *always* care more about things that are closer. The full story is likely to be very complicated.

[20] For simplicity, I will ignore this qualification in the following when I talk about the true link between value and fitting attitudes.

to be successfully pursued are good, for such states of affairs vacuously satisfy the conditional in the definiens.

The value primitivist can easily take into account parental partiality without using *ad hocery*, since she can acknowledge that the fact that *my daughter* is suffering provides a *further* reason to disfavor any instance of my daughter's suffering. This further reason explains why I may have stronger overall reason to prefer an instance of a stranger's suffering to an instance of my daughter's suffering. More generally, the value primitivist can maintain that the overall reason to favour or disfavor x is a function of both the intrinsic value of x and facts about the *modal*, *temporal*, or *personal* 'closeness' to x.[21]

4. What Is So Fitting About the Fitting Attitude Analysis?

4.1 The Only Option for Certain Thick Values of the 'X-able' Form

It is commonly argued that even if the FA-analysis is controversial as a wholesale analysis of all evaluative terms, it *must* be true for value concepts of the '-able' form, such as desirable, admirable, enviable, and preferable, and so on.[22] Each of these value concepts involves an attitude: desiring, admiring, envying, and preferring, respectively. It is then argued that the obvious alternative is to define these concepts in terms of some deontic notion, more specifically, in terms of *reason* or *ought*. The desirable is defined as what we ought to or have reason to desire, the admirable as what we ought to or have reason to admire, and so on.

This argument moves too fast, however. It is right that we must all accept that desirable, admirable and their likes all involve attitudes. What is much more contestable is the claim that we need to appeal to something deontic, such as reasons or ought, rather than something evaluative. It is instructive to see that the standard dictionaries define being desirable, being admirable and its likes not as being something we ought to or have reason to desire, or admire and so on, but as being *worthy* of desire, or admiration. Now, the notion of being worthy is not obviously deontic rather than evaluative. Indeed, being worthy of desire is to have sufficient *worth* to be desired, and being worthy of admiration is to have sufficient worth to be admired. The general pattern here is that being worthy of X is the same as having sufficient worth to be Xed. As far as I have some intuitive grip on the distinction between the deontic and the evaluative, I would put the concept of worth in (or at least closer to) the evaluative category. Of course, it is open to the FA-proponent to go on to define worth in terms of reason or ought. But that is not something we all

[21] This is far from a complete list of objections to the FA-analysis. For some objections to Brentano-style accounts, see Bykvist (2021).

[22] See, for instance, Schroeder (2010).

have to agree on just because we agree that the X-able is worth being Xed. We need a positive argument to be convinced that worth should be analyzed in terms of clear-cut deontic notions such as reason and ought.

4.2 Conceptual Parsimony

Another popular argument for the FA-analysis is that it achieves conceptual parsimony. Instead of having two distinct conceptual categories, the evaluative and the deontic, we can reduce the evaluative to the deontic and only assume one primitive notion.

Even if no one would deny that, *all other things being equal*, we should prefer a more conceptually parsimonious account of value, it is not clear that all other things are equal in the case of the FA-analysis of goodness. Since the FA-analysis is fraught with problems, it is far from clear that it is on the whole preferable to a less conceptually parsimonious account of value. If there is a less parsimonious account – for example, one according to which the evaluative and the deontic are both primitive categories – that can avoid the problems that afflict the FA-analysis and still reap its benefits, then adopting the less parsimonious account seems to be the preferable option.

4.3 Explains Why We Are Justified in our Concern for Valuable Things

It is popular to claim that the FA-analysis explains why we are justified in our concern for valuable things. As Rabinowicz and Rønnow-Rasmussen claim:

> The virtues of the FA analysis are considerable: it demystifies value and explains why we are justified in our concern for valuable objects. The justification is immediately forthcoming if value is nothing but the existence of reasons for such a concern.[23]

In a similar vein, Danielsson and Olson claim:

> One feature of the Brentano-style approach that we find particularly attractive is that a kind of internalism will be included in the bargain. This will be a kind of internalism that establishes a necessary link between values and attitudes: necessarily, to claim that an object is valuable is to claim that a pro-attitude towards that object is (would be) correct.[24]

To think that this is an advantage of the FA-analysis over value primitivism is to ignore the resources available to the latter theory. As I explained in Section 3.2, a

[23] Rabinowicz and Rønnow-Rasmussen (2004), p. 400.

[24] Danielsson and Olson (2007), p. 520.

value primitivist view need not deny that there is a necessary link between values and attitudes. They can claim that this is a necessary *conceptual* truth. Even if goodness is primitive it need not be conceptually isolated. It can stand in interesting analytical relations to other concepts, such as the concept of reason to favour.[25]

4.4 Demystifies Value

It has been claimed that the FA-analysis demystifies value by reducing it to something less mysterious: a deontic concept. This is a questionable claim. First, it is not clear why – quite generally – evaluative notions should be seen as more mysterious than deontic concepts. After all, they both fall on the 'value' side of the traditional fact-value divide, where the fact-side seems less mysterious.

Second, it is a mistake to think that demystification requires reduction. You can demystify a concept without analyzing it. You can demystify a primitive concept by pointing out its analytical links to other concepts. A value primitivist who accepts that it is a conceptual truth that we have reason to favour the good seems to have a strong claim to conceptual demystification.

4.5 Buck-Passing: Explains Why Goodness Is Not In Itself Reason-Giving

All sides of the debate agree that there is some kind of link between goodness and reasons. We often have strong reason to promote what is good, and prevent what is bad, for example. One famous argument in favour of the FA-analysis is that it allows for a buck-passing account of this link. The goodness of something does not provide any reasons to favour it (at least no independent reasons). Rather, what it is to be good is to have some other features that provide reasons to favour it.

It is very doubtful whether buck-passing is a great advantage of the FA-analysis. First of all, it is not clear whether the buck can always be passed to something clearly non-evaluative.[26] For example, a good painting can merit appreciation because of the *elegance* and *balance* of its composition. But elegance and balance are thick value notions, and thus not purely descriptive notions.

Second, the value primitivist does not have to say that the *goodness* of an object provides (independent) reason. It is open for her to say that what provide reasons are the features that make the object good. If these features are themselves non-evaluative, the buck is ultimately passed to something non-evaluative. Indeed, the

[25] Reisner (2009) questions whether the FA-analysis really has the advantage of *explaining* why we are justified in our concern for valuable objects. If the FA-analysis provides a *definition* of the goodness *concept* in terms of being justified in caring about what is good, it is unclear whether we automatically have an explanation of this justification. In general, it seems odd to say that the definiens explains the definiendum. To take a non-evaluative example, it is odd to explain that someone is a bachelor by pointing out that he is an unmarried man.

[26] Crisp (2005).

value primitivist can even say that it is a *conceptual* truth that the features that make an object good provide some reason to favour it. Remember that the value primitivist can accept that it is a necessary conceptual truth that if x is good, then x has some features that make x good which also provide some reason to favour x. On this view, it is not the goodness of an item that provides the reasons to favour it but its good-making features. That the item is good *conceptually entails* that it has good-making features which also provide reasons to favour.

Even if the value primitivist can avoid the charge of treating goodness itself as reason-providing, it has been accused of making value *redundant*. If what provide the reason to favour an item are its non-evaluative features, it seems redundant to postulate a primitive notion of goodness. Why not just say that to be good *is* just to have features that provide reason to favour?

One obvious reason why we should balk at this strong identification of goodness with the property of having reason-giving features is the list of problems presented in Section 3 (e.g., the problem of solitary goods and the 'wrong kind of reason' problem), it seems unlikely that we can find an illuminating non-circular analysis of goodness. If this is right, then it is not redundant to postulate a primitive notion of goodness; it is indispensable.

4.6 Explains the Plurality and Incomparability of Values and Also What Unites All Values

It has been argued that the FA-analysis has an easy time explaining the plurality and incomparability of value without giving up on a unifying account of goodness. By 'plurality of value' is meant the idea that 'different goods are properly valued in different ways'.[27] For example, virtuous acts call for admiration and pleasant experiences for rejoice. Incomparability is expected to occur when two items differ radically in what kind of responses is called for. For example, since it does not seem possible to always compare the degrees of admiration to the degrees of rejoice, it is to be expected that some virtuous acts cannot be compared in value to some pleasant experiences. The FA-analysis could then say that what unifies all *good* things is that it is fitting to have some kind of *pro*-attitude towards them, and what unifies all *bad* things is that it is fitting to have some kind of *con*-attitude towards them.

These considerations do not univocally speak in favour of the FA-analysis, however. Starting with the point about unification, one could wonder what makes all the pro-attitudes 'pro' and all the con-attitudes 'con'. An FA-analysist with reductionist ambitions cannot say that an attitude is 'pro', if it is a fitting response to a positive value, and 'con', if it is a fitting attitude to a negative value. Nor can she say that an attitude is 'pro', if it involves a positive value judgement, and 'con',

[27] Anderson (1993), p. 10.

if it involves a negative value judgement. Note that a value primitivist could happily embrace any of these accounts.

It is also tricky to define pro-attitudes and con-attitudes in terms of some specific phenomenological feel, such as pleasure and displeasure, respectively. If we do, it is not clear that we can maintain the idea that different goods can call for radically different responses, not all involving pleasure or displeasure. Furthermore, it does not seem right to say that a higher degree of a pro-attitude always comes with a higher degree of some phenomenological feel. For example, it is doubtful that a higher degree of admiration always comes with a greater intensity of felt pleasure.

What about the plurality of values? The value primitivist could explain how different values are properly valued in different ways by simply stating that it is a necessary conceptual truth that if x is good in a certain sense or way, then this sense or way of value also determines what kind of response to x that is adequate. For example, if the sense of goodness has to do with the achievements and strengths of a person, then the proper response to a person who exemplifies this value will be admiration. Or, if we are talking about personal goodness, in Rønnow-Rasmussen's sense, then if x has this value for a person, then the proper response has to be *for the sake of her*.[28]

What about incomparability? The FA-analysis explains value incomparability in terms of incomparability of attitudinal strength. Can the value primitivist provide an alternative explanation of value incomparability? Well, nothing prevents the value primitivist from distinguishing different kinds of value: better *as an achievement*, better *as a pleasure*, better as *an artist*, better *as a philosopher*. She does not have to think that there must be some 'covering value' according to which the pleasure and the achievement can be compared. Obviously, they cannot be compared under the value as pleasures or the value as achievements.[29]

5. Concluding Remarks

I have argued that the best version of value primitivism accepts that there are important analytical links between goodness and fitting attitudes. More exactly, if x is good and we can reasonably favour x – that is, favour it successfully without undermining ourselves – then we have some reason to favour x. How strong this reason is may depend on other factors, such as the 'distance' – modal, temporal, or

[28] Rønnow-Rasmussen (2011), p. 46, (2021), p. 131.

[29] It has also been argued that an FA-analysis can provide a plausible account of *parity*. See, for instance, Rabinowicz (2008) and Gert (2004). I think that parity can be explained well without presupposing an FA-analysis, but space constraints prevent me from elaborating on this point. I have a discussion of this point in a longer version of this paper.

personal – between the favourer and x. I have also argued that this form of value primitivism can avoid the standard objections to the FA-account.

This form of value primitivism has all the important virtues that have been ascribed to the FA-account. It is true that an FA-account that reduces the evaluative to the deontic is more conceptually parsimonious, assuming that we are not going for a value-first approach. However, since the FA-account is fraught with serious problems that value primitivism can solve or sidestep, value primitivism seems on the whole to be the superior alternative.[30]

References

Anderson, E. (1993), *Value in ethics and economics*, Harvard University Press.

Bykvist, K. (2021), 'Brentano's fallacy: Moore's argument against Brentano's analysis of value', *History of Philosophy Quarterly* 38(3): 243-259.

Bykvist, K. (2009), 'No good fit – why the fitting attitude analysis of value fails', *Mind* 118(496): 1-30.

Crisp, R. (2000), 'Value …And What Follows. By Joel Kupperman', *Philosophy* 75(3): 458-62.

Crisp, R. (2005), 'How to Avoid Passing the Buck', *Analysis* 65(1): 80-85.

Danielsson, Sven and Olson, Jonas, (2007), 'Brentano and the Buck Passers', *Mind* 116(463): 511-522.

D'Arms, J. and Jacobson, D. (2000), 'Sentiment and Value', *Ethics* 110(4): 722-48.

Gert, J. (2004), 'Value and Parity', *Ethics* 114(3):492-510

Humberstone, (1997), 'Two Types of Circles', *Philosophical and Phenomenological Research* 57(2): 249-80.

Oddie, G. (2005), *Value, Reality, and Desire*, New York: Oxford University Press.

Olson, J. (2004), 'Buck-Passing and The Wrong Kind of Reasons', *The Philosophical Quarterly*, 54(215): 295-300.

Parfit, D. (2001), 'Reasons and Rationality' in *Exploring Practical Philosophy*, Eds. Dan Egonsson, Jonas Josefsson, Björn Petersson, and Toni Rønnow-Rasmussen, Ashgate Press.

Rabinowicz W. and Rønnow-Rasmussen T. (2004), 'The Strike of The Demon. On Fitting Pro-Attitudes and Value', *Ethics* 114(3): 391-423.

Rabinowicz W. and Rønnow-Rasmussen, (2006), 'Buck-passing and the right kind of reason', *The Philosophical Quarterly*, 56(222): 114-20.

[30] I would like to thank Andrew Reisner for very helpful comments on an earlier draft of this paper. For very useful comments on drafts of predecessors of this paper, I would like to thank the audiences at the Pacific APA, 6 April, 2012, and SOPHA, Paris, May 16, 2012. I am especially grateful to Wlodek Rabinowicz, Ruth Chang, Jonas Olson, Graham Oddie, Michael Zimmerman, and Christine Tappolet.

Rabinowicz, W. (2008), 'Value Relations', *Theoria* 74(1):18-49.

Reisner, A. (2009), 'Abandoning the Buck Passing Analysis of Final Value', *Ethical Theory and Moral Practice* 12(4):379-395.

W. D. Ross, (1939), *Foundations of Ethics*, Oxford: Clarendon Press.

Rønnow-Rasmussen T. (2011), *Personal Values*, Oxford: Oxford University Press.

Rønnow-Rasmussen T. (2021), *The Value Gap*, Oxford: Oxford University Press.

Schroeder, M. (2010), 'Value and the right kind of reason', *Oxford Studies in Metaethics* 5: 25-55.

Skorupski, J. (2007), 'Buck-Passing about Goodness', in [online resource]: *Hommage á Wlodek: Philosophical Essays Dedicated to Wlodek Rabinowicz.* Lund: Lund University.Available at <www.fil.lu.se/hommageawlodek>. Eds. Josefsson, Petersson, and Rønnow-Rasmussen.

Sylvan, N. (2021), *Value, Fittingness, and Partiality: On the Partiality Problem for Fitting Attitude Analysis of Value*, PhD dissertation, Stockholm University Press.

Zimmerman, M. (2001), *The Nature of Intrinsic Value*, Lanham: Rowman & Littlefield Publishers.

Zimmerman, M. (2011), 'Partiality and intrinsic value', *Mind* 120(478): 447-483.

Williamson, T. (2000), *Knowledge and Its Limits*, Oxford: Oxford University Press.

An Account of Instrumental Value

Erik Carlson[1]

Abstract. In this paper, I tentatively suggest an account of how the instrumental value of a state of affairs derives from the intrinsic value of other states. According to this account, a state's instrumental value depends on how its outcome compares to the outcomes of its best and its worst alternative. Further, I briefly discuss similar accounts of personal instrumental value, and of harm and benefit.

1. Introduction

Some things are good or bad for their own sake. Such things have *intrinsic* value. Other things are good or bad because they lead to or prevent something that has intrinsic value. These things have *instrumental* value. In this paper I shall tentatively suggest an account of how a thing's instrumental value derives from the intrinsic value of other things. [2]

To make this task somewhat more tractable, I will make a number of simplifying assumptions. I will assume that the bearers of value are contingent states of affairs,

[1] I dedicate this paper to Toni Rønnow-Rasmussen, although I am afraid it lacks much of the philosophical subtlety and sophistication characteristic of Toni's work in value theory.

[2] My usage of the terms 'intrinsic value' and 'instrumental value' is to some extent stipulative. Some authors prefer 'final value', to denote value for a thing's own sake, and 'instrumental value' is sometimes used in both broader and narrower senses than mine. Often, a main distinction is drawn between intrinsic and *extrinsic* value. Some philosophers equate extrinsic value with instrumental value, whereas others regard extrinsic value as a broader category. There are many suggestions about how to sharpen and elaborate on these and related distinctions. Rønnow-Rasmussen (2002, 2015) and Zimmerman & Bradley (2019) contain excellent discussions and overviews of the literature.

which may be either atomic or conjunctive, and that a possible world is a maximal consistent conjunctive state. As concerns instrumental value in particular, it is often natural to regard events, including actions, as value bearers. To accommodate this possibility, I shall view events as a species of states of affairs. Alternatively, one could assume that instrumental value is borne by the state of affairs that a certain event occurs, rather than by the event itself.

Further, I will assume that intrinsic value can be measured on a real-valued ratio scale, such that the value of an intrinsically good (bad) state of affairs is represented by a positive (negative) number and the value of an intrinsically neutral state by zero. The intrinsic value of a conjunctive state of affairs *S* is, I will suppose, the sum of the basic intrinsic values of its atomic or conjunctive parts, including *S* itself.[3] Finally, I will assume that for any contingent state of affairs, there is a possible world that would be actual if this state were to obtain, and a possible world that would be actual if it were not to obtain.[4] Some of these assumptions may not be very realistic, but they allow us to avoid a number of difficulties that are not directly relevant to the main issues.

2. The Simple Account

It might be suggested that the instrumental value of a state of affairs is simply the intrinsic value there would be in the universe if the state were to obtain. Thus, let us start by considering the following account, letting W_S denote the possible world that would be actual if state of affairs *S* were to obtain:

> *The Simple Account*. The instrumental value of a state of affairs *S* is equal to the intrinsic value of W_S minus the intrinsic value of *S* (which may be zero).

Although appealingly simple, the Simple Account will not do. It implies that all states with the same intrinsic value that obtain in a given possible world have the same instrumental value in that world. (Note that in a world where states *S* and *S** both obtain, W_S and W_{S*} are identical.) Further, in an intrinsically good (bad) world all obtaining intrinsically neutral states are instrumentally good (bad). This is very implausible. Surely, states with equal intrinsic value may differ in instrumental value, and an intrinsically good or bad world may contain both instrumentally good and instrumentally bad states, which are intrinsically neutral.

These objections indicate that the connection between intrinsic and instrumental value posited by the Simple Account is too tenuous. The mere fact a certain

[3] Intuitively, a thing's basic intrinsic value is that part of its intrinsic value that does not derive from any of its proper parts. See, e.g., Feldman (2000) and Zimmerman (2001, chapter 5).

[4] The last assumption will be relaxed in section 6.

intrinsically good or bad state would obtain if a state S were to obtain is insufficient to confer positive or negative instrumental value on S. At the very least, what would be the case were S *not* to obtain also seems relevant as regards the instrumental value of S.

3. The Revised Simple Account

This suggests the following account:

> *The Revised Simple Account.* The instrumental value of a state of affairs S is equal to the intrinsic value of the conjunction of the states of affairs $S^* \neq S$, such that S^* would obtain if and only if S were to obtain.

By including the "only if" clause, this account avoids the most obvious flaws of the Simple Account. It allows that states with the same intrinsic value differ in instrumental value, and that intrinsically good (bad) worlds contain states that are intrinsically neutral and instrumentally bad (good).

However, the Revised Simple Account faces other serious problems. Suppose the world would have contained no intrinsically good or bad states of affairs had there not been life on Earth, and let S be the state that an asteroid hits the Earth early in its history, preventing life from ever evolving. Suppose also that the actual world is intrinsically very good. Intuitively, S is then instrumentally bad. According to the Revised Simple Account, however, it is instrumentally neutral.

Moreover, the Revised Simple Account also yields implausible results concerning the relative ranking of states of affairs in terms of instrumental value. Consider the following case:

> *Levers.* God offers you to pull one of three levers, labelled L_1 to L_3. You cannot refuse God's offer. Pulling a lever has no intrinsic value. If you pull L_i possible world W_i will be actual. W_1 and W_2 are very good worlds, containing many and exactly the same intrinsically good states of affairs, and no intrinsically bad ones. W_3 is not nearly as good, containing no intrinsically bad states, but only one intrinsically slightly good state. This state is not included in W_1 or W_2, and its intrinsic value is 1. Suppose also that you pull L_1, and that you would have pulled L_2, had you not pulled L_1.

In this case there is no obtaining intrinsically good or bad state that would not have obtained if you had not pulled L_1. Hence, the Revised Simple Account implies that the instrumental value of pulling L_1 is zero. The instrumental value of pulling L_3, on the other hand, is 1, since the only intrinsically good state in W_3 would obtain just in case you were to pull L_3. But the conclusion that pulling L_3 is instrumentally better than pulling L_1 is surely false.

4. The Counterfactual Comparative Account

The Revised Simple Account is insensitive to the fact that the asteroid's hitting the Earth, or your pulling L_3 in *Levers*, would *prevent* many intrinsically good states from obtaining. The remedy, it may be thought, is the following further revision:

> *The Re-Revised Simple Account.* The instrumental value of a state of affairs *S* is equal to the intrinsic value of the conjunction of the states $S^* \neq S$, such that S^* would obtain if and only if *S* were to obtain, minus the intrinsic value of the conjunction of the states S^{**}, such that S^{**} would obtain if and only if *S* were *not* to obtain.

This revision lets us take into account the intrinsically good or bad states that a state *S* prevents, when calculating the instrumental value of *S*. Thus revised, the account still implies that pulling L_1 in *Levers* has zero instrumental value. But the new revision implies that pulling L_3 has negative instrumental value. Relative to W_3, the nearest world where you do not pull L_3 is either W_1 or W_2. The Re-Revised Simple Account hence implies that the intrinsic value of W_1 or W_2 (which is the same) should be subtracted from the intrinsic value of W_3, which is 1, in order to arrive at the instrumental value of pulling L_3. Thus, the Re-Revised Simple Account yields the intuitively correct verdict that pulling L_1 is instrumentally better than pulling L_3.[5] (One might still object, of course, to the conclusion that pulling L_1 is instrumentally neutral, rather than instrumentally good.)

Given the assumption that the intrinsic value of a possible world is the sum of the basic intrinsic values of its parts, the Re-Revised Simple Account can be stated in a simpler way, letting $W_{\neg S}$ denote the possible world that would be actual were state *S* not to obtain:[6]

> *The Counterfactual Comparative Account.* The instrumental value of a state of affairs *S* is the difference between the intrinsic value of *Ws* and that of $W_{\neg S}$, minus the intrinsic value of *S*.

I have renamed the account in order to highlight its close similarity to the much-discussed Counterfactual Comparative Account in the literature on harm and personal value.[7]

I believe, however, that this account also faces fatal counterexamples. This is one:

---

[5] Like the Revised Simple Account, this account also avoids the above-mentioned problems for the Simple Account.

[6] To clarify, $W_{\neg S}$ is assumed to be the non-*S*-world that is nearest to W_S, rather than the non-*S*-world that is nearest to the actual world. These two worlds may be different, if the actual world is a non-*S*-world.

[7] See section 8.

> *Buttons.* God offers you to push one of four buttons, labelled B_1 to B_4. You cannot refuse God's offer. Pushing a button has no intrinsic value. If you push B_i possible world W_i will be actual. W_1 is an extremely good world, and W_2 is almost as good. W_3 is an extremely bad world, and W_4 is even worse. In the nearest possible world where you push B_2 it is true that if you were not to do so, you would push B_1. Further, in the nearest possible world where you push B_3 it is true that if you were not to do so, you would push B_4.[8]

The Counterfactual Comparative Account implies that pushing B_2 is instrumentally bad, while pushing B_3 is instrumentally good. This conjunction of claims is highly implausible in itself, and it has the even more implausible implication that pushing B_3 is instrumentally *better* than pushing B_2. This follows if we assume the principle, which I take to be a conceptual truth, that any good bearer of a certain kind of value is better, as regards this kind of value, than any bad bearer of the same kind of value.

5. Contextualism and Contrastivism

Ben Bradley has suggested a contextualist version of the Counterfactual Comparative Account.[9] On this account, different conversational contexts pick out different similarity relations between possible worlds.[10] It is hence context-dependent what the nearest non-*S*-world is, for a given state *S*. Therefore, Bradley's account does not imply, in *Buttons*, that pushing B_2 is instrumentally bad, or that pushing B_3 is instrumentally good *simpliciter*. Rather, pushing B_2 is instrumentally good relative to contexts where it is true that you would otherwise push B_3 or B_4, and instrumentally bad relative to contexts where it is true that you would otherwise push B_1. Similarly, pushing B_3 is instrumentally good relative to contexts where it is true that you would otherwise push B_4, and instrumentally bad relative to contexts where it is true that you would otherwise push B_1 or B_2.

This contextualist element does not save Bradley's account from trouble in *Buttons*. In the stipulated context, call it *C*, it is true in the nearest world where you push B_2 that you would otherwise push B_1, and also true in the nearest world where you push B_3 that you would otherwise push B_4. Hence, Bradley's account implies that pushing B_3 is instrumentally good and that pushing B_2 is instrumentally bad, relative to *C*. It follows that pushing B_3 is instrumentally better than pushing B_2,

[8] Essentially this example is given in Carlson (2020: 409), as part of an argument against the Counterfactual Comparative Account of harm and benefit. See also Carlson, Johansson & Risberg (2021, forthcoming).

[9] Bradley (1998). He intends his account to cover extrinsic value in general, considered as a broader category than instrumental value (see footnote 2). In his (2009: 50-52), Bradley proposes a similar account for personal extrinsic value.

[10] Bradley (1998: 116); cf. Bradley (2009: 50).

relative to *C*. But, it seems to me, pushing B_3 is not instrumentally better than pushing B_2 relative to *any* context.

Bradley might object that *C* is for some reason an unrealistic context. But this does not seem to be the case. To make the stipulated counterfactuals plausible, suppose, for instance, that you can reach B_1 and B_2 most easily with your left hand, while B_3 and B_4 are most easily reached with your right hand. Suppose you just pick a button, say B_2. (Maybe you are unaware of the effects of pushing the buttons.) Had you not pushed B_2, you would still have used your left hand and pushed B_1. Had you pushed B_3, on the other hand, it would have been true that if you had not done so, you would still have used your right hand and pushed B_4.

An idea in the vicinity of Bradley's contextualism is to formulate the Counterfactual Comparative Account as a contrastivist account.[11] According to such an account, a state's instrumental value is relativized to a relevant contrast state. Thus, a state *S* may be instrumentally good relative to state *S** (if W_S is intrinsically better than W_{S*}), but instrumentally bad relative to state *S*** (if W_{S**} is intrinsically better than W_S). Another way to express these contrastive evaluations is to say that it is instrumentally good that *S* obtains rather than *S**, but instrumentally bad that *S* obtains rather than *S***.

Applied to *Buttons*, this account avoids the implausible result that pushing B_2 is instrumentally bad and pushing B_3 is instrumentally good. Hence, we cannot draw the even more implausible conclusion that pushing B_3 is instrumentally better than pushing B_2. What the contrastive account implies is that pushing B_2 rather than B_1 is instrumentally bad, that pushing B_2 rather than B_3 or B_4 is instrumentally good, that pushing B_3 rather than B_1 or B_2 is instrumentally bad, and that pushing B_3 rather than B_4 is instrumentally good.

My main objection to this account is that it is too uninformative. Suppose we are asking whether pushing B_2 is instrumentally good or bad. The reply that pushing B_2 rather than B_3 or B_4 is instrumentally good, whereas pushing B_2 rather than B_1 is instrumentally bad, does not really seem to answer our question. A possible response to this objection would be to claim that for any state of affairs, there is only one relevant contrast state. This would preclude that a state is instrumentally good relative to one contrast state and instrumentally bad relative to another, but it would make the account even less informative. Given the counterfactuals stipulated in *Buttons*, the contrast state to pushing B_2 would have to be pushing B_1, and the contrast state to pushing B_3 would have to be pushing B_4. All we would be able to say about the instrumental value of pushing B_2, then, would be that pushing B_2 rather than B_1 is instrumentally bad. Similarly, all we could say about pushing B_3 would be that pushing B_3 rather than B_4 would be instrumentally good. No comparison could be made between the instrumental value of pushing B_2 and that of pushing B_3.

[11] Comments by an anonymous reviewer prompted me to discuss this possibility. Alastair Norcross (2005) has suggested a contrastive version of the Counterfactual Comparative Account of harm and benefit.

This seems unsatisfactory. (It might be suggested that if a state is instrumentally good relative to its contrast state, then it is instrumentally good *simpliciter*. But this move would take us back to the standard Counterfactual Comparative Account.)

6. The Midpoint Account

A potential lesson to draw from the failure of the Counterfactual Comparative Account is that the relevant comparison, for determining the instrumental value of a state *S*, is not what *would* be the case if *S* were not to obtain, but rather what *could* be the case. Thus, in *Buttons* it seems that we should compare the outcome of pushing a certain button with the respective outcomes of pushing the other buttons, and not just with that of not pushing the button in question.[12] More generally, we should compare a given state *S* to the states that are, in some sense, alternatives to *S*.

In order to capture this idea, let us assume that for any state *S*, there is a finite set of mutually exclusive states that contains *S* and its alternatives. Call such a set an *alternative-set*. The alternatives to *S* are the states that might obtain instead of *S*. (Somewhat more will be said about this assumption below.) Let A_S = {*S*, *S**, …, *S***} be the alternative-set to which *S* belongs, and let A_{WS} = {W_S, W_{S*}, …, W_{S**}} be the corresponding set of possible worlds. A straightforward suggestion is that the instrumental value of *S* is determined by comparing the intrinsic value of W_S to the intrinsic value of the best and the worst world in A_{WS}. Thus, add the intrinsic values of these two worlds, and divide this sum by 2.[13] Call the result the *midpoint* of A_{WS}. We can now consider:

> *The Midpoint Account.* The instrumental value of a state of affairs *S* is the difference between the intrinsic value of *Ws* and the midpoint of A_{WS}, minus the intrinsic value of *S*.[14]

[12] By the "outcome" of a state of affairs I mean the possible world that would be actual were the state to obtain.

[13] If two or more worlds are tied for best (worst) in A_{WS}, choose any of the best (worst) worlds.

[14] Why not instead choose the *average* intrinsic value of the worlds in A_{WS} as the baseline, and define the instrumental value of *S* as the difference between the intrinsic value of W_S and this average, minus the intrinsic value of *S*? A drawback of this account is that it makes instrumental value depend on the number of alternatives, in an arguably implausible way. Consider a situation in which states S_1 and S_2, which both have zero intrinsic value, are the only alternatives, and assume that the intrinsic values of W_{S1} and W_{S2} are 10 and −10, respectively. Choosing the average as the baseline yields the result that the instrumental values of S_1 and S_2 are, respectively, 10 and −10. Now suppose that the alternative-set is expanded with S_3, S_4 and S_5, and that the intrinsic values of W_{S3}, W_{S4} and W_{S5} are all −10. In this second situation, the instrumental values of S_1 and S_2 are 16 and −4, respectively. It seems, however, that the instrumental values of S_1 and S_2 should not vary, solely depending on whether S_3, S_4 and S_5 are included in the alternative-set.

This account yields plausible results in the cases we have discussed so far. In *Levers*, it implies that pulling L_1 and pulling L_2 are instrumentally good, while pulling L_3 is instrumentally bad. In *Buttons*, the implications are that pushing B_1 and pushing B_2 are instrumentally good, whereas pushing B_3 and pushing B_4 are instrumentally bad.

As compared to the Counterfactual Comparative Account, a further advantage of the Midpoint Account is that it does not require the questionable assumption that there is, for any state of affairs, a possible world that *would* be actual if this state were not to obtain. The set A_{WS} can be taken to include W_S and the set of worlds that *might* be actual, were *S* not to obtain. The alternatives to *S* are then the set of states that might obtain, instead of *S*, were *S* not to obtain. We need not assume that one of these states is such that it *would* obtain, in the absence of *S*. If *S* is an action, the alternatives to *S* are plausibly taken to be the other actions, incompatible with *S* and with each other, that are available to the agent in the situation. If *S* is an event but not an action, its alternatives might be the set of events, incompatible with *S* and with each other, whose occurrence at the same time and place is consistent with the past and the laws of nature of W_S.[15]

Concerning states of affairs other than events, it may often be unclear what states should be included in an alternative-set. Consider, for example, the state that Joe Biden is the present President of the United States. Who might have been President now instead of Biden? It is natural to include Donald Trump among the alternatives, and to exclude Abraham Lincoln. But what about Sarah Palin, say? Whether or not she should be included is arguably a context-dependent matter. We might want to consider only persons who actually ran for President in 2020, or we might be willing to consider a larger group of persons. It seems difficult to argue that one choice is objectively more correct than the other. The most feasible fully general version of the Midpoint Account may therefore be one that does not assign instrumental value to states of affairs *simpliciter*, but rather to states relative to an alternative-set, determined by a context of utterance. This allows for the possibility that a state is instrumentally good relative to one alternative-set and instrumentally bad relative to another.

7. Two Objections to the Midpoint Account

To be sure, the Midpoint Account is not unassailable. It has somewhat counterintuitive implications in cases like the following:

> *Knobs.* God offers you to turn one of three knobs, labelled K_1 to K_3. You cannot refuse God's offer. Turning a knob has no intrinsic value. If you turn K_i possible

[15] If physical determinism is true this condition has to be relaxed, in order to avoid the conclusion that no event has any alternatives.

> world W_i will be actual. W_1 is a very good world, having an intrinsic value of 60. W_2 is a very bad world, having an intrinsic value of −110. W_3, finally, is an extremely bad world, having an intrinsic value of −300.

The midpoint of $\{W_1, W_2, W_3\}$ is −120. Hence, the Midpoint Account implies that turning K_2 has an instrumental value of 10, thereby being instrumentally good. But it may seem that turning K_1 is the only instrumentally good alternative in *Knobs*, and that turning K_2 and turning K_3 are both instrumentally bad.

I think, however, that it is defensible to claim that turning K_2 is instrumentally good. After all, it prevents an extremely bad world from being actual. Of course, it also prevents a very good world from being actual. But since the difference in intrinsic value between W_2 and W_3 is greater than that between W_1 and W_2, the former, good aspect of turning K_2 arguably outweighs the latter, bad aspect.

In general terms, the Midpoint Account implies that no matter how bad the outcome of a state S is, and no matter how good alternative outcomes there are, S can be instrumentally good, provided that there is an alternative with an outcome bad enough to lower the midpoint below the intrinsic value of W_S. Conversely, a state with an extremely good outcome, and some extremely bad alternative outcomes, can still be instrumentally bad, if there is an alternative with an enormously good outcome that raises the midpoint high enough.

I am not sure that these implications are unacceptable. In any case, it is worth noting that the Counterfactual Comparative Account faces a similar problem. According to that account, too, a state S with an extremely bad (good) outcome can be instrumentally good (bad), if $W_{\neg S}$ is intrinsically even worse (better) than W_S.

Another objection to the Midpoint Account is that it fails to reflect the importance of *causation*, as regards instrumental value.[16] In one situation, let us suppose, actions a and b are your only alternatives. Both actions would cause a state of affairs S with intrinsic value 10 to obtain, and have no other intrinsically good or bad states in their outcomes. In another possible situation, actions c and d are your only alternatives. They would both cause a state S^* with intrinsic value −10 to obtain, and have no other intrinsically good or bad states in their outcomes. The Midpoint Account implies that a, b, c and d are all instrumentally neutral. But, the objection goes, a and b are in fact instrumentally good, since they would cause an intrinsically good outcome to obtain, and c and d are in fact instrumentally bad, since they would cause an intrinsically bad outcome to obtain.

This objection presupposes controversial claims about causation. Since S is unavoidable in the first situation, the assumption that a and b would each cause S to obtain seems difficult to square with theories of causation honouring the slogan that "causation is difference-making". And likewise regarding c, d and S^* in the second situation. But suppose, for the sake of argument, that the causal claims involved are consistent. Then my inclination is to conclude that causation is less relevant for

[16] This objection stems from comments by Olle Risberg.

instrumental value than one might think. If exactly the same intrinsically good or bad states of affairs will obtain whatever you do in a situation, I find it plausible to conclude that all your alternatives have neutral instrumental value. Whatever is true of causation, it would seem that instrumental goodness and badness require difference-making.

8. Personal Instrumental Value, Harm and Benefit

Several philosophers have proposed the Counterfactual Comparative Account as an account of *personal* instrumental value.[17] In our framework, this proposal can be put as follows:

> *The Counterfactual Comparative Account of personal instrumental value.* The instrumental value for a person *P* of a state of affairs *S* is the difference between the intrinsic value for *P* of *Ws* and that of $W_{\neg S}$, minus the intrinsic value for *P* of *S*.[18]

It is easy to see that this account is vulnerable to a variant of *Buttons*, in which pushing the buttons affects your, or someone else's, personal intrinsic value. As in the case of impersonal instrumental value, the Midpoint Account fares better (although the objections discussed in section 7 are relevant). Define the set A_{WS} as in section 6, and add the intrinsic values for *P* of the best and the worst world for *P* in A_{WS}. Let the *midpoint for P* of A_{WS} be this sum divided by 2. We can now state:

> *The Midpoint Account of personal instrumental value.* The instrumental value for a person *P* of a state of affairs *S* is the difference between the intrinsic value for *P* of *Ws* and the midpoint for *P* of A_{WS}, minus the intrinsic value for *P* of *S*.

As far as I can see, this account is equally plausible for personal as for impersonal instrumental value.

The Counterfactual Comparative Account is even more popular as an account of *harm* and *benefit*:

> *The Counterfactual Comparative Account of harm and benefit.* A state of affairs *S* harms (benefits) a person *P* if and only if the intrinsic value for *P* of *Ws* is lower (higher) than the intrinsic value for *P* of $W_{\neg S}$.[19]

[17] See Bradley (2009: 50); Feit (2016: 138f); Feldman (1991: 214f, 1992).

[18] Personal intrinsic value is often equated with welfare.

[19] For defences of this account, see, e.g., Boonin (2014); Bradley (2009); Jedenheim Edling (2021); Feit (2015, 2016, 2019); Klocksiem (2012, 2019); Parfit (1984: 69); Petersson (2018); Purshouse (2016); Timmerman (2019). Not all of these authors give an explicit account of benefit, but in most

One of several problems with this account is that it is vulnerable to variants of *Buttons*. If we take the value assumptions in that case to concern your personal intrinsic value, the Counterfactual Comparative Account implies that pushing B_2 would harm you, whereas pushing B_3 would benefit you. This runs afoul of a very plausible principle, stating that if *a* and *a** are alternative actions open to you in a situation, and doing *a* would benefit you while doing *a** would harm you, then you have a prudential reason to do *a* rather than *a**. In *Buttons*, there seems to be absolutely no reason for you to push B_3 rather than B_2. Moreover, the account also violates another very plausible principle, to the effect that if states *S* and *S** belong to the same alternative-set and the intrinsic value for *P* of W_S is much higher than that of W_{S*}, then *S* would harm *P* only if *S** would, and *S** would benefit *P* only if *S* would.[20]

Again, the Midpoint Account seems more promising:

> *The Midpoint Account of harm and benefit.* A state of affairs *S* harms (benefits) a person *P* if and only if the intrinsic value for *P* of *Ws* is lower (higher) than the midpoint for *P* of A_{WS}.

Assuming that it is your personal intrinsic value that is at stake in the cases we have considered, this account implies that pulling L_1 or L_2 would benefit you in *Levers*, while pulling L_3 would harm you. In *Buttons*, pushing B_1 or B_2 would benefit you, whereas pushing B_3 or B_4 would harm you. In *Knobs*, finally, turning K_1 or K_2 would benefit you, while turning K_3 would harm you. Of these results, the only one that is not intuitively quite plausible is that turning K_2 would benefit you. (Obviously, this is closely connected to the first objection discussed in section 7.)

I am, nevertheless, unsure whether the Midpoint Account is acceptable as an account of harm and benefit.[21] Its plausibility will largely depend on how well it handles variants of much-discussed difficulties for the Counterfactual Comparative Account; in particular the "preemption" and "failure to benefit" problems.[22] Pursuing these matters here would, however, take us too far afield.

---

cases it is clear that they take benefit to be analogous to harm. The Counterfactual Comparative Account is typically taken to be an account of *overall*, rather than *pro tanto*, and *extrinsic*, rather than *intrinsic*, harm and benefit. A state of affairs is intrinsically (extrinsically) harmful or beneficial to the extent that it is harmful or beneficial because of its intrinsic (extrinsic) properties.

[20] These criticisms are developed in Carlson (2019, 2020) and in Carlson, Johansson & Risberg (2021).

[21] A general argument against "well-being counterfactualist" accounts of harm and benefit, to which category the Midpoint Account belongs, is stated in Carlson, Johansson & Risberg (2021: 171-73).

[22] For a thorough discussion of the preemption problem, see Johansson & Risberg (2019). The failure to benefit problem is discussed in, e.g., Feit (2019); Purves (2019); Johansson & Risberg (2020); Klocksiem (2022).

9. Concluding Remarks

I have tentatively suggested the Midpoint Account as an account of impersonal and personal instrumental value, and also floated it as a possible account of harm and benefit. Even if these accounts should ultimately be rejected, there may be some weaker positive results to be salvaged. The central idea behind the suggested accounts is that the instrumental value of a state of affairs depends on how its outcome compares to those of alternative states, in terms of intrinsic value. If this idea is sound, we may at least have arrived at a partial account of instrumental *betterness*. According to this partial account, a state S is impersonally instrumentally better than an alternative state S^* if and only if W_S is intrinsically better than W_{S^*}. And analogously for personal instrumental value. To this partial account, the Midpoint Account adds a zero point or baseline, categorizing states as instrumentally good, bad or neutral, and allowing for comparisons of instrumental value across alternative-sets. Clearly, the partial betterness account may be correct also if the Midpoint Account mislocates the baseline. Similarly, even if the Midpoint Account of harm and benefit is wrong about the baseline separating beneficial states from harmful ones, it may nevertheless be true that a state S is less harmful or more beneficial than an alternative state S^*, for a person P, just in case W_S is intrinsically better for P than W_{S^*}. If so, we have at least obtained a partial account of the relation "less harmful or more beneficial than".

A possible and somewhat skeptical position is that these partial accounts of instrumental betterness and relative harmfulness are accurate, but that there is no general way to correctly locate the baseline. The factors relevant for determining the baseline may be different, or have different relative weights, for different alternative-sets.[23]

References

Boonin, David (2014) *The Non-Identity Problem and the Ethics of Future People*. New York: Oxford University Press.

Bradley, Ben (1998) "Extrinsic Value". *Philosophical Studies*, 91(2): 109-26.

Bradley, Ben (2009) *Well-Being and Death.* New York: Oxford University Press.

Bradley, Ben (2012) "Doing Away with Harm". *Philosophy and Phenomenological Research*, 85(2): 390-412.

[23] Jens Johansson, Olle Risberg, and two anonymous reviewers gave very helpful comments on earlier versions of this paper. I also wish to acknowledge financial support from Riksbankens Jubileumsfond, Grant P21-0462, and Vetenskapsrådet, Grant 2018-01361.

Carlson, Erik (2019) "More Problems for the Counterfactual Comparative Account of Harm and Benefit". *Ethical Theory and Moral Practice*, 22(4): 795-807.

Carlson, Erik (2020) "Reply to Klocksiem on the Counterfactual Comparative Account of Harm." *Ethical Theory and Moral Practice*, 23(2): 407-13.

Carlson, Erik, Jens Johansson & Olle Risberg (2021) "Well-Being Counterfactualist Accounts of Harm and Benefit". *Australasian Journal of Philosophy*, 99(1): 164-74.

Carlson, Erik, Jens Johansson & Olle Risberg (forthcoming) "Benefits Are Better than Harms: A Reply to Feit". *Australasian Journal of Philosophy*.

Feit, Neil (2015) "Plural Harm". *Philosophy and Phenomenological Research*, 90(2): 361–88.

Feit, Neil (2016) "Comparative Harm, Creation and Death". *Utilitas*, 28(2): 136–63.

Feit, Neil (2019) "Harming by Failing to Benefit". *Ethical Theory and Moral Practice*, 22(4): 809–23.

Feldman, Fred (1991) "Some Puzzles About the Evil of Death". *The Philosophical Review*, 100(2): 205–27.

Feldman, Fred (1992) *Confrontations with the Reaper*. New York: Oxford University Press.

Feldman, Fred. (2000) "Basic Intrinsic Value". *Philosophical Studies*, 99(3): 319-46.

Jedenheim Edling, Magnus (2022) "A New Principle of Plural Harm". *Philosophical Studies*, 179(6): 1853-72.

Johansson, Jens & Olle Risberg (2019) "The Preemption Problem". *Philosophical Studies*, 176(2): 351-65.

Johansson, Jens & Olle Risberg (2020) "Harming and Failing to Benefit: A Reply to Purves". *Philosophical Studies*, 177(6): 1539-48.

Klocksiem, Justin (2012) "A Defense of the Counterfactual Comparative Account of Harm". *American Philosophical Quarterly*, 49(4): 285–300.

Klocksiem, Justin (2019) "The Counterfactual Comparative Account of Harm and Reasons for Action and Preference: Reply to Carlson". *Ethical Theory and Moral Practice*, 22(3): 673-77.

Klocksiem, Justin (2022) "Harm, Failing to Benefit, and the Counterfactual Comparative Account". *Utilitas*, 34(4): 428-44.

Norcross, Alastair (2005) "Harming in Context". *Philosophical Studies*, 123(1/2): 149-73.

Parfit, Derek (1984) *Reasons and Persons*. Oxford: Oxford University Press.

Petersson, Björn (2018) "Over-Determined Harms and Harmless Pluralities". *Ethical Theory and Moral Practice*, 21(4): 841–50.

Purshouse, Craig (2016) "A Defence of the Counterfactual Account of Harm". *Bioethics*, 30(4): 251-59.

Purves, Duncan (2019) "Harming as Making Worse Off". *Philosophical Studies*, 176(10): 2629–56.

Rønnow-Rasmussen, Toni (2002) "Instrumental Values – Strong and Weak". *Ethical Theory and Moral Practice*, 5(1): 23-43.

Rønnow-Rasmussen, Toni (2015) "Intrinsic and Extrinsic Value" in I. Hirose & J. Olson (Eds.) *The Oxford Handbook of Value Theory* (29-43). New York: Oxford University Press.

Timmerman, Travis (2019) "A Dilemma for Epicureanism". *Philosophical Studies*, 176(1): 241–57.

Zimmerman, Michael J. (2001) *The Nature of Intrinsic Value*. Lanham, MD: Rowman and Littlefield.

Zimmerman, Michael J. & Ben Bradley (2019) "Intrinsic vs. Extrinsic Value" in Edward N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy*.

“I Owe You”

Accountability in Finance and Morality

Stephen Darwall

Recently, two books have appeared in which the words ‘accountability’ and its relatives, ‘accountable’, ‘account’, and the like, play prominent and perhaps unexpected roles. The books are Alva Noë’s *Infinite Baseball* and Robert Hockett and Aaron James’s *Money from Nothing* (Noë 2019; Hockett and James 2020). Both books argue that accountability, in something close to the moral sense, are essential to their respective subjects, baseball and finance respectively. This essay investigates Hockett and James’s claim.[1] How close is financial accountability to moral accountability?

Hockett and James are primarily concerned to understand the nature of money and its relation to credit and debt. A central goal is to argue that national debt is very different from household debt, and that we should be much less concerned with the absolute value of national debt than people often are. Hockett was one of the architects of the Green New Deal, and *Money from Nothing* argues that it and similar projects for the common good can be publicly financed even with substantially increased federal deficits and debt without unacceptable inflationary risks of the sort these are frequently thought to incur.

What grounds their case is a view about the nature of money, credit, and debt. Accountability language enters centrally here also, as it does in morality. “We humans seem to be natural accountants,” they say. “We hold each other and ourselves ‘accountable,’ keeping track of where things stand between us by our best bookkeeping” (26). Money is simply, they conjecture, “that thing, whatever it happens to be, that a community agrees to count as settling accounts between them” (27). To make this claim most plausible, we might understand it as restricted to

[1] Investigate the relation between baseball and moral accountability also in “‘It’s On You’: Accountability in Baseball, Finance, and Morality” (Darwall unpublished).

economic or financial debt, since that is the kind with which their economic arguments are concerned. But Hockett and James suggest that their general model of money, credit, and debt can be applied also to morality, or at least to the part of morality that is concerned with "what we owe to each other" (301).

An important strand of their argument is that debt and credit are "two sides of the same coin" (88). Here they quote A. Mitchell Innes:

> Credit is simply the correlative of debt. What A owes to B is A's debt to B and B's credit on A . . . The words 'credit' and 'debt' express a legal relationship between two parties, . . . the same legal relationship seen from two opposite sides (88).

It is obvious how to understand this in the economic case. A's financial debt to B is the very same legal-economic relation as B's credit with A. If A owes B ten dollars, then B has a ten-dollar credit with A, and *vice versa*. This has profound consequences, Hockett and James argue, when we consider national debt, at least when it is owed to a nation's own citizens (rather than to foreign creditors). Every increase in domestically held national debt is, by logical necessity, accompanied by an identical increase in the assets of citizens to whom it is owed. For this reason, Hockett and James recommend a rebranding of "national debt clocks" as "private wealth clocks" (217).

The bipolar structure of financial debts and credits, liabilities and assets recalls what Michael Thompson calls the "bipolar normativity" of directed or bipolar legal and moral obligations. When one person (A) owes a duty to another (B), then B has a correlative right against A, and *vice versa*. These "represent," Thompson says, "the same 'legal' or 'jural' relation from the different points of view of the legal persons caught up in it" (Thompson 2004: 370). This, of course, is just the familiar "correlativity" of (bipolar) duties and rights pointed out by Wesley Hohfeld almost a century ago (Hohfeld 1923: 40, as noted by Thompson 2004: 370; see also Darwall 2013).

Money, Hockett and James claim, is whatever a community "agrees to count as settling accounts." It is clear enough how to understand this in the case of financial debts and credits. Money is something like a general IOU that can be cashed not just with the person who owes one a debt but with anyone for anything they are willing to exchange. Hockett and James suggest, however, that something like it is also necessary when someone owes a moral obligation or debt to another. There must be something that plays the "account settling" role in the case of moral debts also. So, they conclude, "money is not so far from morality as it seems" (Hockett and James: 108).[2] It would seem to follow from their general definition, indeed, that whatever

[2] Taken by itself, this would seem to be consistent with financial accountability nonetheless being different from moral accountability in crucial respects, just not as different as it might seem. Nonetheless, I interpret Hockett and James as claiming that moral and financial accountability are formally identical. The context in which their remark takes place is an extended discussion of Nietzsche's talk of responsibility and debt from Essay 2 of *On the Genealogy of Morals* in which

can settle accounts in the case of moral accountability, or at least that can be publicly recognized as doing so, will count as a kind of money by virtue of that fact.

But how close is financial accounting and accountability to moral accountability, really? My aim in what follows is to consider the forms that accountability takes in these different arenas with a view to seeing what we can learn about the nature of moral accountability and morality. It will not matter either whether Hockett and James intend to *identify* financial accountability with moral accountability, either in whole or even in part. I shall understand them rather as pointing to something in their respective domains that is undeniably like moral accountability and that employs very similar language with it remaining an open question how close to moral accountability financial accountability actually is. That is the question I wish to investigate.

Finance: Credit and Debt as Voluntarily Assumed Fungible Assets and Liabilities

The classic nexus in finance is economic exchange and the acquiring of assets and liabilities through the extension of credit and the acquiring of debt. Suppose, for example, that A loans $10 to B. Since there is no "free lunch," neither A nor B is made any better off in financial terms by the loan pure and simple. A is now out $10, but has acquired a $10 credit with A that exactly balances this out. And B now has $10 more than previously, but has simultaneously acquired a $10 debt to A. A's and B's bottom lines are unchanged. Even so, A and B enter into the creditor/debtor relation voluntarily. It is normally assumed by both parties that, for whatever reasons, both prefer the *status quo post* to the *status quo ante*.

Since financial debt and credit are "two sides of the same coin," the extension of credit and the acquiring of debt not only exactly balance out the financial situation

---

Nietzsche says that "bad [moral] conscience" derives from "the oldest and most primitive relationship there is, . . . the relationship between buyer and seller, *creditor* and *debtor* (107). Hockett and James then note that Nietzsche ties the development of this more primitive notion of responsibility to the form it takes in morality under the influence of the "priestly caste" and the Christian idea of "a debt" that "we can never repay," owing to Jesus's giving his life for our sins. They reply that the idea of moral accountability can be secularized through ordinary social practices of settling accounts. They then describe some illustrative examples in which people keep track of what they owe one another through the doing of favors, making promises, accepting the benefits of cooperation, and so on. Their conclusion would seem to be that even in these moral cases, which are not explicitly financial, there will have to exist some socially recognized way of setting accounts—of what people owe to one another—and that this is what moral accountability must be (121-136). "Why," they ask, "must moral accounting be cosmic and theological?" (108). I will be arguing that we can agree with them that it need not presuppose anything cosmic or theological, but that it nonetheless must involve something essential *normative* that outruns any actual social practices, or even, for that matter, any actual supernatural theological facts.

of creditor and debtor *intra*-personally between *status quo ante* and *status quo post*, but also their financial situations *inter*-personally *ex post*. A's credit with B is exactly matched by B's debt to A: $10. Putting aside questions of interest, B can pay off their debt to A simply by giving A $10, thereby cancelling A's credit with B.

Financial debt and credit always involve a common *content* (in this case, $10) that is simultaneously owed by the debtor to the creditor and credited to the creditor with the debtor. The creditor/debtor relation is created by the voluntary transfer of the content from creditor to debtor and dissolved by its return (perhaps with interest) from debtor to creditor.

Much of Hockett and James's discussion is concerned with the nature of money and how even the $10 in this simple example is itself really a representation and medium of a whole nexus of credits and debts ultimately backed by claims on the national Treasury that can be relied upon to be honored because the federal government can require that its own claims for taxes, fines, and the like be paid in its own currency. This is all wonderfully interesting and insightful, but we can ignore it for our purposes, since we are concerned with how similar financial accountability is to moral accountability. For that purpose, we can consider simpler cases of the kind we have just imagined.

Now it is an important feature of loans of the kind we are considering that they are assumed by both parties to be entered into voluntarily. Moreover, a loan requires, like a promise, a commonly presupposed normative structure in the background. The making and accepting of a loan itself involves a kind of implicit agreement or promise. Debtor and creditor agree, explicitly or implicitly, to the loan along with terms of repayment. The debtor implicitly promises to repay and the creditor agrees to accept repayment on certain terms. Loans, like promises and agreements, therefore, require a background normative structure that both parties must represent themselves as accepting as common ground even to create credits and debts.

The powers to make and to accept loans are, like the powers to make and accept promises, "normative powers" (Raz 1972). Promises and loans can only come into existence through exchanges or transactions in which the parties reciprocally recognize each other's respective normative powers or authorities respectively to make and accept the promise or loan (see, e.g., Watson 2009, Darwall 2013d). The powers are called normative because their exercise affects the reciprocal obligations and rights that exercising them brings into existence. The credit is a right to repayment held against the creditor and the debt is an obligation to repay owed to the creditor.

This means that whenever there is financial credit or debt, say, of $10 (the common *content*) that is itself financial, the parties must represent themselves to one another as assuming in common that there are background obligations and rights that are normative or moral. Perhaps, like an insincere promiser, a lender or borrower might make or accept a loan without actually accepting the relevant normative moral obligations or rights. But they cannot intelligibly be understood as

making and accepting the loan without presenting themselves to one another as doing so. Otherwise, we have no distinction between loaning the money and simply transferring it to them without an expectation of repayment. It follows that although, when A makes a loan of $10 to B, the *content* of their respective credit and debt is entirely financial, A and B must represent themselves to one another as assuming obligations and rights that are normative or moral. In other words, alongside the purely financial debt content, there is the represented normative or moral form of the debt—its being *owed*.

I argue, moreover, that normative powers can themselves exist only if their exercise occurs against the background of reciprocal rights and obligations that do not result from their exercise (Darwall 2006: 200-203; 2013). We have already noted that in our imagined case A and B must represent themselves to one another as having the normative power to loan and borrow, respectively, and as exercising that power. That is needed to be able to distinguish between A's loaning $10 to B and A's simply *giving* it to B.

But we and, not least, A and B, need to be able to distinguish also between A's loaning $10 to B and B's simply *taking* $10 from A (without A's voluntary consent). It is part of the idea of a loan that what is loaned is taken with the voluntary agreement or consent of the lender. And that means that the terms on which A and B must represent themselves as relating are such that were B simply to take $10 from A without his consent, B would thereby wrong A. A and B must present themselves to one another, that is, as *already* having correlative normative or moral obligations and rights, independently of any they can acquire through the exercise of the normative powers to promise, lend, and acquire financial debt. It follows that financial loans, debts, and credits, require independent, commonly recognized moral obligations and rights in order to come into existence in the first place.

Something similar is assumed as part of the "common ground" in any voluntary economic exchange whatsoever, whether bartered or mediated by money. Both parties present themselves to one another as having the normative powers to offer and to receive in exchange what the other offers. But like the power to offer and receive a loan, the normative powers involved in any voluntary exchange cannot exist without there existing already a background of assumed obligations and rights that are independent of the powers exercised in the exchange. If I offer you my extra copy of *Money From Nothing* for your extra copy of Marx's *Economic and Philosophic Manuscripts*, then we represent ourselves to one another as assuming that it would be wrong for each of us simply to take what we are hoping to receive from the other in exchange.

Hockett and James are right, therefore, that there is a deep connection between economic or financial credit and debt and moral debt, at least directed or bipolar obligations that are owed to others. But the order of explanation is the reverse of what they might seem to suggest. We need the normative idea of moral debt (and, I argue, accountability) in order to understand financial debt in the first place. Financial debt is the content of a publicly presupposed moral debt.

Now since we cannot derive an 'ought' from an 'is,' we cannot conclude from the fact that a financial debt exists that it ought to be repaid as a fully normative matter of morality. The point is the same here as it is with other publicly presented putative normative structures like law. Legal positivists may be right that whether a system of law is in place as a social reality is independent of the truth of normative moral facts and that no moral obligation to obey the law can follow simply from law's existence as a matter of positive social fact. But even positivists generally agree that the law must present itself, or be represented socially, as genuinely obligating as a matter of public appearance (Green 2003). That is what distinguishes the social reality of law from the "gunman writ large" (Hart 1961: 7). Similarly, financial credit and debt involve a public presentation or appearance of normative moral debt. The latter is the assumed public social medium necessary to give any financial offer or exchange its financial content.

Now money, for Hockett and James, is "whatever a community agrees to count as settling accounts" (Hockett and James 2020: 27). In the financial case, this is straightforward enough. The content of the debt and credit resulting from a loan is itself expressed in monetary terms, and it is transferred instantaneously. When A loans $10 to B, the resulting financial transfer of $10 from A to B simultaneously creates A's credit with B and B's debt to A. Moving $10 from A's account to B's creates a $10 debt to A in B's account and a $10 credit with B in A's account. Credit and debt are the very same financial fact viewed from the poles of creditor (A) and debtor (B), respectively.

But what about exchanges that are neither financial nor simultaneous? Here things are more complicated. If we agree that I will give you my copy of *Money From Nothing* for your copy of Marx's *Economic and Philosophic Manuscripts*, then, even if we hand these to one another simultaneously, there is no common content that we have agreed to exchange. By agreement, we have exchanged one book for another, not financial debt for financial credit. Suppose, now, that although you give me your copy of Marx straightway, I wait to deliver my copy of Hockett and James. You and I now have an account that needs to be settled. You owe me nothing, and I owe you a specific book. The easiest way to settle the account, of course, is simply for me to keep our bargain by giving you the book. But suppose I do not. Then we need some way of settling my outstanding debt of a specific book to you and your credit of that same book against me, and neither the debt nor the credit is itself financial. Neither credit nor debt is of a kind that can necessarily be fully discharged or paid with currency.

Of course, perhaps it can. If you do not care which copy of Hockett and James you receive, perhaps we can settle accounts by my giving you a different copy or even by my giving you money to buy a new one yourself, if you do not mind. The fact that we live in a market economy where books are bought and sold may enable us to settle accounts, and money, as we ordinarily understand it, will play a significant role. The important point, however, is that the credit and debt that we have created by our agreement are not themselves financial in the way those created

by a financial loan are. They are, or at least are represented by us as being, a normative moral debt and credit (in the sense of a right of receipt), and how to settle accounts when the debt is not discharged is an irreducibly normative moral question.

Suppose that you do not want just any copy, but the very one I had been offering, and that, for whatever reason, I am no longer willing to give it to you. How are we to settle accounts? I might try to find something else I am willing to offer that you might be willing to receive in exchange for your credit of the book with me. Perhaps, although the copy of *Money From Nothing* you wanted was the very one I was offering, there is something else you might like as well or more. Money might play a role here, since I might not own that myself but be able to acquire it in exchange from someone else and then settle our accounts by offering it to you in place of what we had agreed I would give you before. Or you might accept an offer of money pure and simple, relying on something appropriate you might want to purchase coming along.

Even so, the credit with me that I created when we made our deal was, we were presupposing, a moral rather simply a financial credit. It was a right to receive the specific book copy I had offered. So the account we have to settle when I do not keep my part of the bargain is a moral rather than a financial account. Even financial loans involve, again, a presupposed normative moral structure, so that settling them is never *simply* financial. However, since the content of financial loans is itself financial, financial accounts can always be settled financially. Our case is different. But even though the credit you have with me is, we presuppose, irreducibly moral, we might nonetheless come to agreement in settling our account if there are things you would be willing to take in exchange that I would be willing to offer for your credit.

So long as all that is needed to settle our accounts is finding mutually agreeable "exchange value," as Marx would put it, it would seem that it can, in principle, be settled by money as we ordinarily understand it (Marx 1991: 139). Even if I own nothing you would take in exchange for my book, somebody might, and they might be willing to sell what they own in exchange for money I would be willing to offer you now.

"But wait a minute," you say. "Our deal was not that you would give me your copy of *Money From Nothing* or something of equal or greater exchange value." "You agreed to give me that specific book, and you thereby gave me a right against, or a credit with, you to receive it from you. And you add, "Not just a financial credit, but a *moral right*." So the real question is, "What are we going to do about the fact that you have violated that *right*?" "How can we settle this question of *right* by looking to a measure or medium market value?" By definition, that can only concern the good rather than the right. When you have been looking for something of comparable value, hoping that I might take it in exchange, you have been relying on an implicit principle of right, namely, that part of my right of receipt is to waive it voluntarily if I would prefer something else in its place.

To see the point even more vividly, suppose that without even any suggestion of a voluntary exchange, I simply *take* your copy of Marx. (Maybe, I mutter "property is theft" under my breath (Proudhon 2011).) In the alternative scenario we were imagining, we necessarily presented ourselves to one another as having ownership rights in our respective books that such a taking would violate. So now it has happened. I have taken your book, but, importantly, I have also *violated your right* to your book. And even if simply returning the book might settle the *content* aspect of our accounts once I have taken it, it does nothing to address the fact that I violated your right.

Again, money, as Hockett and James understand it, is whatever a community agrees to recognize as settling accounts. When accounts can be settled through voluntary exchange and all that is in question is exchange value, then all we require is a common currency. But so far that may do nothing to address any fundamental questions of right at issue. And even when it suffices to satisfy claims of right through voluntary agreement, it will do so because of assumed autonomy rights that license right holders to transfer or give up their rights through voluntary exchange that are in the background. We cannot do without some way of settling *moral* accounts.

When money as we ordinarily understand it does not suffice as a commonly recognized means for settling accounts, then what can? In societies that are governed by the rule of law, the obvious answer seems to be: procedures of civil and criminal law. If you and I signed a contract governing the exchange of our books, then you can take me to court to get satisfaction for your violated claim right. And if, without any agreement, one of us simply takes the book they want from the other, then the other can report the crime to the police and seek damages under tort law through the courts. What is at stake in either case is a matter of right that can ultimately be settled, not by any medium of exchange, but only by procedures of justice

Of course, legal procedures cannot completely settle the substantive questions of moral right and moral accountability that would be at issue. These are irreducibly normative questions of morality and not of law. But for issues of right like these, legal procedures are as close as we can come to publicly recognized ways of settling contested accounts of right. That is the very reason we have systems of law. It is arguably the case, moreover, as Kant maintained, that establishing common public law as a social reality is something morality itself requires.[3] Even so, issues of moral right and rights necessarily outstrip any socially constructed practices of law, however morally justified those practices might be.

[3] In the *Doctrine of Right* (Kant 1996).

Financial and Moral Accountability Compared

Considering financial accounts and the role of money in settling them throws into relief the fact that the question of how to settle accounts is always ultimately a moral one in which questions of moral obligation and right are inevitably at issue. And even when we establish collective practices and institutionalize the rule of law to publicly recognize these moral obligations and rights and hold ourselves accountable, as best we can, for complying with them, these cannot decisively resolve the normative questions of moral accountability that always remain in the background.

The very ideas of moral obligation and right are tied to moral accountability conceptually. What is morally obligatory is, as a conceptual matter, what we are accountable to one another *as representative persons* or members of the moral community for doing (Darwall 2006, 2013a). And moral obligations we owe to specific individuals, entailing rights these very individuals have against us, whether resulting from financial transactions, other voluntary exchanges, or even independently of voluntary exchange, are things we are accountable to them for as the specific individuals to whom we stand in these bipolar moral relations (Darwall 2013a). Here the authority they have to hold us accountable is not the representative authority that any person might have, but the *individual authority* of a specific individual related to us through bipolar normativity (Thomson 2004, Darwall 2013a, Wallace 2019).

Moral accountability and authority of either of these kinds is always ultimately mutual and reciprocal. I can be morally accountable to you as a representative person, if, and only if, you are reciprocally accountable to me. And you can be personally accountable to me as the specific individual related to you by bipolar normativity, as when we agree to a voluntary exchange, if, and only if, I am reciprocally accountable to you. The standing to enter into these relations of mutual accountability rests, moreover, on the "participant" agential capacities to which Strawson so influentially drew our attention.

The upshot is that moral accountability is always necessarily *relational*, not just in a logical sense that is common with financial accountability, or even in the topical sense of concerning our relations, including our financial transactions, with one another. Moral accountability is itself always *to* others who have the capacity and authority to relate and hold themselves accountable as fellow persons having these very capacities and authorities (Darwall 2006).

Financial debts are relational in a logical sense. Any credits you have must be against someone who thereby has a debt to you. So far, this need not involve anything second personal, as is shown increasingly in our globalized world in which not only does one not have to relate to someone to pay one's debts, one may have no idea of the identity of the institutions and persons to whom one ultimately owes them. But even financial accountability ultimately presupposes the possibility of

moral accountability; the latter is always assumed in the background. It is simply impossible, therefore, to reduce moral accountability to financial accountability; the latter, indeed, presupposes the former.

Finally, moral accountability and accounting must necessarily outstrip any social procedure, even indeed any morally justified social procedure for settling accounts. We can see this by reflecting on the conceptual connection between moral obligation and blame. It is a conceptual truth, I argue, that an action is morally obligatory if, and only if, it is an act of a kind that it would be blameworthy to fail to perform without excuse (Darwall 2006, 2013b, 2016). One might think, therefore, that moral accounts can be accomplished by social practices of blame. But that is not so. Suppose that I refuse to give you the copy of the book I had promised, and some appropriate fine is added to the social opprobrium of being made the object of society's blame. The problem remains that because moral accountability is necessarily a normative matter it cannot be determined even by any social currency of blame in addition to the financial currency, or money that is at issue with financial debt and credit.[4]

References

Darwall, Stephen (2006). *The Second-Person Standpoint: Morality, Respect, and Accountability*. Cambridge, MA: Harvard University Press.

Darwall, Stephen (2013a). "Bipolar Obligation," in Darwall 2013b.

Darwall, Stephen (2013b). *Morality, Authority, and Law: Essays in Second-Personal Ethics I*. Oxford: Oxford University Press.

Darwall, Stephen (2013c). *Honor, History, and Relationship: Essays in Second-Personal Ethics II.* Oxford: Oxford University Press.

Darwall, Stephen (2013d). "De Mystifying Promises," in Darwall 2013c.

Darwall, Stephen (2016). "Making the 'Hard' Problem of Moral Normativity Easier," in *Weighing Reasons*, eds., Errol Lord and Barry Maguire. Oxford: Oxford University Press.

Darwall, Stephen (2019). "What Are Moral Reasons?," The 2017 Amherst Lecture in Philosophy (http://www.amherstlecture.org/darwall2017/index.html).

Darwall, Stephen (unpublished). "'It's On You': Accountability in Baseball, Finance, and Morality."

Hart, H. L. A. (1961). *The Concept of Law*. Oxford: Clarendon Press.

Hockett, Robert and Aaron James (2020). *Money From Nothing: Or, Why We Should Stop Worrying About Debt and Learn To Love The Federal Reserve*. Brooklyn, NY: Melville House Publishing.

[4] I am indebted to a referee for pressing me to clarify this.

Hohfeld, Wesley Newcomb (1923). *Fundamental Legal Conceptions*, ed. Walter Wheeler Cook. New Haven, CT: Yale University Press.

Kant, Immanuel (1996). *Practical Philosophy*, trans. and ed. Mary J. Gregor. Cambridge: Cambridge University Press.

Marx, Karl (1991). *Capital, Volume One: A Critique of Political Economy*. London: Penguin Books.

McKenna, Michael (2018). *Conversation and Responsibility.* Oxford: Oxford University Press.

Proudhon, Pierre-Joseph (2011). *Property is Theft: A Pierre-Joseph Proudhon Reader*, ed. Iain McKay. Edinburgh, Scotland: AK Press.

Raz, Joseph (1972). "Voluntary Obligations and Normative Powers," *Proceedings of the Aristotelian Society*, 46: 79-101.

Strawson, P. F. (1968). "Freedom and Resentment," in *Studies in the Philosophy of Thought and Action*. London: Oxford University Press.

Thompson, Michael (2004). "What Is It To Wrong Someone?: A :Puzzle About Justice," In *Reason and Value: Themes from the Philosophy of Joseph Raz*, eds., R. Jay Wallace, Philip Pettit, Samuel Scheffler, Michael Smith. Oxford: Oxford University Press.

Wallace, R. Jay (1994). *Responsibility and the Moral Sentiments.* Cambridge, MA: Harvard University Press.

Wallace, R. Jay (2019). *The Moral Nexus*. Princeton, NJ: Princeton University Press.

Watson, Gary (1987). "Responsibility and the Limits of Evil: Variations on a Strawsonian Theme," in *Responsibility, Character, and the Emotions: New Essays in Moral Psychology*, ed. F. D. Schoeman. Cambridge: Cambridge University Press.

Watson, Gary (2009). "Promises, Reasons, and Normative Powers." In *Reasons for Action*, eds., David Sobel and Steven Wall. Cambridge: Cambridge University Press, Pp. 155-178.

Theodicy as Axiology and More

Seyyed Mohsen Eslami

Abstract. The literature on the problem of evil does not draw enough upon the relevant debates in (meta)ethics, and ethical theorists (broadly understood) can engage with the problem of evil as a way of inquiry in their field. I review how the problem of evil is essentially formed based on (evaluative and deontic) ethical judgments, and how responses to it, either theistic or atheistic, are mainly based on the relevant ethical judgments. Meanwhile, though contemporary debates in metaphysics and epistemology have influenced the literature on the problem of evil, the same does not hold true for ethics. This suggests that there are ways to engage with the problem of evil as doing axiology or ethical theory more generally, which may be fruitful regardless of their being theodicy. I end by briefly discussing an example focused on the idea of moral progress.

Introduction

The problem of evil is both an important theoretical question and a fundamental human concern that is dealt with as a philosophical problem mainly in the philosophy of religion.[1] To begin with, it should not be a surprise if the problem of evil is not, in a sense, essentially a question of "philosophy of religion". Philosophy

[1] There *are* problems around evil that have been a concern of moral philosophers. For example, Socrates' claim about impossibility of knowingly choosing to do wrong which is now discussed as the guise of the good, or the debate among modern philosophers about the nature of human beings and whether it is primarily good or bad. All this is much expected considering the subject matter of the discipline.

of religion is where (so-called) core philosophical questions are raised concerning religion. Therefore, questions of the existence of God or the rationality of faith are at bottom metaphysical and epistemological questions.

What about the problem of evil? The rich literature on the subject emphasizes the important metaphysical and epistemological questions involved in the problem of evil, including how to understand divine omnipotence, free will, possible worlds as well as probabilities of knowing God's reasons, and so on. True as it is, the role ethics plays in this area is not as one might expect. For one thing, it is not hard to find great contemporary metaphysicians and epistemologists engaged with such issues, though this is not exactly the same with regard to ethics, broadly understood.

In the following, I briefly review the ways the problem of evil is formed and dealt with mainly ethically. The aim is not to claim that the current literature on the subject does its work without ethics. But the worry is rather whether the role ethics plays is acknowledged consciously, and whether the relevant ethical literature is used properly and effectively.

Formulating the Problem of Evil

The problem of evil is primarily a moral problem in its formation. This is how we come to the problem of evil. On the one hand, there is God. Traditionally, God is omniscient, omnipotent, and benevolent. According to the last one, *i.e.* benevolence, God is expected to be good and do right, normally understood as to help people, to prevent people from suffering, and the like. On the other hand, there is the world. The world as we see and experience it includes evils. Bad things happen in the world, and that is what is meant by "evil" (van Inwagen, 2006). And, thus, there is the ordinary question "why does God allow us to suffer?" This is the question that leads to the different versions of evil, either focused on gratuitous evil or the amount, kind, or distribution of evils (see: Trakakis, 2007), formulated in different ways.[2]

It is clear how the framework in which these questions are formed is ethical. By "ethical" I mean any sort of deontic and evaluative judgment which are not that of other normative domains (say, legal, aesthetic, or else.). Deontic judgments are claims about which category of forbidden, required, or permissible (or else) an action belongs to. On the other hand, evaluative judgments mainly focus on matters of value, whether something is good, bad, or neutral.[3]

[2] Here the problem of evil is mostly presented in terms of consequentialism, and the deontological considerations are mentioned along the way. This is partly a narrative choice, not meant to accurately formulate the problem or represent the literature.

[3] The same is true about other categories of the normative, such as characterological (Miller, 2011) or fittingness (Berker, MS).

On the one hand, regarding God, it seems that it is assumed that God is an agent, able to be good or bad and do right or wrong. These are in part metaphysical assumptions, with ethical implications, such as what is required for an agent to be a morally good one. Furthermore, it is assumed that not only God is good, but God's goodness also requires God to treat us in a certain way. Either God has obligations toward us and should respect our rights, or considering God is morally perfect God should do as best as God can, say, as a matter of supererogation. Be that as it may, these are all deontic judgments.

Regarding the world, first, there is the idea of what is good and what is bad, or evil. It is assumed that, for example, innocent people being killed in wars or natural disasters killing lots of people is bad. Furthermore, there are claims about the world as a whole. A world without war, cancer, and earthquakes would be better than the one we live in. These two groups of claims about God and the world are enough for some versions of the problem of evil.

The problem of evil might be focused on why there are evils that seem to be pointless, say, not having any positive role. For example, perhaps one thinks pain is something bad, or even has negative value, and that the world would be better without it. But it may turn out that pain brings about good consequences, and therefore leads to something of value. In that case, we ought to stop complaining about why there is pain, indeed we should be thankful. Sure, to this end, we need some explanation, such as whether this makes pain itself valuable in some sense, or that it has a specific relation with the good consequences. Certainly merely leading to a good outcome is not sufficient for justification. More on this follows in the discussion of the greater good theodicies.

In the same vein, we wonder why there are evils, such as innocent kids getting killed in wars, looking for explanations. Furthermore, we can imagine that what ignites the problem of evil to be different claims: about the very existence of evil, some kinds of evils, or the huge amount of evils – say, many kids getting killed in many different wars at different times and places.

For example, concerning the very existence of pointless evil, van Inwagen argues that we might be able to tolerate some gratuitous evil. In response, others have attempted to argue that the problem of evil challenges theism even if we allow some gratuitous evils (Russell, 2017). Some other arguments from evil maintain that there are "intense sufferings" in this world that are pointless (Rowe, 1979). Children being tortured, raped, and killed for example. These are not only bad things but too bad and horrible, in terms of quantity as well as quality (Adams and Sutherland, 1989). Perhaps they also lead to some good consequences, though the amount of evil may still be weightier than the good outcomes, and a specific relation between them is required. (More on this later.)

Be that as it may, in all of these judgments, not only are some things evaluated as good and bad, but there are also claims about value comparisons ("what good outcomes would be weightier that the evil?") and actions ("what God has good reason to do, to permit, or to avoid?").

One might think that this claim about the ethical nature of the problem of evil has something to do with a familiar distinction between "moral evils" and natural ones (for a discussion of the distinction, see: Trakakis, 2007). This is not the case. Evil in the problem of evil can include all bad things. The distinction between natural and moral evils is due to the source of evil, or what has led to the evil – whether it includes some intention raised from an agent who enjoys some kind of freedom. In either case, the result is some evil. And judging the result to be evil is the moral claim which is central to the problem of evil. But the distinction between moral and natural evils works at another level, dealing with specific proposals in response to the problem of evil, such as the free will defense (Plantinga, [1974] 2002). In that context, we can ask whether the theodicy in question can deal with both kinds of evils.

Another worry might be whether all versions of the problem of evil are ethical, *i.e.*, inquiring about them could count as ethical inquiry. Here is one classification. The problem of evil can be of these sorts: (a) theoretical, (b) practical, and (c) existential. The theoretical problem of evil concerns the consistency of the (belief in) existence of God and the existence of evils in the world. The questions are about the possibility of consistent acceptance of both, what epistemic stance we should take in this regard, and what implications it has for what we believe (and, consequently, do). This is primarily a theoretical problem. Atheists can also find the question theoretically interesting: if God with such and such attributes did exist, would it be possible that this world, with its evils, was God's creation?

As illustrated, the theoretical problem of evil, which is the more familiar form, is ethical. Note that the theoretical problem of evil can include non-theological versions. The problem of evil can arise from within a non-theological worldview. This also suggests that the problem is ethical rather than theological. Imagine that there is no God (as understood in the Abrahamic religions). Still, the fact that there are evils combined with some judgments about the world leads to the problem of evil.

For example, consider the belief that the world is governed by some sort of "moral order" – that doing good will return to you, "the world" does justice to everyone, and the like. Another option is the belief that the world is getting better and better. (cf. Nagasawa, 2018 on "existential optimism"). To get to the problem of evil we need two classes of moral judgments: judgments about the world that set our expectations of it, and judgments about the things we find in the world. But no discussion of God (as is common to the debate in the literature) is necessary. This being said, in the following "the problem of evil" is the familiar debate, usually involving God with the aforementioned attributes.

Aside from the theoretical problem of evil, there is the practical problem of evil. In this case, God is not at the center of the debate. Here, the challenge is that there are evils in the world - What can or should we do about it? Consider theists who believe that there is no inconsistency between admitting the existence of evils in the world and the existence of God and the rationality of belief in God. Still, the

practical problem of evil is there to be dealt with. In a sense, a considerable part of moral philosophy is concerned with this practical problem.

Obviously, the practical problem of evil is ethical, though this is not what philosophers of religion are typically concerned with. Again, the evaluation that the world (that it should not be as it is, and that it should be as it is not) is ethical. Furthermore, in this case, there is the "practical" element – that something should be done. This requires both a description of an ideal (or a better) situation to work toward it and a discussion of how this should be done on our part. Morally speaking, not all means are justified to reach the ideal (or better) situation. There is no way to deal with the practical problem of evil without getting into ethical inquiry, though many non-ethical (and non-philosophical) inquiries are needed too.

What about the existential problem of evil? The existential problem of evil refers to the phenomenon that people lose faith or go through a huge and deep change after confronting some evil (for a review, see: Peterson, 1998, ch. 7). Note that the theoretical solutions to the problem of evil will not be necessarily effective here, since in this case psychological aspects are also involved. Yet, this might be itself an instance of evil to deal with, the evil being people being hurt in this way and being unable to deal with the suffering. Similarly, this counts as evil for the theist if people lose faith without responding properly to the relevant considerations.

Furthermore, there are other ethical aspects to the existential problem of evil. Although the main issue might seem psychological, one can argue that the psychological reaction is (at least partly) grounded on the person's moral perceptions. A normal person facing a brutal killing would not merely come to the belief that something morally wrong is happening but would be, say, horrified and angered. Of course, the belief underlying the emotional reaction need not always be explicit. Still, it is plausible to think that people's emotional reactions, though psychological, are affected by their deep-rooted moral views.

It is noteworthy that it is rather obvious that ethical judgments play a central role in the formation and formulation of the problem of evil. The point to be emphasized is that there are relevant debates in the context of (meta)ethics, not properly reflected in the literature on the subject in the philosophy of religion. There seems to be a gap to be bridged.

Reactions to the Problem of Evil

The problem of evil (henceforth meaning "the theoretical problem of evil", unless otherwise stated) has invited different responses. Roughly speaking, the responses are of two kinds: theistic and atheistic. I take theistic responses to include all reactions to the effect that the problem of evil does not threaten theism. On the other hand, atheistic responses include all responses to the effect that the problem

provides a challenge for theism. Though, in principle, it seems possible for a theist to accept some atheistic response.

Theistic Approaches

No doubt, some responses philosophers have adopted to deal with the problem are mainly metaphysical or epistemological. For example, consider an approach that relies only (or at least mainly) on the idea of possible worlds. One might say that although it might *seem* that a better world could be created instead of this one, it is not the case, and since the actual world is the best possible world. Defending that this world is the best possible world and that no better world is possible might be a theistic response to the problem of evil. A classic example is Leibnitz (Murray and Greenberg, 2016). Such maneuvers are metaphysical.

Similarly, there are epistemological approaches. One might resist the conclusion of the problem of evil, insisting that we do not know enough and we don't have enough epistemic capabilities to know the premises of the argument, and thus cannot acknowledge its conclusion. Some forms of skeptical theism are examples. These approaches suggest that perhaps not all responses to the problem of evil are mainly based on normative judgments. Although, it may turn out that the development of such responses would, in the end, need normative judgments.

Be that as it may, it seems that the majority of the theistic reactions to the problem of evil are ethical. Meanwhile, note that even the aforementioned examples of metaphysical and epistemological approaches will require some ethical discussions, or would benefit from them. Consider the first one, the best possible world response. Perhaps this response can resist the conclusion of a simple version of the problem of evil – why didn't God create a better world? However, a further question will arise –why did this world have to be created, at all? This can be not merely an inquiry out of curiosity, but a challenge, implicitly suggesting that if this is the best possible world, and that the best possible world includes such and such evils, no world at all would be morally preferable. Or, at least, a moral agent would not allow this best possible world, considering the evils it contains. The best possible world response is not enough on its own, and some elements that it can benefit from are ethical.

The same is true about the skeptical theistic approach. The skeptical theist needs to explain where and why we have epistemic limitations. For example, one might appeal to the ethical aspects of the problem to emphasize our limited knowledge – how can we know about the value comparisons, or how can we know the moral obligations of an agent such as God? Again, some maneuvers of ethical nature would be helpful to further such approaches.[4]

[4] I have put aside the point that all such judgments (say, that if we do have epistemic limitations then we should not rely on our limited knowledge and conclude things about God) might be understood as deontic judgments.

While not all responses to the problem of evil are merely or mainly ethical, the majority of responses to the problem of evil are so. This being said, such attempts and debates about them may be considered inquiries in ethical theory, broadly construed. In many cases, proponents of theistic responses to the problem of evil propose an ethical claim regarding either God or the world, which is supposed to eliminate the alleged conflict between the two. Here are some examples.

One might argue that (i) God is not *morally* good. The source of the problem, according to these philosophers, is that people have understood God as an agent like us, therefore expecting God to be caring or loving. God's goodness, if saved, is interpreted as metaphysical goodness. This is understood in different ways, including perfection. But what would justify giving up the common conception of God as morally good? This line of thought might be motivated by ideas about the source or nature of ethical norms.

For example, if God is just another (but maximally perfect) agent, and morality is a matter of rational agents, then moral principles apply to God. In that case, it is not easy to claim that God is not morally good. However, other views about the nature and source of morality may open other possibilities. Perhaps endorsing other views that ground morality on facts about human beings (which may be developed in various realist or non-realist forms), then there is room for God not to be morally good.[5]

This brings us to a more familiar approach: divine command theory. If what is good and what should be done is grounded on God's will, then the arguments from evil must be problematic. There is no way for God to go wrong. Compare this with some form of personal subjectivism to the effect that one should do what one wants to do at the moment. In that case, with some conceptions of "want", one hardly can go wrong, since one always does what one wants at the moment. Similarly, assuming some forms of divine command theory, there is no worry about the morality of God's actions.[6] Of course, it is not our concern here whether such an approach has any plausibility.

Another route to take is that (ii) God is morally good, but is not omnipotent. At face value, if the theistic response to the problem of evil is based on the limitations of God's power, it might seem to be metaphysical. However, it is also important how the limit on God's power is explained. One line of argument for limiting God's power is insisting on God's goodness. Theologically speaking, this sounds better than sheer limitation, which counts as a weakness. For example, one can propose a

[5] To defend an atheistic response, Maizen (2017) argues that this line of thought fails. In fact, it doesn't seem easy for theists to let go of the idea of God's moral goodness.

[6] It is noteworthy that a complete defence of this position requires an explanation about our intuitive ethical judgments that have *led* to the problem of evil. That is, it seems that there are evils in the world. It seems that according to this position, those are not in fact "evils", since they are somehow willed by God. Now, we need to know more about our intuitive ethical judgments – are they trustworthy at all, how we learn about morality, etc. In a similar line of thought, some have argued that some, many, or all of the theistic responses lead to scepticism about ethics. For example, see: Maitzen, 2009.

moral principle that limits God's world: interfering with the events of the world violates the autonomy of adult human beings or our valuable freedom, which is not permissible. Technically, this might not be a limitation on the side of power, indicating weakness. Be that as it may, these kinds of responses are also rooted in ethical ideas and invite further ethical inquiry.

Yet another line of thought is that (iii) God does not have any obligation to us, at least not the obligations which the arguments from evil assume. Obligations have grounds. I cannot blame people for not helping me if they do not have any obligation to help me. Similarly, to expect God to help us (or to not allow evil and the like), one needs to explain why exactly God is obligated to help us. Some have argued that God is morally good and omnipotent, yet there is no implication that God should, say, stop the evils or avoid creating a world with evils. Such an approach can be developed in different ways and directions. Recent examples include challenging the idea of the "perfect love" of God toward human beings (Rea, 2018, ch. 5) and God's specific relation to the world (Mooney, 2022).

An objection might be raised to the effect that if God is morally perfect, there is no need for any obligation. A perfect being should go beyond the call of duty, and do not only what is required, but what is supererogatory. Considering that the idea of supererogation is controversial among moral philosophers, evaluating these suggestions depends on the relevant debates.

Similar attempts have been done on the other side of the problem – regarding the world. First, one can go into detail about each instance of evil. For example, if the problem of evil is based on the badness of pain, an assessment of the badness of pain is helpful. It might turn out that the pain is not bad. Considering pain's evolutionary role and its survival value, it is arguably good for us. However, this line of defense has its limits. Perhaps some proposed evils are not ultimately evils. It doesn't sound promising to apply the same strategy to horrible particular cases of, say, rape and murder.

A rather relevant approach is appealing to the privation theory of evil. Traditionally, one way to deal with the problem of evil has been to propose a theory of value, to see if the alleged evils are in fact evils. For example, the privation theory is mainly about the nature of evil, claiming that it is a form of absence. However, it seems that the main motivation and application for the theory have been in the context of the problem of evil. If the alleged evils are mere privations, according to the proponents of this approach, they might not need much explanation. Evaluation of this theory of value, both its formulation and its relation with other positions in axiology, is an ethical inquiry.

Second, one popular approach is to develop a greater good theodicy (Langtry, 1998).[7] Philosophers have looked for valuable things in the world which are

[7] Here I use the theodicy in the broad sense, meaning any theistic response to the problem of evil. It includes both theodicy and defence, as the distinction is not relevant here – and it might be unimportant basically. See: van Inwagen, 2006.

dependent on the alleged evils, and greater than them. Some of the most famous theodicies are in this group. Soul-making theodicy of John Hick is an example. According to a famous formulation of the problem of evil by William Rowe (1979), the challenge is to find a "greater good" for which the evils exist. Soul-making is the good we get, according to Hick ([1966] 2010, III and IV). Similar approaches are proposed based on the value of sympathy, care, etc.

In the same vein, the free will defense could be understood in this way. It seems that we should understand Plantinga's free will defense (1977) as proposing a *greater good*. The idea is that a world with freedom would be better than a world without freedom, even if the former has more evils in it. (Then, it is argued that the valuable freedom has features, limiting what is possible for God to do.) Evils are bad, but freedom is worth it. As Plantinga writes, "[t]he Free Will Defense can be looked upon as an effort to show that there may be a very different kind of good that God can't bring about without permitting evil" (Plantinga, [1977] 2002).

For these theodicies to work, three conditions need to be met. First, the theodicist needs to pick a value. Next, the evils in question (or God permitting them) must be necessary for that value to obtain. Finally, the value must be greater than the evils. A greater good theodicy might be challenged in each part. For example, imagine if "coming back to God" is introduced to be the greater good. Not everyone *agrees* that such things are valuable to help make sense of the evils in the world. In this regard, values that are *not* dependent on theological assumptions have a better chance. Similarly, a greater good theodicy that introduces "freedom" as its central value might be challenged concerning the second condition. It is not easy to show the necessary relation between all (kinds or instances) of the evils (or God permitting them) and that value. Finally, there is the worry about being a "greater" value. Consider the greater theodicy which finds soul-making to be the greater good. One can easily acknowledge that soul-making is valuable and even if one accepts that there is some necessary relation between evils and soul-making, the question remains whether, ultimately, this outweighs the evils.

Atheistic Approaches

Atheistic approaches, insisting on the *problem* of evil for theism, are also mainly ethical. First, they are direct challenges to the abovementioned claims in the theistic responses, especially regarding the world, such as the question of whether the evils of the world are in fact evil, or what can or cannot outweigh such evils. Furthermore, there can be more explicit and direct moral attacks on theodicies, such as on the greater good theodicies.

One challenge against theistic responses to the problem of evil is whether it is morally permissible to prefer some greater good, even if it requires the suffering of innocent human beings. This is mainly the same question mentioned before, *i.e.*, even if this is the best possible world, is it necessarily better than no world at all?

That is, if the moral principles applying to God are similar to ours, it is not clear that one can allow evils affecting innocent people merely to achieve a greater good.[8]

This being said, anyone who defends some atheistic response is, first, defending some ethical judgments involved in the very problem of evil, and, second, defending some ethical judgments to show that theistic attempts to resist the problem of evil fail.

So far, some standard lines of argument in the debate are overviewed. However, there are additional worries about theodicies, shaping a growing area in the recent literature. Some philosophers have raised worries against theodicies in general, instead of focusing on this or that proposal or its specific argument.

One main challenge is the ethical implications of theodicies. For example, it is argued that some or all of the theodicies have the potential to lead to moral skepticism. These theodicies try to convince us that what seems to be evil is not actually evil; overall, it is good as it brings about a great amount of value. Therefore, accepting such theodicies makes us suspicious of our moral intuition in general (see: footnote 6).

This can be done in two different ways. First, according to the argument, the theodicy can lead to skepticism as long as our first-order intuitions play a central role in our understanding of the right and the good. In that case, the very problem of evil may resolve, as there are not the required normative ingredients to have the problem in the first place.[9] Second, this may be understood as an answer to the problem of evil by rejecting its normative premises. This is compatible with having some normative judgments, very different from commonsense morality, say, thinking that the source of underrating right and wrong is some specific religious manuscript or procedure. Either way, intuitions of commonsense morality are undermined, which is not a welcomed consequence.

Another ethical challenge about the theodicy is that it weakens people's moral motivations. If suffering helps to improve sympathy or soul-making, why should one try to eliminate it? Or, more worrisome, why should one avoid making people suffer? These are sides of the same question: if God exists and God has knowledge, power, and benevolence, why does not God intervene and help people?

In response, one might suggest that God does not do this so that we can do it and gain the good. This is not as helpful as it might seem. For this proposal to work, one needs to explain when it is morally permissible for an agent (God) not to act and let us act, even when there is not much chance that we can do it. Consider a parallel

[8] Elsewhere (Eslami and Saeedimehr, 2021) I have discussed the moral objection to greater good theodicies. To argue against theodicies based on this argument, one needs to defend the idea that the moral principles appealed to (say, do not harm, or do not harm to bring about greater good) are absolute. However, it seems that such principles are not exceptionless. Meanwhile, this is not enough for a theodicist. Even if we agree that humans (limited in all sorts of aspects) may be permitted to violate the moral principles in question, it is not automatically clear if the evils in question can also be exceptions and why the same route is open to God.

[9] Thanks to Jakob Werkmäster for this point.

case by Maitzen (2017). A kid is drowning. I am there and I can help. But there is also someone there that is much more equipped and better suited to do the job in all respects. Is it permissible for that person to just wait and let me act? And if something goes wrong, who is blameworthy? Whatever answers we give to these questions, we can agree that these are normative claims, playing important roles in the debate on the problem of evil.

Similarly, and more generally, there are discussions about the relation between ordinary moral rules and normative principles and how they apply (or do not apply) to God. One example in the Christian tradition is the Pauline Principle (Romans 3:8), according to which evil is not allowed even for the sake of good (Sterba, 2019). Therefore, one can ask whether it is permissible for an agent to do badly even though good consequences are predicted. Another issue is the Principle of Double Effect. According to this principle, one's intentions are also relevant to the evaluation of actions. For example, it differs if one intends bad consequences, or if one intends to do good and at the same time *foresee* that there would be unintended bad consequences. Aside from the controversy about the principle itself, some have argued that this principle does not apply to God and therefore cannot be appealed to in theodicies, because God has everything in perfection (see: Sterba, 2017).

Be that as it may, it seems that discussion of ethical principles and judgments, specifically in the context of the problem of evil, would hugely impact the debate. This is, of course, acknowledged in theory and practice by some authors, though still seems to be the minority approach.[10]

Illustrated in this way, it might seem that it is obvious and even trivial that the problem of evil is mainly an ethical problem, or at least has important ethical aspects. However, this is not how the literature on the subject has perceived the issue. There are even explicit statements to that effect. As James P. Sterba writes in his edited volume *Ethics and the Problem of Evil*,

> What is a bit surprising, however, is that philosophers currently working on the problem of evil have yet to avail themselves of relevant resources from ethical theory that could similarly advance the discussion of the problem. (Sterba, 2017)

There are also other ways to point out the problem, roughly sketched here and in need of independent evaluation. For one thing, we can look into textbooks. By reading the textbooks of philosophy of religion, we learn a good deal about contemporary epistemology and metaphysics. These issues are either explicitly

[10] In this regard, we can also think of questions of metatheodicy, about the very practice of developing and evaluating theodicies. Here again there seem to be ethical considerations concerning what we do and what we ought to be doing when doing theodicy. For example, one may argue against the very practice of developing theodicies (Trakakis, 2013). On the other hand, one may think of considerations for developing theodicies, hoping for consoling suffering people, as might be inspired Miguel de Unamuno's "San Manuel Bueno, Mártir". Though here I take the suggestion of theodicy as axiology (and, ethical inquiry, more generally) counting for theodicy, I do not claim that it settles this metatheodicy question about theodicizing.

discussed, suggested in the readings, or cited along the way. Similarly, many works of epistemology and metaphysics are included in the bibliographies at the end of well-cited books and papers on the issue. The same is not true about ethical theory. Another way to depict the problem is to consider the authors. There are more prominent metaphysicians and epistemologists who also discuss the problem of evil. This is not the same in the case of ethicists, though there are exceptions.

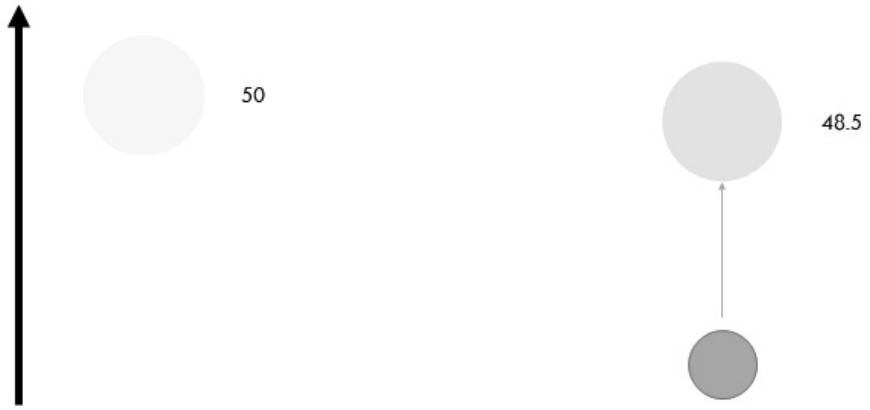
Furthermore, we see how the debate on the problem of evil and the relevant metaphysical and epistemological issues have had mutual influences on each other. Familiar examples might be the cornea principle in epistemology or the relevance of conceptions of free will in metaphysics. The same is not obvious in the case of ethics. One clear example of this gap is how little of the recent changes in the ethics literature has entered the debates on the problem of evil. As Scanlon (2014) points out, one of the main shifts in ethical theory of the last few decades is the centrality of reasons. Many of the main questions in ethics are reconstructed and explored in terms of reasons. Furthermore, this has made it possible to connect different areas of inquiry about the normative, now in a unified field of inquiry. Therefore, some rich, vast, and growing literature on this subject has developed. Yet, not many of these shifts are reflected in the current literature on the problem of evil.

Case of the Moral Progress Approach

So far, it has been suggested that there is a way to consider the problem of evil as an ethical problem. That is, the literature on the problem of evil could benefit from the various debates in ethics. Furthermore, this suggests that we can consider some inquiries into theodicy as ethical inquiries. That is, such attempts may be relevant and even fruitful for ethical theory, even if all theodicies are doomed to fail. Here is a brief example.

Dan Egonsson and I (2021) have proposed a Moral Progress Approach to the problem of evil. Evils in the world have led us to think that a world without evils would be better than this one. This thought could be developed into variations of the problem of evil. In contrast, we have argued that it may be possible for a world, with evils, to be better than its counterpart without evils. In doing so, we consciously rely on axiological claims about progress. The basic idea is that among the sources of value of progress is progress itself (Egonsson, 2018). Therefore, a world that has progressed from a lower point, B, to a better one, A, may be more valuable than a world just being at point A, due to the value of progress itself.

A parallel case in the personal domain can better illustrate and motivate the thought: a life including some progress from B to A may be more valuable than a life without any progress, just being at level A, or even A+, which is a bit higher than A. This suggests that progress is valuable, and its value is not merely from its endpoint.



And to develop this idea into a theodicy and argue for its distinctiveness, we have to address various other questions from the axiology of the world to contrasting features of progress (compared with, say, soul-making). Note that it is one thing to argue for the possibility of a world with evils being better than a counterpart without evils, and it is another to claim that our world is better than a counterpart without evils. The aim is to defend the more modest claim while motivating further developments of the view, moving toward a form of theodicy.[11]

In the same vein, to further develop and defend the Moral Progress Approach one needs to get into the axiological questions involved. Here is an example. It seems that a more specific conception of progress is required for such an approach to have any viability. For one thing, to rely on the value of moral progress in the context of the problem of evil raises an extra question about how to value progress (or how to measure progress, so that we can compare different instances of it and their values).

Imagine that I have progressed from point A to point B, and this is valuable. But does it affect the value of the progress I have had for *how long* I have been at point B?[12] Here is the parallel case regarding worlds: granted that a progressing world has some value that the non-progressing world lacks, does it matter for *how long* the world has been at the lower point and after how long it has progressed? This is important, because even if we accept the rough claim of the Moral Progress Approach, we may wonder whether we may ask for a world with the same amount of progress, but with less time waiting for the progress to come.

It seems that applying axiological claims about progress in the context of the problem of evil may turn out to be helpful to better explore and understand progress,

[11] Whether such a theodicy is successful is another issue. To that end, there are, to be sure, important objections to deal with. However, the relative strength of the view compared to other similar options in the literature (say, soul-making theodicy) is important. The Moral Progress Approach may be easier to defend in this regard.

[12] I have benefited from conversations with Mahmoud Morvarid on this.

whatever the result be as a *theodicy*. And the abovementioned suggestion also was merely one aspect of the issue. The same seems to be true for other questions in ethics, broadly understood. Such attempts though explored in the context of the problem of evil, and even if there is no hope for theodicies to work, may be seen as inquiries into axiology of the world and more. *This* is the world we live in, and it matters to our relationship with it how valuable it is.[13]

References

Adams, Marilyn McCord, & Stewart Sutherland (1989) "Horrendous evils and the goodness of God". *Proceedings of the Aristotelian Society, Supplementary Volumes*, 63, 297-323.

Berker, Selim (MS). "The Deontic, the Evaluative, and the Fitting".

Egonsson, Dan (2018) "Moral progress" in *International encyclopedia of ethics* (edited by Hugh LaFollette), 1-9.

Eslami, Seyyed Mohsen, & Dan Egonsson (2021) "Progress on the Problem of Evil". *International Journal of Philosophical Studies*, 29(2), 221-235.

Eslami, Seyyed Mohsen, & Mohammad Saeedimehr (2021) "The Problem of Evil, Greater Good, and the Moral Objection" [in Farsi]. Philosophy of Religion *Research*, 19(1), 69-92.

Hick, John ([1966] 2010) *Evil and the God of Love*. London: Palgrave MacMillan.

Langtry, Bruce (1998) "Structures of greater good theodicies: The objection from alternative goods". *Sophia*, 37(2), 1-17.

Maitzen, Stephen (2009) "Ordinary Morality Implies Atheism". *European Journal for Philosophy of Religion*, 1 (2): 107-126.

Miller, Christian (2011) "Overview of Contemporary Metaethics and Normative Ethical Theory" in *The Continuum Companion to Ethics* (edited by Christian Miller). London: Continuum.

Mooney, Justin (2022) "The Nonconsequentialist Argument from Evil". *Philosophical Studies*.

Murray, Michael J. and Sean Greenberg. (2016). "Leibniz on the Problem of Evil" in *The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2016/entries/leibniz-evil/>.

Nagasawa, Yujin (2018) "The Problem of Evil for Atheists" in *The Problem of Evil: Eight Views in Dialogue* (edited by Nick Trakakis). New York: Oxford University Press.

[13] Thanks to the editors and readers of the *Festschrift,* especially Jakob Werkmäster and Frits Gåvertsson, for their invitation and comments on the earlier draft. I am sincerely happy to have the chance to express my gratitude to the three philosophers this volume is for – Dan Egonsson, Toni Rønnow-Rasmussen, and Björn Petersson.

Peterson, Michael L. (1998) *God and Evil: An Introduction to the Issues*. Colorado: Westview Press.

Plantinga, Alvin ([1977] 2002) *God, Freedom, and Evil*. Michigan: William B. Eerdmans Publishing Company.

Rea, Michael C. (2018) *The Hiddenness of God*. New York: Oxford University Press.

Rowe, William L. (1979) "The problem of evil and some varieties of atheism". *American Philosophical Quarterly*, 16(4), 335-341.

Sterba, James P. (2019) *Is a Good God Logically Possible?*. Cham: Palgrave Macmillan.

Sterba, James P. (ed.) (2017) *Ethics and the Problem of Evil*. Indiana: Indiana University Press.

Trakakis, Nick (2007) *The God beyond belief: In Defence of William Rowe's Evidential Argument from Evil*. Dordrecht: Springer Science & Business Media.

Trakakis, Nick (2013) "Anti-theodicy" *in The Blackwell Companion to the Problem of Evil* (edited by Justin McBrayer & Daniel Howard-Snyder). Wiley-Blackwell.

Van Inwagen, Peter (2006) *The Problem of Evil*. New York: Oxford University Press.

Rock-Bottom Reasons

Cathrine V. Felix

Since morals… have an influence on the actions and affections, it follows, that they cannot be derived from reason; and that because reason alone, as we have already proved, can never have any such influence. Morals excite passions, and produce or prevent actions. Reason of itself is utterly impotent in this particular.

– David Hume, *A Treatise on Human Nature*

The only rationality of action is the rationality of internal reasons.

– Bernard Williams, "Internal and External Reasons"

Abstract. Bernard Williams inspired the debate about reasons for action. He argues that the only genuine reasons for action are internal. I criticize his account and argue for external, *rock-bottom reasons*. Rock-bottom reasons are not psychological states of the agent but responses to the external world and, indeed, the fundamental reasons for action. They are *motivating* reasons in that they motivate action, and they are *explanatory* reasons in that they explain why action occurred. What they are not are *moral* reasons: they do not involve psychological states, including values and the like. Seeing things this way allows for the common-sense view that agents' behavior can be understood without reference to psychology. My view is that human beings are fact-responsive creatures.

Introduction

Bernard Williams' (1981) views on explanatory reasons for action have not received as much attention as his views on normative reasons from the same paper. This is so even though Williams (1999: 102) himself states that the explanatory dimension is very important.[1]

I wish to take the explanatory dimension deeply into account. The multiple dimensions of Williams' views on reasons are thoroughly intertwined even as it is not always clear how they relate. Williams' conclusion *is* clear: the very idea of external reasons for action is empty. An agent cannot have reasons to act that are independent from every aspect of the agent's psychological states (the agent's *subjective motivational set,* dubbed *S*; Williams, 1999: 102)). Without motive*,* one cannot explain action. In consequence, for a reason to explain action, it must be internal. Thus, Bernard Williams is a skeptic about external reasons.[2]

Williams begins by pointing out that there are two contrasting statements that can be made of an agent's reasons for action:

1. *A* has reason to Φ, and
2. There is a reason for *A* to Φ.

The first statement establishes an essential connection between reasons and motivation; as such it relates to the agent's psychology, and so falls under the heading 'internalist' – in accordance with the Humean claim that only desires can motivate. The statement can be falsified by *A*'s lack of "motive which will be served or furthered by his Φ-ing…There is a condition relating to the agent's aims, and if this is not satisfied it is not true to say, on this interpretation, that he has a reason to Φ" (Williams: 1999: 101).

The second statement establishes no such connection between reasons and motivation. Williams calls it 'externalist'.[3] So, on the one hand, there may be

[1] Williams does not distinguish between normative and explanatory reasons: "some writers make a distinction between «normative» and «explanatory» reasons, but this does not seem to me to be helpful, because normative and explanatory considerations are closely involved with one another" (Williams, 2001).

[2] Sobel (2001: 218) writes: "Williams' claim that reasons must be interrelated with explanation in a particular way… does not support internalism as he supposes". Sobel shifts focus from the externalist-internalist distinction to that between objectivism and subjectivism. "The interesting question is… not whether to embrace internalism or externalism, but whether to embrace objectivism or subjectivism – a debate that may boil down to a dispute about the powers of practical reason to bring about consensus in the motivations of people who start out with radically different motivations" (234-5). Although I find Sobel's approach intriguing, my focus is on the internalist-externalist distinction put forth by Williams.

[3] Williams does not claim that there are two kinds of reasons: "I shall… for convenience refer sometimes to 'internal reasons' and 'external reasons', as I do in the title, but this is to be taken only as a convenience" (Williams, 1999: 101).

internal reasons for action (Statement 1): reasons substantially connected to *S*. On the other, there may be external reasons not related to *S* (Statement 2). Williams (1999: 111) writes that "the only real claims about reasons for action will be internal claims": i.e., externalist claims are false.

Motivation

The Humean idea inspiring Williams, that people are fundamentally driven by inner states, is widespread and considered by many to reflect common sense.[4] On the internalist view, reasons for action are determined by states of mind. Donald Davidson writes (2001: 3-4): "whenever someone does something for a reason, therefore, he can be characterized as a) having some sort of pro attitude towards actions of a certain kind, and b) believing…that his action is of that kind".[5]

Every reason for acting must include a desire and a belief[6] : whenever an agent performs an action, part of her reason must be that she wants to achieve some result and another part must be her belief that performing the action will achieve the desired result. Say that I have a desire to catch your attention and believe that waving is a way of achieving that result; I further believe that *this* is an act of waving. I need the belief to guide my action and achieve my goal. Crucially, the belief merely guides action and cannot, on the internalist view, motivate action. Williams starts off from this view, but he makes some adjustments.[7]

---

[4] Millgram (1996: 197) neatly expresses the popularity of this view: "experience shows that the internalist take on reasons can appear enormously compelling to a very wide range of the philosophical community, from freshmen in their first ethics class to seasoned professionals in their most tough minded moods".

[5] The term "pro-attitude" was introduced to serve as a broader term than "desire" in the sense used by Hobbes, who treated all motives as desires. It was coined by Nowell-Smith (1954: 112) and adopted by Davidson among others. (Nowell-Smith also writes of *con-attitudes*.) For Davidson (2001), a pro attitude consists of "desires, wantings, urges, promptings, and a great variety of moral views, aesthetic principles, economic prejudices, social conventions, and public and private goals and values in so far as these can be interpreted as attitudes of an agent directed towards actions of a certain kind." See also Petersson (2000).

[6] It is widely recognized that Davidson (2001) resurrected the belief-desire view for contemporary analytic philosophy. Its historical source is normally taken to be Hume, so it is sometimes called the *Humean story* (Smith, 1998). It is also referred to as the *desire/belief-thesis* (Bittner, 2001), the *Standard View* (Dretske, Stoecker & Sandis, the *Standard Model* and *Standard Story* (Hornsby, 2003). and the *Received View* (Stoutland, 2007). The most common label until recently was the *Belief Desire Model* (Petersson, 2000). Bratman's version is sometimes referred to as the BDI Model for belief-desire-intention. Bittner (2001: 29) traces the idea of "desire providing the impulse and reason guiding it" far beyond Hume all the way back to Socrates' speech in Plato's *Phaedrus* (246a6-7).

[7] Ingmar Persson (1997) has argued it's doubtful that Hume himself was a Humean in the modern sense. Thanks to Björn Petersson for this information.

Williams' on Internalist Reasons

Williams (1999: 101) offers a simplified version of the Humean view: "*A* has a reason to Φ iff *A* has some desire the satisfaction of which will be served by his Φ-ing", which he calls the *sub-Humean model*. A large part of Williams' subsequent discussion amounts to a refinement of the sub-Humean model to make it easier to set up against the externalist account.

Williams (1999: 102) emphasizes that "basically, and by definition, any model for the internal interpretation must display a relativity of the reason statement to the agent's *subjective motivational set*" (emphasis original). Still, he says that even though the agent's subjective motivational set is discussed "primarily in terms of desire" (105), this is only intended in a formal philosophical sense. According to Williams (102), for something to be an internal reason, the action in question must be suitably related to an element in *S,* whose members are *not* restricted to desires[8]: "*S* can contain such things as dispositions of evaluation, patterns of emotional reaction, personal loyalties, and various projects, as they may be abstractly called, embodying commitments of the agent" (105).

Williams (1999) offers two main arguments against externalism: one concerning explanations of actions, and one based on the role of practical reasoning in action.

1. "If there are reasons for action, it must be that people sometimes act for those reasons, and if they do, their reasons must figure in some correct explanation of their action" (102).

2. "…No external reason statement could *by itself* offer an explanation of anyone's action… The whole point of external reason statements is that they can be true independently of the agent's motivations. But nothing can explain an agent's (intentional) actions except something that motivates him so to act" (106-107).

To explain what it takes for something to be an external reason for action, Williams offers the example of an agent who has no interest whatsoever in obtaining what she wants/needs. He reasons that if there really is some such person, then she also "has no reason to pursue these things" (Williams, 1999: 105). If she needs life-saving medicine to survive yet refuses to take it, and one continues thinking that she has a reason to take the medicine, then one cannot mean reason for action in an internal sense (the agent is fully unmotivated), so reason in this case must be *external* reason

[8] Contrary to what one might suppose, a desire need not have any connection to strong emotions; what is meant by "desire" is simply that which an agent has whenever motivated to act. Consider the story (Reuters, 2007) that went viral in 2007 of the Buddhist monk in Thailand who hacked off his penis with a machete because he had an erection during meditation. He refused to allow the doctors to sew it back on because he wanted to avoid such distractions in the future. It can be argued that, for the ascetic, it can be rational – given her motivations – to deny herself something that arouses in her powerful feelings.

for action. Williams reasons that it is an assumed part of everyday communication that "people… say things that ask to be taken in the external interpretation" (Williams, 1999: 106). If one says to the person who lacks interest in her own well-being that she has a reason to take the medicine, one means an external reason. Williams offers another example: namely Henry James' story of Owen Wingrave, who lacks the motivation to follow family tradition and join the army: "Owen's family urge on him the necessity and importance of his joining the army, since all his male ancestors were soldiers, and family pride requires him to do the same. Owen Wingrave has no motivation to join the army at all, and all his desires lead in another direction: he hates everything about military life and what it means. His family might have expressed themselves by saying *that there was a reason for Owen to join the army*" (Williams, 1999: 106; emphasis original). At the same time, "…no external reason statement could *by itself* offer an explanation of anyone's action" (Williams, 1999: 106-107). That there is a reason (his family's reason) for Owen to join the army cannot explain Owen's reason not to: "for if it was true at all, it was true when Owen was not motivated to join the army" (Williams, 1999: 107). Although Williams does not find explanation by external reasons useful, he does find it necessary to make improvements to the original Humean model.

Williams' Refined Sub-Humean Model

First, *A* can have reason for acting without believing this to be the case. There might be something in the agent's situation that she fails to recognize as a means to reach her ends. If *A* wants to win a game of chess, there are moves she has a reason to make even though she might not see them.

Second, *A* might have reason to act even when *A* lacks any desire to do so. Consider Iris, who was not motivated to stop smoking in 1999, yet had reason to stop smoking already then, even though she did not realize that she has such a reason until 2013 when she – through practical deliberation – realized it to be the case. Iris might see herself as having been a nonsmoker at heart in 1999 and having valued quitting cigarettes already then. On Williams' account, this is acceptable, because *stop smoking in 1999* belongs to Iris's overall motivational structure.

By contrast, Ronny, who was also a smoker in 1999 and remained one in 2013 might well have deliberated and come to the conclusion in 2013 that he had reason to stop smoking – without also concluding that he had reason to do so already in 1999. Back then, Ronny valued smoking over quitting and had no reason – within his overall motivational structure – to stop.

Williams' interpretation of practical reasoning is wider than that of the original Humean model. Part of his refinement is that, on his view, deliberation is not limited to becoming aware of causal means to an end through overt reasoning (1999: 104):

> A clear example of practical reasoning is that leading to the conclusion that one has reason to Φ because Φ-ing would be the most convenient, economic, pleasant etc. way of satisfying some element in *S,* and this of course is controlled by other elements in *S*, if not necessarily in a very clear or determinate way. But there are much wider possibilities for deliberation, such as: thinking how the satisfaction of elements in *S* can be combined, e.g. by time-ordering; where there is some irresoluble conflict among the elements of *S*, considering which one attaches most weight to (which, importantly, does not imply that there is some one commodity of which they provide varying amounts); or, again, finding constitutive solutions, such as deciding what would make for an entertaining evening, granted that one wants entertainment.

As noted above, it is compatible with Williams' approach that an agent's rational reasoning can make her aware of reasons she had at some point in the past that she was not then consciously aware of: An agent can come to see that he has reason to do something which he did not see he had reason to do at all. In this way, the reasoning process can add new actions for which there are internal reasons (Williams, 1999: 104).

Williams envisions a more fruitful role for practical reasoning than the original Humean model. "Practical reasoning is a heuristic process, and an imaginative one" (Williams, 1999: 110). Practical reasoning is far more than means/ends reasoning to satisfy desires. At the same time, something can count as a reason for an agent if and only if the reason bears a link to aspects of the agent's psychology. This is where Williams departs from the externalist, because "the whole point of external reason statements is that they can be true independently of the agent's motivations" (Williams, 1999: 107). It is time to spell out and then criticize Williams' views on external reasons.

Internal vs. External Reasons

Williams (1999: 108-109) claims that "…an element which the externalist theorist essentially wants [i.e., lacks], [is] that the agent should acquire the motivation because he comes to believe the reason statement, and that he should do the latter, moreover, because, in some way, he is considering the matter aright." He continues (109):

> If the theorist is to hold on to these conditions, he will, I think, have to make the condition under which the agent appropriately comes to have the motivation something, that he should deliberate correctly; and the external reasons statement itself will have to be taken as roughly equivalent to, or at least as entailing, the claim that if the agent rationally deliberated, then, whatever motivations he originally had, he would come to be motivated to Φ.

Williams assumes that the external-reasons theorist should accept Hume's basic point that reason alone cannot motivate: "…there is no motivation for the agent to deliberate *from*, to reach this new motivation" (Williams; 1999: 109). The only cases Williams sees as potentially supported by externalism about reasons-for-acting are those that involve persuasion or manipulation (Williams 1999: 107):

> The basic case must be that in which A Φ's, not because he believes only that there is some reason or other for him to Φ, but because he believes of some determinate consideration that it constitutes a reason to Φ. …Owen Wingrave might come to join the army because (now) he believes that it is a reason for him to do so that his family has a tradition of family honour.

Four points bear noting. First, if Owen's family convinces him to join the army, he does so by giving in to pressure: he acts as a puppet on a string, lacking (so the argument goes) any personal motivation to act as he does. As commonly understood, external reasons appear to constitute too narrow a concept, unlikely to explain all one might want to explain when explaining actions. On the other hand, understanding them on Williams' terms – allowing in the exceptions that he does – not only does not help them but risks widening their scope too far. If this is what external reasons are about, the internalist position seems a better alternative. The question is, must the externalist accept Hume's linking of reasons to motivation? Such a view fails to conceive of reasons as external in what one might see as a meaningful sense.

Second, if Owen is moved by other persons' reasons, he is indeed moved to action by external *factors*, but, crucially, I would argue, not by external *reasons*. To be driven by external reasons should be reserved for responding to facts about the external world. It is not to respond to the manipulation or interests of others.

Third, Williams presents his case as if responding to external reasons is equivalent to responding to systems of justification held by others, such as Owen's family. According to their values, joining the army is the right thing to do. Doing so is internal to their values but external to Owen's. The question here is, why should it be a problem for the externalist about reasons if a person fails to be personally motivated by a set of values held by some group of people, such as family? Indeed, Williams argues elsewhere (1999b) for such relativism about reasons.

Fourth, one can well ask why it is so important for Williams to appeal to an agent's personal motivational set. The answer seems to be an assumption that the act must somehow stem from a choice: it must be up to the agent. This is probably why Williams contrasts internal and external reasons in the way he does. For him, internal reasons stem somehow from free will, while external reasons are due to outside forces.

Williams does not tell the whole story about Owen Wingrave. It is highly likely that Owen was homosexual. (He also opposed his family's stated wish for him to get married.) Being a homosexual in the army was at the time – and still is in many

parts of the world – incredibly challenging. Presuming Owen was gay, his resistance to signing up is probably attributable to his sexual orientation: a part of him deeply rooted in his character and identity as a human being. In that case, Owen responded rationally to external reasons: namely, the poor treatment many gay people received in the military.

What are internal reasons if not the reasons described above? What is more intrinsic to a person than her sexual orientation and love interests? If one is sympathetic to this way of thinking though, it makes a problem for Williams' focus on motivation and personal choice. In most people's experience as backed by accumulating scientific research, sexual orientation is not a choice (never mind a choice that can be altered)[9] unless one believes in the discredited practice of conversion therapy. If internal reasons can come down to something like sexual orientation, then Williams' preferred internal reasons appear just as narrow and problematic as the external reasons he criticizes, for the agent's motivation to act / not act can have its wellspring in something *not* freely chosen, being deeply rooted in character and identity.

It might be that an agent is genuinely motivated to Φ but, due to her core nature, is simultaneously motivated not to Φ: i.e., she is conflicted. Consider an extremely shy person motivated to speak up against her tyrannical boss but most likely to keep quiet. One could say that she manipulates herself to do something different from what she would like to do.

Of external reasons for action, Williams (1999: 111) concludes that "there is, I suggest, a great unclarity about what is meant", outside cases of external manipulation and the like. Much of the problem though is of his own creation.

Internal reasons are not the kind of things that can be true or false the way external reasons can be and generally are. Values are not true or false. Core identity is not true or false: it just *is*. External reasons are of another ilk. They are not grounded in the psyche of an agent or a shared system of values as with Owen's family. External reasons for action are responses to facts that the agent knows and responds to in the world, more or less consciously. An agent acts according to external reasons when she responds to how things are in the world.

It is time to make the positive case for external reasons. Most of the time and unusual cases aside, the reasons why people act are in the world and not in their minds. Human beings are, at heart, fact-responsive creatures driven not by psychology but response to reality, which offers up the "rock-bottom" reasons: the *real* reasons for action.

[9] Depending on where one lies along a sexual spectrum or within a space of sexual possibilities, some people do have a choice whether or not to explore a side of their identity they have not previously examined. In any case, how one identifies sexually is a highly complex matter far beyond the purview of this paper.

Fact Responsiveness

Consider Maria, who is driving a car (Stoutland, 2007). Approaching an intersection, she sees a red traffic light ahead and stops. The explanation for her action is nothing more than that the traffic light was red: she took *this* fact[10] about the world to favor her stopping the car.

Yet this answer is not available for Williams, who would be compelled to identify some desire, the satisfaction of which the agent is acting in accordance with. Remember: *A* has reason to Φ iff *A* has some desire the satisfaction of which will be served by his Φ-ing (Williams, 1999: 101)). Even though Williams' view is significantly more sophisticated in the end, it demands a reason to act that it be substantially rooted in the psychology of the agent. To explain Maria's behavior, Williams must add something like a goal within Maria's overall subjective motivational structure *S*, something like an overarching goal of obeying the law: her desire and thus motivation points towards that goal; she has an accompanying belief that, if she stops at the light, her desire will be satisfied; etc. Of course, such a story might be true in some cases; but it is a bold claim indeed to say that *all* cases are like this. Such a claim would be as audacious as Hume is in the introductory quote, where he takes his point about the impotence of reason as proven. As Anscombe (1981) has pointed out, it is highly doubtful that Hume or anyone else has proven any such thing (or could).

There are reasons why an internalist account of reasons for action nevertheless retains appeal, notably in cases of false belief. Say that Belinda walks across the room to open her door; it is fitting to appeal to a belief such as "somebody rang the doorbell" – both in the case where somebody actually has rung her doorbell and where she is mistaken. Such an example is easily explained by the internalist by appealing to the agent's inner psychological state. The externalist runs into trouble with cases of false belief, because – on her view – reasons are external to the agent, and it seems odd that something that is not the case can nevertheless be a reason.

This is where I see the appeal to rock-bottom reasons being most useful. Instances of reasons for acting based on false beliefs[11] can either be due to the agent's psychological profile – which neither internalists nor externalists need have problems with – or the false belief can be traced back to a response to external circumstances: what I am calling rock-bottom reasons. It is, I think, just a matter of Gricean conversational implicature that one rarely mentions the external triggers that give rise

[10] The literature reveals differing preferences regarding use of terms. Many prefer "state of affairs" (e.g., Dancy, 2000; Stoutland, 2007), or "true propositions" (Smith, 1997). Suffice to say that I prefer the term "facts" because facts are what I see human beings being reason responsive to, favoring them to act in certain ways and not others. *Pace* Dancy and Stoutland, I do not accept the possibility of "non-factive" states.

[11] Of course, there can be cases where the agent's belief is true even though this is not what moves the agent to act, and they can be handled in a similar way.

to false beliefs; they seem not worth mentioning. Instead, one focuses on that which is most relevant to the context one finds oneself in. Anscombe (2000: 8) writes:

> I am sitting in a chair writing, and anyone grown to the age of reason in the same world would know this as soon as he saw me, and in general it would be his first account of what I was doing; if this were something he arrived at with difficulty, and what he knew straight off were precisely how I was affecting the acoustic properties of the room…then communication between us would be rather severely impaired.

In ordinary discourse, it is not difficult to communicate or understand what is going on in false-belief cases. In philosophy or psychology though, one ignores the triggering reasons at one's peril. They have important theoretical ramifications, as in the debate over internal vs. external reasons.

It is important briefly to consider those unusual cases identified by Dancy (2000) where what appropriately is taken as reason for acting is a false belief *per se*, untethered to external facts. Dancy (2004: 124) writes: "normally, if things are not as I believe them to be, I do not in fact have the reason that I take myself to have".[12] That the mere fact that an agent has a belief is normally *not* a reason for acting is neatly illustrated by Dancy's countering example of a hill walker and a crumbling cliff: the walker knows himself well enough that, aware that he *believes* it is dangerous to climb a cliff, he knows, too, he will get so nervous he is likely to panic and fall. He sees the cliff and out of compulsion forms the belief that the cliff is crumbling – whether it is or not – thus dangerous to climb. In this case, that the agent has this belief is a good reason to avoid climbing the cliff, regardless of actual circumstances.

What makes Williams'' internalism on reasons for action most problematic is that people lack direct access to the motives of others; they must infer motives instead, judging on the basis of statements and observable actions. Given their interpretative skills and ability to cooperate, they nevertheless manage, more often than not, to understand others' motives. Morton (2003: 1) writes:

> You are walking towards a closed door, with your arms full of groceries. Another person is also approaching the door, slightly ahead of you. He accelerates his pace slightly. This generates an expectation in you. He has either seen the problem you face and intends to solve it by opening the door for you, or he sees that you might expect him to open the door and is rushing to get through before the issue arises.

It is part and parcel of everyday interactions with others that people form expectations this way, judging as best they can whether someone is cooperative or not, adjusting to one other's behavior as needed. Serious questions remain about how exactly they do this. It seems as though it cannot be because they directly *see* the motive of others. Williams, who seeks to explain actions in terms of agents' motives has difficulty cashing out exactly what happens when understanding others'

[12] Dancy (2004: 124) credits Joseph Raz (1986: 142-143) for these sorts of arguments.

actions. In particular, he struggles to account for cases where the agent's action does not reflect the agent's motive. As author Aksel Sandemose complained: "it is so annoying when my neighbour approaches me when I'm working in the garden. He doesn't understand that I'm busy writing my next book".

When explaining reasons for action, it is not only possible but often necessary to explain behavior without reference to psychology. This is because one often has no way of knowing what exactly goes on in the agent's mind when performing an action, and, moreover, there are actions where the agent's motivation is no help in understanding the agent's reasons for acting. Instead, one must focus on details of the agent's situation at the time, relating these to one's background knowledge. In this way, one can form an understanding of which external circumstances the agent is responding to and so arrive at her rock-bottom reasons.

One could object that, if one fails to consider the possible psychological motivations that make an agent act as she does, one would be downplaying the importance of the agent's rationality in justifying her behavior and so failing to distinguish properly between her actively intentional actions and mere happenings for which she is somehow responsible. There are good reasons why one does not normally attribute full agency to young children, most non-human animals, or the severely brain damaged.

The objection is fair, as far as it does. What does not follow is that *all* action explanations must be linked to an agent's psychological states so as to constitute proper explanations. All one needs to counter Williams' internalism is some cases that require no link to *S*. Although it is reasonable to require agents to fill certain rational requirements before attributing actions to them, that need not mean knowing every motivation for every action for every agent – or even assuming that relevant psychological states could be determined, even in principle.

Ascribing psychological explanations to actions is something people often if not indeed generally do, lending Williams' account its unquestionable appeal. When interacting with others, people try to imagine the other person's point of view – adopt her perspective – in explaining to themselves why the person acts and responds as she does. People judge one another on their understanding of the person's character; *ceteris paribus,* one trusts one's gut feelings about what someone is "really" like regardless of externally observable signs one can point to and name. It feels natural to describe people as envious, jealous, joyful, absentminded, etc. I have no quarrel with any of this. At the same time, it's important to remember that even the most skillful judge of others can be mistaken in her judgments. Maybe the sweet, shy colleague at work abuses her son? Maybe the seemingly brilliant shopkeeper rarely pays her bills? The moral of the story is, psychological explanations for actions can be useful, even necessary, more often than not correct. If they proved often wrong, cooperation would be difficult to imagine. At the same time, they are *not* always useful, *not* always needed, *not* always to be counted on. Reasons for any number of actions can best be determined without resort to psychological states.

Conclusions

Behind agents' reasons to act lie rock-bottom reasons, responding to facts in the world, that serve as the ultimate motivating force for actions. The most significant distinction between my approach and Williams' is that he rejects external reasons for action while I support them and, indeed, consider them primary. Next up is his focus on grounding reasons in ethical considerations. Even though Williams wants to be seen as speaking in general of reasons to act, ethical considerations are always close at hand, whereas I take them to be secondary. That can make it seem as though Williams and I are talking past each other even though, really, we are not. I am not making claims one way or another about ethical considerations for reasons to act.

If I have managed to show that Williams' arguments against externalism do not hold and that external reasons for action are needed at least in some instances, I rest content. I give the final words to the poet Gunnar Ekelöf (1941): *det finns ingen annan styrka än inre styrka / och den kommer utifrån* ("there is no other strength but inner strength, / and it comes from outside"; *translation mine*).

References

Adler, J. E. (2008). Presupposition, attention, and why-questions. In J. E. Adler & L. J. Rips (Eds.), *Reasoning: Studies of human inference and its foundations.* Cambridge University Press.

Anscombe, G. E. M. (2000). *Intention* (2nd Edition). Harvard University Press.

Anscombe, G. E. M. (1981). "Modern Moral Philosophy". In *Ethics, Religion and Politics*. Blackwell.

Bittner, R. (2001). *Doing things for reasons*. Oxford University Press.

Bratman, M. E. (1999). *Faces of intention*. Cambridge University Press.

Cohon, R. (1986). Are external reasons impossible? *Ethics*, *96*(3): 545-556. https://doi.org/10.1086/292774

Dancy, J. (2004). *Practical reality*. Oxford University Press.

Davidson, D. (2001). *Essays on actions and events*. Clarendon Press.

Dretske, F. (2009). What must actions be for reasons to explain them? In C. Sandis (Ed.), *New essays on the explanation of action.* Palgrave Macmillan.

Ekelöf, G. (1941) *Färgesong.* Bonniers.

Finlay, S. (2009). The obscurity of internal reasons. *Philosophers' Imprint*, *9*(7): 1-22.

Grice, H.P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Speech acts* (pp. 41-58). Accessed 14 November 2022 from http://www.ucl.ac.uk/ls/studypacks/Grice-Logic.pdf

Hobbes, T. (2004). *Leviathan*, ed. R. Tuck. Cambridge University Press.

Hooker, B. (2001). Williams' argument against external reasons. In E. Millgram (Ed.), *Varieties of practical reasoning.* MIT Press.

Hornsby, J. (2003). Agency and actions. In J. Hyman & H. Steward (Eds.), *Agency and action* Cambridge University Press.

Hume, D. (1978). *A treatise of human nature* (2nd Edition). Ed. L. A. Selby-Bigge and P. H. Nidditch. Clarendon Press.

James, W. (1890). *The principles of psychology*. Macmillan.

Lubin, D. (2009). External reasons. *Metaphilosophy, 40*(2): 273-291. https://doi.org/10.1111/j.1467-9973.2009.01580.x

McDowell, J. (1998). Might there be external reasons. In J. McDowell (Ed.), *Mind, Value & Reality.* Harvard University Press.

Millgram, E. (1996). Williams' argument against external reasons. *Noûs*, *30*(2): 197-220. https://doi.org/10.2307/2216293

Morton, A. (2003). *The importance of being understood*. Routledge.

Nowell-Smith, P. H. (1954). *Ethics*. Penguin Books.

Persson, I. (1997). Hume – Not a Humean about Motivation. *History of Philosophy Quarterly*, 14(2): 189-206.

Petersson, B. (2000). Belief & Desire: The Standard Model of Intentional Actions – Critique and Defence.

Reuters staff (2007, January 20). Buddhist monk cuts off penis and renounces refix. Retrieved 11 November 2022 from https://www.reuters.com/article/oukoe-uk-life-thailand-monk-idUKBKK28492620061122.

Sandis, C. (2009). Introduction. In C. Sandis (Ed.), *New essays on the explanation of action* Palgrave Macmillan.

Setiya, K. & Pakkunainen, H. (eds.) (2011). *Internal reasons: Contemporary readings*. MIT Press.

Smith, M. (1997). *The moral problem*. Blackwell.

Sobel, D. (2001). Explanation, internalism, and reasons for action. *Social Philosophy & Policy*, *18*(2): 218-235. https://doi.org/10.1017/S026505250000296X

Sperber, D. & Wilson, D. (1996). *Relevance: Communication and cognition*. Blackwell.

Stoutland, F. (2007). Reasons for action and psychological states. In A. Leist (Ed.), *Action in context.* Walter de Gruyter.

Williams, B. (1998). Internal reasons for action and the obscurity of blame. In B. Williams, *Making sense of humanity and other philosophical papers 1982-1993* (pp. 35-45). Cambridge University Press.

Williams, B. (1999). Internal and external reasons. In B. Williams, *Moral luck: Philosophical papers 1973-1980* (pp. 101-113). Cambridge University Press.

Williams, B. (2001). Postscript: Some further notes on internal and external reasons. In E. Millgram (Ed.), *Varieties of practical reasoning.* MIT Press.

Individually Fitting but Collectively Unfitting Blame

Andrés G. Garcia

Abstract. People that produce bad outcomes can thereby become the fitting targets of blame, the fitting intensity of which is determined by the badness of the outcomes. In the following paper, I suggest that the amount of blame instances that people are fitting targets of is also determined by the weight of the badness of the outcomes. I use the example of online blame as a paradigmatic case where the amount of blame instances that people are made targets of risks being excessive, even when each instance of blame is fittingly held and expressed.

Introduction

Debbie is morally responsible for having produced an outcome of moral value and has made herself the fitting target of reactive attitudes of a certain kind, duration, and intensity. The kinds of reactive attitudes that Debbie is the fitting target of is partly determined by the character of the moral value that she has produced. If the outcome is morally bad, then Debbie is typically the fitting target of blame, and if the outcome is morally good, then Debbie is typically the fitting target of praise.[1]

[1] The qualifier "typically" leaves room for the view that having produced bad outcomes is not *sufficient* to make agents fitting targets of blame. Agents are blameworthy when they fail to meet up to a normative standard. How we should cash this out in detail is controversial, but philosophers have offered additional constraints in the literature. These are meant to rule out cases where agents produce bad outcomes accidentally—or in other ways that are not suitably reflective of their character and understanding. Strawson (1962) insists that what he refers to as the "quality of will" of agents helps determine their blameworthiness. Others have argued that agents being fitting targets of blame entails

The duration and intensity of the reactive attitudes that Debbie is the fitting target of is partly determined by the weight of the moral value that she has produced. If the outcome is very bad, then Debbie might be the fitting target of a relatively long period of intense blame; and if the outcome is only slightly good, then Debbie might be the fitting target of a relatively short period of mild praise.[2]

There is a normative intuition suggesting that just as the weight of the moral value that Debbie has produced is relevant to the duration and intensity of the reactive attitudes of which she is the fitting target, it is also relevant to the *quantity* of reactive attitudes that can be fittingly directed at her. Suppose Debbie has made insensitive comments about a group of people on an online social media platform. We can imagine that once enough people learn of her moral transgression, she becomes the target of thousands of instances of blame, each of which is individually fitting in terms of duration and intensity. In the following paper, I shall explore and try to make sense of the suggestion that while each instance of blame that is directed at Debbie may also fit her, she can nonetheless be the unfitting target of the *total amount* of blame instances that is collectively directed at her.

I shall start out by elaborating on the normative intuition that there can be something problematic about a collection of attitudes. I illustrate it by appealing to a type of case where people do something online and become targets of reactions that seem disproportionate in terms of collective scope. Previous works on the dangers of moral sanctions on the internet have often been limited to their harmful effects on their targets (Tosi & Warmke 2020: 103; Billingham & Parr 2020), e.g., by considering how expressions of blame (understood broadly, to include speech acts as well as punitive practices) may affect the blameworthy. I will take a slightly different perspective by looking at the issue whether there can be something problematic about a quantity of blame instances as such, irrespective of its potential for psychologically harmful effects *vis-à-vis* the blamed.[3]

Excessive Blame

The paper proceeds from the normative intuition that not everything is as it should be regarding the quantity of blame that is sometimes heaped on individuals, even though they may have produced morally bad outcomes and made themselves the fitting targets of blame. Real-life examples that illustrate the phenomenon are inevitably controversial, but the infamous case of Justine Sacco springs to my mind.

---

that they have sufficient knowledge, understand their own actions, and are aware of relevant reasons and values (e.g., Held 1970; Pettit 2007; Coates & Swenson 2013; Nelkin 2016; Tierney 2019).

[2] For more on how the weight of value relates to the intensity and duration of fitting attitudes, see Andersson & Green Werkmäster (2020).

[3] See also Aitchison & Meckled-Garcia (2021) for the relevance of non-consequentialist *disrespect*.

Sacco was an American woman on her way to South Africa for work. Before she boarded her flight from the United States, she logged into the social media platform *Twitter* and posted a public message in the form of an insensitive joke: "Going to Africa. Hope I don't get AIDS. Just kidding. I'm white!" (Ronson 2015). By the time Sacco woke up in Cape Town, she had lost her job and was the target of a massive online campaign involving tens and thousands of enraged people.

I will leave aside issues about whether Sacco's joke was just a crude satire about racism and privilege or whether the joke itself was an expression of these things. I shall also bracket questions about whether the intensity and means by which people subsequently expressed their blame toward Sacco were appropriate.[4] I will instead look at the question whether there could be something about the quantity of blame that was directed at her that was not as it should have been. Even if each individual that blamed Sacco may have done so fittingly, the normative intuition is that we should be given pause by the sheer quantity of blame that ended up being directed at her in the end. *The unfittingness view* states that while Sacco was the fitting target of blame, her overall moral character could still have made her the unfitting target of the collection of blame that ended up being targeted at her.

To avoid being stuck in the specifics of the Sacco example, I will continue to phrase my discussion in general terms and focus on the hypothetical case of Debbie. My hope is that readers will recognize the ubiquity of the general phenomenon to which I am alluding, even if they happen to have qualms about the example I have just used. The question is whether there is some way of capturing and vindicating the normative intuition that I have emphasized without adopting the unfittingness view. In other words, is it possible to explain what it means that not everything is as it should be regarding the quantity of blame that is sometimes heaped on individuals, without thereby accepting that collections of blame instances themselves can in some sense be unfitting? Before I attempt to answer this question, I need to lay out some relevant assumptions that I will be making about *blameworthiness*.

I assume that an agent that produces bad outcomes is typically blameworthy and that this means that they are the fitting targets of blame.[5] Fittingness should of course be understood as a normative relation holding between a response and a target. Blame fits blameworthy agents much like admiration fits admirable artworks, respect fits respectable achievements, and love fits lovable people. Certain objects call for certain responses and *vice versa*, by virtue of the former fulfilling the inherent standards of the latter. While the notion of fittingness is a normative one,

[4] My suspicion was that they were not always so. For more details, see Ronson (2015).

[5] This fits well but does not entail a fitting-attitudes analysis of blameworthiness. For more on this pattern of analysis, see, e.g., Rabinowicz & Rønnow-Rasmussen (2004). Also, note that philosophers tend to speak of blame as a complex phenomenon involving both a judgment about the blameworthy and a collection of attitudes, such as moral resentment and anger. I am primarily interested in the *attitudinal* part of blame, meaning that I shall be using the term as a placeholder for those particular non-doxastic and motivational mental states that we direct toward people for the bad outcomes that they produce and for which they thereby become fitting targets.

this does not mean that it must involve a relation of *requirement*. Instead, the notion of fittingness can involve a less demanding sense of congruence, such that if objects fulfil the inherent standards of certain responses, then the objects become *recommended* targets of those responses.[6]

In their work on the fittingness relation, D'Arms & Jacobson (2000) distinguish between a narrow and wide sense in which objects can be fitted to certain responses. The former involves a congruence between responses and their targets, exclusively grounded on facts about how well the former fulfils the inherent standards of the latter. By contrast, the wide notion of fittingness is guided by a more general sense of what is good and right and thus takes into consideration the *effects* of adopting certain responses as well.[7] Similar ambiguities presumably occur for physical notions of fit. A Harley Davidson vest could be visually fitted to me in the narrow sense that I have the right measurements to wear it, but because such a vest is meant to convey certain things about the look and demeanour of its bearer, wider contextual factors may cause the vest to *not* fit me visually.

The unfittingness view states that while Debbie may have produced bad outcomes for which she is the fitting target of blame, her overall moral character can still make her the unfitting target of the quantity of blame that is directed at her. This could now be interpreted as saying that while Debbie is the fitting target of blame in the narrow congruence sense, she is the unfitting target of a collection of blame in the sense that involves our general sense of what is good and right. I think that this is true as far as it goes, but we need to be careful here as the point is not just that Debbie can be psychologically affected by becoming a knowing target of a quantity of blame. The intuition is that not everything is as it should be regarding the fact that she is a target of so many blame instances *irrespective* of the emotional harm that she might suffer from her awareness of being a target.[8]

I will return to the two different fittingness notions momentarily, but before that it might be helpful for me to take a quick detour and consider the main rivals to the unfittingness view. The aim of such accounts would be to explain how there might be something problematic about the kind of cases to which I have alluded without

[6] The idea that attitudes have "inherent standards" that determine their fittingness to certain targets comes from McHugh & Way (2016).

[7] Actually, D'arms & Jacobson distinguish between a narrow and broad sense of *appropriateness* and reserve the term 'fittingness' for the narrow sense. I do not think that this difference marks an important disagreement between us. The same distinction that they have in mind seems to apply to the everyday notion of fittingness.

[8] It would be uncontroversial to suggest that an object could be fitting as a target of a collection of attitudes that are adopted by an individual agent (i.e., fitting attitudes can be *intrapersonally* aggregated). For example, a lovable person may to that extent also be admirable and respectable, which typically entails that they are the fitting target of a collection of attitudes that includes love, admiration, and respect. The question is if a person can be the fitting target of a collection of attitudes that are adopted by a single agent, why could a person not be the unfitting target of a collection of attitudes that is dispersed among many (i.e., fitting attitudes that are *interpersonally* aggregated)?

having to assign any normative status to collections of blame instances *as such*. Considering the general features of such accounts will help us see the commitments of the unfittingness view and what it might entail for our understanding of the normative dimensions of blame. In the next section, I will start out with an account that puts emphasis on the notion of *moral standing*.[9]

Atomistic Understandings

It seems clear that for an individual to be a fitting blamer in a given situation, they must have moral standing to blame in that situation. For example, a person that has produced a morally bad outcome can lack moral standing to blame someone else for producing that outcome—or for producing an outcome of the same general type. This suggests the possibility that there might be cases where there is a limited number of places of moral standing that would make people fitting blamers. Debbie has produced a morally bad outcome and is to that extent blameworthy, but potential blamers need to make sure that there is enough space for them to be fitting blamers of her. Debbie is the fitting target of blame, but it could still be unfitting *for most people* to blame her. If this is right, then we can explain her situation without saying that she is the unfitting target of a quantity of blame instances *as such*.

The standing view insists that insofar as there is something problematic about Debbie's case, this is because not all the instances of blame that are directed at her are fitting in the first place. The idea is that the fittingness of an individual instance of blame depends, *inter alia*, on the number of blame instances of which someone is already a target.[10] One immediate problem is that whether someone has moral standing to blame typically depends on the contents of their character, abilities, or quality of will—in short, who they are as people—as well as on facts about how they have been affected by the actions of the blameworthy (Todd 2017). This explains why people that have produced a bad outcome may lack the moral standing to blame someone else for producing that same outcome—or for producing an outcome of the same type. Put simply, they are hypocrites.

[9] For an excellent discussion about the notion of moral standing, see Todd (2017).

[10] This version of the standing view takes the blame game to be like a game of musical chairs. There are a limited number of chairs placed in a circle, each of which represents a place of moral standing that can enable its occupant to be a fitting blamer. Music is played and people dance around the chairs, following one another until a moral transgression is committed. As soon as Debbie produces a bad outcome, the music stops, and people scramble for seats. Only those that manage to occupy a chair find themselves with the moral standing to blame Debbie for producing the bad outcome. Those that find themselves without a seat cannot adopt an attitude of blame at Debbie, on pain of doing so unfittingly. Of course, one difference is that when the actual case involves thousands of anonymous strangers, as may easily happen online, it can be difficult to ascertain with accuracy who has moral standing. In fact, the number of places of moral standing could even be *indeterminate*.

Given this perspective, it would be surprising if someone could lack moral standing to blame Debbie simply for arriving too late to the blame game. The order in which someone finds themselves among other blamers does not necessarily say anything about their character, abilities, or quality of will—nor about how they have been affected by the actions of the blameworthy. Of course, this points in the direction of a version of the standing view that puts more emphasis on how people are personally affected by bad outcomes. In many cases where we might be given pause by the quantity of blame that is directed at individuals, this might be because there is a piling on from outsiders that do not have anything to do with the case at hand. Most of the blame might come from people that have not been personally affected by the relevant outcomes produced by the blameworthy (cf., Radzik 2011).

While this may capture *many* examples of the relevant type of problem case, it is unclear whether it applies to all—especially if the relevant moral transgressions can be plausibly understood as contributing toward some structural injustice that affects a large portion of the moral community. For example, it seems perfectly conceivable that the amount of people that were in fact fitting blamers of Sacco is still sufficient to give us pause and a sense that not everything is as it should be regarding the quantity of blame that ends up being directed at her. Nevertheless, the standing view might get *something* right in that it refuses to assign a negative normative status to a collection of blame instances. Perhaps we should at least consider whether there is room for other accounts that manage to be just as individualistic.

An alternative suggestion makes use of the aforementioned distinction between two types of fittingness. Recall that attitudes are made fitting to targets in the congruence sense by the properties of attitudes and the properties of their targets.[11] For example, the explanation for why love is the fitting response to lovable people has to do with the nature of love as well as the nature of the people (i.e., they have certain descriptive characteristics that *make* them lovable) (Howard 2019). Wider considerations about the effects of loving someone do not influence whether that person is a fitting target of love in the congruence sense (for they do not say anything about whether the people fulfil the inherent standards of love), but it might still be relevant to the question whether it is morally right to love them. To answer this question, we need to take a wider view of love and its *consequences*.

If these sorts of observations hold for blame as much as they do for love, then there is room for *the wrongness view*. It states that while each individual instance of blame that is directed at Debbie could be fitting, some of them may nevertheless be morally wrong, all things considered. The account therefore tries to explain the problem case in individualistic terms but does so while avoiding appeals to the notion of moral standing.[12] Personally, it also seems to me intuitive to suggest that

[11] For more on the explanation of the fittingness of attitudes and its implications for the relation between the fittingness of attitude and value, see, e.g., Orsi & Garcia (2021, 2022).

[12] The standing view and the wrongness views are reminiscent of the axiological theory *conditionalism*. While the two former views state that whether an attitude is fitting or right can depend on how many

while a person may indeed be a fitting blamer, it can nevertheless be morally wrong for them to adopt an attitude of blame, all things considered. That there can be such cases seems an attractive notion irrespective of its use in the current context. The question then is whether an account like this can capture everything there is to say about cases where too much blame is heaped on individuals.

One potential worry is that, insofar as individual instances of blame appear morally problematic, this is at least *sometimes* because of the contribution they have to a collection. The reason it is wrong for me to blame Debbie is that I would then be contributing to an even larger quantity of blame instances, and it is *this* that should give us pause. This means that when the wrongness view attempts to explain the relevant problem case by appealing to the normative status of individual blame instances, it risks putting the cart before the horse. While this is not a devastating problem, it at least invites us to consider accounts that do not deny the normative intuition from which we proceeded. We should at least be open to a more collectivist view of the relevant problem case. Let us therefore return to the unfittingness view and consider some of its advantages to the individualistic accounts.

Holistic Understandings

Let us remind ourselves of the general features of the unfittingness view. It allows that while Debbie may have produced bad outcomes for which she is the fitting target of blame, her overall character can still make her the unfitting target of the collective blame that ends up being directed at her.[13] The sense in which her character makes her the unfitting target of a large amount of blame instances is wider than the sense in which she nevertheless fulfils the inherent standards of blame. However, it is not meant to be so wide as to take into consideration all the morally relevant effects that might follow from her becoming the knowing target of so many blame instances. Debbie becomes the victim of a kind of collective harm as a result

other attitudes of that type are already held, conditionalism states that whether an object has value can depend on how many other objects of that type already exist. More precisely, conditionalism states that the value of objects depends on context, so that the same object can be good in one situation and bad in another. Crucially, this is supposed to hold for non-instrumental values as well, meaning that whether an object is good or bad *for its own sake* could also depend on the context in which it occurs, including how many other objects of that type already exist. For discussions, see, e.g., Korsgaard (1983), Kagan (1998), Hurka (1998), Dancy (1993, 2000, 2003, 2004), Rabinowicz & Rønnow-Rasmussen (2001), Olson (2004), and Orsi (2015).

[13] The unfittingness view also shares certain similarities with a more general view in axiology, namely *organicism*. While the unfittingness view states that a collection of attitudes can be unfitting even if each attitude within the collection is fitting, the latter states that we cannot calculate the values of wholes on the basis of the values of their parts (cf., G. E. Moore 1993/1903: 79). Typically, organicism also insists that the values of parts are not sensitive to context in the way that conditionalism claims—at least if said values are meant to be something other than instrumental.

of her being the target of so many instances of blame, even if she remains unaware of the fact and does not suffer emotionally or psychologically.

The suggestion that there is this wider notion of fittingness starts to look particularly attractive when we consider the connection between reactive attitudes and *dispositions*. It seems plausible to suggest that the attitude of blame correlates with dispositions to act in certain ways *vis-à-vis* the blamed. For one thing, a person that blames another will typically be disposed to treat them in ways that are punitive or that otherwise increases the social distance between the one who blames and the one who is blamed. Indeed, this seems plausible even in the event that reactive attitudes cannot be reduced to dispositions. If this is on the right lines, then the unfittingness view could be supported by an argument stating that a person can be the unfitting target of too many such dispositions, even if they should result from individual instances of blame that are all fitting. My argument is that it is a bad thing if someone with a good moral character is put in a situation of *social fragility*.

Suppose that Debbie has produced a slightly bad outcome but that she remains a morally good person overall. For the most part, she acts toward others in ways that are kind, caring, and compassionate. While she fulfils the inherent standards of blame, Debbie has a history and moral character that makes her undeserving of a situation where thousands of strangers are disposed to act toward her in ways that are punitive, or that would increase the social distance between her and the rest of the moral community. I submit that there is a sense in which such a state of affairs, where Debbie is put in a situation of social fragility, would not be fitted to her. This is so irrespective of whether she "breaks" in the sense that the dispositions are ever realized. As this notion of fittingness is guided by our general sense of what is good and right, perhaps we can elaborate on it by invoking this sense more directly. Let us briefly consider how a collection of blame instances can be bad.

If an attitude is fitting, then the adoption of the attitude by someone is typically good. Indeed, the fact that a fitting attitude is adopted is itself the fitting target of a positive attitude, such as respect and admiration.[14] I would argue that while each instance of blame that is directed at Debbie may be fitting and its adoption would to that extent be good, the fact that so many instances of fitting blame are directed at her is nevertheless bad *overall*. This means that while it may be fitting and right for anyone that contemplates Debbie and her actions to adopt an attitude of blame toward her, *pace* the standing view and the wrongness view, it may also be fitting for them to lament the collection of blame instances to which they would thereby be

[14] This is not to suggest that the adoption of fitting attitudes is always good *overall*. Suppose an Evil Demon threatens to destroy the universe if people should ever condemn him for his destructive tendencies. It would be fitting to condemn the evil demon for his destructive tendencies (i.e., he fulfils the standards that are inherent to the attitude of condemnation) and yet, condemning the evil demon would be bad overall. Nevertheless, it remains intuitive to suggest that, typically, the adoption of attitudes in cases where they are fitting is to that extent (i.e., *pro-tanto*) good, say, in the sense of being *admirable* or *respectable*. This is so even when the attitudes happen to be of the kind that draws more unpleasant associations, such as envy and hate.

contributing. The unfittingness view can in this way be expanded upon by incorporating a "holistic" view of the values accruing to attitudes.

What is attractive about this elaboration is that it also provides a practical answer to the question what people that are contemplating Debbie and her actions should do. The standing view and the wrongness view provide practical advice as well, but their advice seems incorrect in the relevant type of problem case. Some of them suggest that, depending on the *order* in which people find themselves among other blamers, they might find that they should avoid blaming Debbie even if she is morally responsible for a bad outcome and thus fulfils the inherent standards of blame. By contrast, the unfittingness view suggests that while each person that contemplates Debbie's actions is right to blame her, it could also be fitting for them to direct a negative attitude at the collection of blame of which she is a target—perhaps with an appreciative eye toward her history and moral character.

One objection takes aim at the generality of the unfittingness view. Suppose that Kate is a morally bad person, responsible for having committed a very slight moral transgression. She stole a pencil from someone's desk, say, or threw a pebble at a squirrel and missed. This was caught on film and went viral on social media, with the result that millions of people ended up blaming Kate fittingly. The sheer quantity of people that blame Kate may still seem excessive. She does not deserve so many instances of blame, but not for the reason that she has an overall good moral character. Kate is an overall bad person.[15] I am not entirely sure what to say about this sort of case, though I am tempted to suggest that while Kate has an overall bad character, there is a sense in which the actions for which she is blamed is not indicative of this character. Be that as it may, I also wish to stress that the unfittingness view states that a quantity of blame instances can be unfitting to a certain person, but perhaps it can be liberal when it comes to the explanation of this. One commonplace explanation has to do with the overall moral character of the would-be blamed (this is the case when it comes to Debbie) but other explanations may be required to capture our intuitions about other examples (about Kate).

Another objection to the unfittingness view maintains that, unlike its individualistic rivals, this view does not fit well with commonplace ideas about how fittingness should be understood. One such idea is that notions of fittingness, whether they are of the narrow or wide variety, should be understood in terms of *normative reasons*. This means that whether an attitude is unfitting is ultimately a matter of whether there are normative reasons for some agent to avoid adopting the attitude. This is a problem for the unfittingness view because when we say of a quantity of blame instances that *it* is unfitting in respect of a target like Debbie, it seems that we cannot understand this suggestion in terms of there being some agent that has normative reasons to avoid adopting that quantity of attitudes.

The theoretically expensive choice would be to respond by insisting that there is a *collective entity* for whom there can be reasons to avoid adopting a collection of

[15] I owe this objection to Mattias Gunnemyr, who raised it to me (in private communication).

attitudes, but this appears too costly in the present context.[16] Instead, we might suggest that there are notions of fittingness that cannot be understood in terms of normative reasons in this way. Among other things, these notions allow us to say that Debbie is, by virtue of her history and moral character, the unfitting target of a certain quantity of attitudes, even if there is no one agent that has reasons to avoid adopting any attitudes within the quantity. The manoeuvre would have seemed *ad hoc* if not for the fact that several philosophers have put forward good reasons to think that the fittingness of responses is not generally reducible to normative reasons for responses (e.g., McHugh & Way 2016, 2022).[17]

Another objection states that if blame works in the way that I have suggested, then this would make it unique from other kinds of reactive attitudes. After all, there is nothing unfitting about Debbie becoming the target of large amounts of love, respect, or admiration—unless she was not the fitting target of such reactive attitudes to begin with. I am not convinced that this is true. For example, praise seems to me to mirror the behaviour of blame in this regard. The amount of praise that Debbie becomes the target of could be excessive given the weight of the value that she has produced. There may also be cases where other kinds of reactive attitudes, including admiration, exhibit this sort of behaviour. Personally, I often find that the amount of admiration that is heaped over popular songs and movies is not as it should be—even when the individual instances of admiration happen to be fitting in terms of their intensity and duration.

Perhaps we should also be open to the possibility that there is something special about the attitudes of blame and praise, which results in them becoming subject to these wider fittingness conditions. The question why this might be is a difficult one to which I regrettably have no answer. Perhaps it has something to do with the distinctive roles and social functions that moulded the attitudes of blame and praise in the first place. In the absence of an explanation, there is of course the temptation to be sceptical and insist that quantities do *not* play a role for the fittingness of any reactive attitudes—though they may play a role for determining whether it is fitting to express them. The weight of being a knowing recipient of too much blame is heavy, as is the weight of being constantly reminded of one's moral transgressions. It is certainly easy to explain the salience of *such* factors.

---

[16] There is a wealth of literature on the topic of collective agency going back to the first half of the 20th century. For a small selection of recent and insightful discussions, see, e.g., Held (1970), Gilbert (1989, 2000, 2006, 2013), Tuomela (1989, 2005, 2006, 2013), Smiley (1992), Velleman (1997), Bratman (1999, 2013), Kutz (2007), List & Pettit (2011), and Toumela & Mäkelä (2016). I wish to note that the collective entity that I am here referring to is very unlikely to be integrated enough to be considered a collective agent, even on the most generous accounts of what this means. So, for this type of theoretically expensive account to work, it would have to invoke the idea that not only agents can be subject to reasons. I owe thanks to Mattias Gunnemyr for pressing this issue (in private communication). Unfortunately, I will have to set it aside here.

[17] The assumption is that unless we understand fittingness in terms of normative reasons, then we must understand it in terms of another normative category, like value. We should be neutral here although I am tempted to suggest, as McHugh & Way do (2016), that fittingness is a primitive notion.

In particular, the effects of expressing otherwise fitting attitudes have a big impact on the decision whether to make one's private blame of an individual public. I have so far treated blame as a non-doxastic and motivational mental state that is directed at people for the bad outcomes that they produce. When we talk about blame in everyday life, however, we often have in mind something distinctly public, like acts of punishment as well as written or spoken speech-acts (e.g., "How could you, Debbie!").[18] Sceptics to the unfittingness view might dig in their heels and insist that if we look to quantities of blame and keep in mind the attitudinal understanding, then there can be no question about the unfittingness of such quantities. An agent simply cannot be the unfitting target of a collection of blame instances *as such*, regardless of the kinds of dispositions with which they are then subjects.

Fair enough. The paper proceeds from the normative intuition that there could be something about a collection of blame instances that is not as it should be. What I have argued is that the approach that is most likely to capture the intuition well is something like the unfittingness view. The reason is that it does not explain the normative status of a collection of blame instances by giving up on the idea that the individual instances within it are all morally unproblematic. In other words, if one senses the pull of the intuition, then there are reasons to prefer something like the unfittingness view. That said, I acknowledge that if one does not sense the pull of the intuition—perhaps because one harbours convictions about the reducibility of fittingness to normative reasons—then something like the standing view and the wrongness view may be preferable.

Concluding Remarks

I have suggested that just as the weight of the value that people produce can become relevant to the duration and intensity of the reactive attitudes of which they are the fitting targets, it can be relevant to the amounts of reactive attitudes of which they are the fitting targets as well. The suggestion was meant to help explain and vindicate our reaction to cases where people have said or done something bad and thereby made themselves the targets of disproportionate quantities of reactions. I used the example of online blame as a paradigmatic case where this is a risk. The purpose of the paper was to explore the implications of the intuition just mentioned, and to outline some of the theoretical costs that are associated with its various interpretations. I suspect that further research into this area could yield a better understanding of the social dimensions of moral blame.[19]

[18] See the discussions about private and public blame between McKenna (2012, 2016) and Driver (2016). For more on the normative stakes that are associated with these, see also Gokhale (2019).

[19] This research was funded by the Swedish Research Council, grant number: 2018-06612.

References

Andersson, H., & Werkmäster, J. G. (2020). How Valuable Is It?. *Journal of Value Inquiry, 55*(3), 525–542.

Aitchison, G., & Meckled-Garcia, S. (2021). Against Online Public Shaming: Ethical Problems with Mass Social Media. *Social Theory & Practice, 47*(1), 1–31.

Billingham, P., & Parr, T. (2020). Online public shaming: Virtues and vices. *Journal of Social Philosophy, 51*(3), 371–390.

Bratman, M. (1999). *Faces of Intention: Selected Essays on Intention and Agency*. New York: Cambridge University Press.

——— . (2013). *Shared Agency: A Planning Theory of Acting Together*. Oxford: Oxford University Press.

Coates, D. J., & Swenson, P. (2013). Reasons-responsiveness and degrees of responsibility. *Philosophical studies, 165*(2), 629–645.

Dancy, J. (1993). *Moral Reasons*. Oxford: Blackwell Publishing.

——— . (2000). The Particularists Progress. In: B. Hooker, M. O. Little (Eds.) *Moral Particularism*. Oxford: Oxford University Press, 130–156.

——— . (2003). Are There Organic Unities?. *Ethics, 113*(3), 629–650.

——— . (2004). *Ethics Without Principles*. Oxford: Oxford University Press.

D'Arms, J., & Jacobson, D. (2000). The moralistic fallacy: On the 'appropriateness' of emotions. *Philosophy & Phenomenological Research, 61*(1), 65–90.

Driver, J. (2016). Private blame, *Criminal Law and Philosophy, 10*(2), 215–220.

Gilbert, M. (1989). *On Social Facts*. New York: Routledge

——— . (2000). *Sociality and Responsibility*. Lanham, MD: Rowman and Littlefield

——— . (2006). Who's to blame? Collective moral responsibility and its implications for group members, *Midwest Studies in Philosophy, 30*(1), 94–114.

——— . (2013). *Joint Commitment*. Oxford: Oxford University Press.

Gokhale, S. (2019). Who Needs Blame?: *Answerability Without Expressed Blame*. CUNY Academic Works.

Held, V. (1970). Can a random collection of individuals be responsible?, *Journal of Philosophy, 67*(14): 471–481.

Howard, C. (2019). Fitting love and reasons for loving. In: Timmons, M. (Ed.), *Oxford Studies in Normative Ethics Volume 9*. Oxford: Oxford University Press.

Hurka, T. (1998). Two Kinds of Organic Unity. *The Journal of Ethics, 2*(4), 299–320.

Kagan, S. (1988). The Additive Fallacy, *Ethics, 99*(1), 5–31.

Korsgaard, C. (1983). Two Distinctions in Goodness. *Philosophical Review, 92*(2), 169–195.

Kutz, C. (2007). *Complicity: Ethics and Law for a Collective age*. Cambridge: Cambridge University Press.

List, C., & Pettit, P. (2011). *Group Agency: The Possibility, Design, and Status of Corporate Agents*. Oxford: Oxford University Press.

McHugh, C., & Way, J. (2016). Fittingness first. *Ethics, 126*(3), 575–606.

——— . (2022). Getting things right: Fittingness, reasons, and value. Oxford University Press.

McKenna, M. (2012). *Conversation & Responsibility*. New York: Oxford University Press

——— . (2016). Quality of will, private blame and conversation: Reply to Driver, Shoemaker, and Vargas, *Criminal Law and Philosophy, 10*(2), 243–263.

Moore, G. E. (1993/1903). *Principia Ethica*. Cambridge: Cambridge University Press.

Nelkin, D. K. (2016). Difficulty and degrees of moral praiseworthiness and blameworthiness. *Noûs, 50*(2), 356–378.

Olson, J. (2004). Intrinsicalism and Conditionalism about Final Value. *Ethical Theory & Moral Practice, 7*(1), 31–52.

O'Neill, J. (1992). The Varieties of Intrinsic Value, *The Monist 75*(2), 119–137.

Orsi, F. (2015). *Value Theory*. London: Bloomsbury.

Orsi, F., & Garcia, A. G. (2021). The explanatory objection to the fitting attitude analysis of value. *Philosophical Studies, 178*(4), 1207–1221.

——— . (2022). The new explanatory objection against the fitting attitude account of value. *Philosophia*, 1-16.

Pettit, P. (2007). Responsibility incorporated, *Ethics, 117*(2), 171–201.

Rabinowicz, W., & Rønnow-Rasmussen, T. (2000). A Distinction in Value: Intrinsic and for its own sake. *Proceedings of the Aristotelian Society, 100*(1), 33–51.

——— . (2004). The Strike of the Demon: On Fitting Pro attitudes and Value. *Ethics, 114*(3), 391–423.

Radzik, L. (2011). On Minding Your Own Business: Differentiating Accountability Relations within the Moral Community, *Social Theory and Practice, 37*(4), 574–598.

Ronson, J. (2015) How One Stupid Tweet Blew Up Justine Sacco's Life. *The New York Times Magazine*. Retrieved from URL=http://www.nytimes.com.

Smiley, M. (1992) *Moral Responsibility and the Boundaries of Community*. Chicago: University of Chicago Press.

Strawson, P. (1962). Freedom and resentment, *Proceedings of the British Academy, vol.48*: 1–25.

Tierney, H. (2019). Quality of reasons and degrees of responsibility. *Australasian Journal of Philosophy, 97*(4), 661–672.

Todd, P. (2019). A unified account of the moral standing to blame. *Noûs, 53*(2), 347–374.

Tosi, J. & B. Warmke. (2020). *Grandstanding: The Use and Abuse of Moral Talk*. Oxford University Press.

Tuomela, R. (1989). Actions by collectives, *Philosophical Perspectives, vol. 3*, 471–496

——— .(2005). We-intentions revisited, *Philosophical Studies, 125*(3), 327–269

——— . (2006). Joint intention, we-mode and I-mode, *Midwest Studies in Philosophy*, *30*(1), 35–58.

——— . (2013). *Social Ontology*. New York: Oxford University Press

Tuomela, R., & Mäkelä, P. (2016). Group agents and their responsibility, *The Journal of Ethics, 20*(1–3), 299–316.

Velleman, D. (1997). How to share an intention, *Philosophy & Phenomenological Research, 57*(1), 29–50.

Harming Others

Mattias Gunnemyr

Abstract. There are two standard accounts of what it is to harm someone: the counterfactual comparative account and the non-comparative account. The first gives counterintuitive verdicts in cases of overdetermination and pre-emption, and the second implausibly entails that an event might harm you even though it makes you better off. On some interpretations, the non-comparative account also gives counterintuitive verdicts in switching cases. I suggest that we can combine these accounts in a way that avoids giving the mentioned counterintuitive implications. Roughly, the suggestion is that you harm someone, *S,* if (a) there is a genuine process connecting what you did to *S*'s being worse off (a non-comparative condition), and (b) *S*'s being better off would have been more secure if you had not acted in this way (a counterfactual comparative condition).

Last week, this paper was rejected by the *Journal of Over-Determination.* According to the journal's strict policy, manuscripts are rejected if at least one of the reviewers recommend rejection and accepted if both reviewers recommend accepting the paper. Deeming from the comments I got, both reviewers considered the paper to be of high quality. However, in the end, both decided to recommend rejection. The review process was double-blinded, but out of a coincidence I later learned that it was reviewed by Björn and Dan.

I would have been better off if the paper had been accepted. I would have been happy that my paper finally found its home in a highly ranked philosophy journal, and my chances of getting tenure would have been improved. Still, it does not matter for me whether one or two of the reviewers recommended rejection. The consolation

of having one positive review would have been balanced out by the frustration of almost getting the paper published.[1]

In recommending rejection, did Björn and Dan harm me? We seem to lack a clear intuitive verdict about whether they did. On the one hand, it seems they did. After all, if neither of them had suggested to reject my paper, it would have been accepted and I would have been better off. On the other hand, it seems that neither of them harmed me. I would not have been better off had Björn recommended to accept my paper since Dan would have recommended to reject it anyway. And similarly, I would not have been better off had Dan recommended to accept my paper since Björn would have recommended to reject it anyway.

The Counterfactual Comparative Account of Harming

Can our best theories of what it is to harm someone explain these shifting intuitions? At a first glance, it seems that they cannot. Consider what I take to be the standard view in the literature on harming:

> *The counterfactual comparative account of harming (CCA):* An event *C* harms a person *S* if and only if *S* would have been better off if *C* had not occurred.
>
> (Feinberg 1984; Parfit 1984; Norcross 2005; Bradley 2009; Klocksiem 2012)

CCA is usually taken to be an analysis of when an event *C* is *all-things-considered* harmful for *S;* that is, when an event *C* makes *S* have a lower lifetime well-being level. On this reading, applying CCA to some situation might require quite laborious evaluations. For instance, in *Journal of Over-Determination,* it requires us to consider whether I as a result of Björn's and Dan's recommendations eventually would get the paper published in another even more well-regarded journal and the wellbeing level I would have then, or whether I would have to end my career in philosophy and pursue some other career entirely, and the wellbeing level that would provide me. To keep the discussion focused, I will instead take CCA to be an analysis of what it is for an event *C* to be *pro tanto* harmful for *S*; that is, what it is for *C* to be harmful for *S* in some respect. This allows me for instance to consider the question of whether Björn's and Dan's recommendations harmed me in the sense of making me sad and having fewer chances of getting tenure without considering whether their recommendations also harmed me or benefitted me in some other way. Still, the arguments I give here could be amended to apply to the standard reading of CCA using slightly modified examples, so using this non-standard interpretation of CCA should not make any difference for the conclusions I draw.

[1] I will call this case *Journal of Over-Determination.* The case is a modified version of an example discussed by Petersson (2018). For further discussion of this example, see e.g. Johansson (this volume).

According to CCA, then, neither Björn's nor Dan's recommendation to reject my paper harmed me. Because of Dan's recommendation, I would not have been better off (happy and with better chances of getting tenure) than I am now (sad and with lesser chances of getting tenure) had Björn not decided to recommend rejection, and because of Björn's recommendation, I would not have been better off had Dan not recommended rejection. Thus, CCA can straight-forwardly explain one of the intuitions we might have about the case: the intuition that Björn's recommendation to reject did not harm me since the paper would have been rejected anyway (and that a similar thing could be said about Dan's recommendation). Yet, this verdict seems less than satisfactory. We have not yet found an explanation for the other intuition: that I was harmed as a result of their recommendations.

Upon closer scrutiny, however, there is a close cousin to CCA that seems to go some way to explain this intuition. This alternative principle does not apply to events, but to *sets of events* (Parfit 1984) or *pluralities* (Feit 2015, 2016). A plurality can be understood as several individual events taken together, so when a plurality harms someone, there are several events such that they harm this person.[2]

> *The plural harm principle (PH):* A set of events – or a plurality – *C* harms a person *S* if and only if *S* would have been better off if *C* had not occurred.[3]

According to Feit and Parfit, PH entails that the two recommendations harmed me: Had not this set or plurality occurred, I would have been better off. That is, had not Björn's *and* Dan's recommendations to reject occurred, they would both have recommended the journal to accept my paper, and I would have been better off. Moreover, PH retains some of CCA's implications: it entails that neither Björn's nor Dan's recommendation harms me. It is still the case that had Björn's recommendation to reject (we can think of this as a one-event plurality) not occurred I would not have been better off, and the same goes for Dan's recommendation.

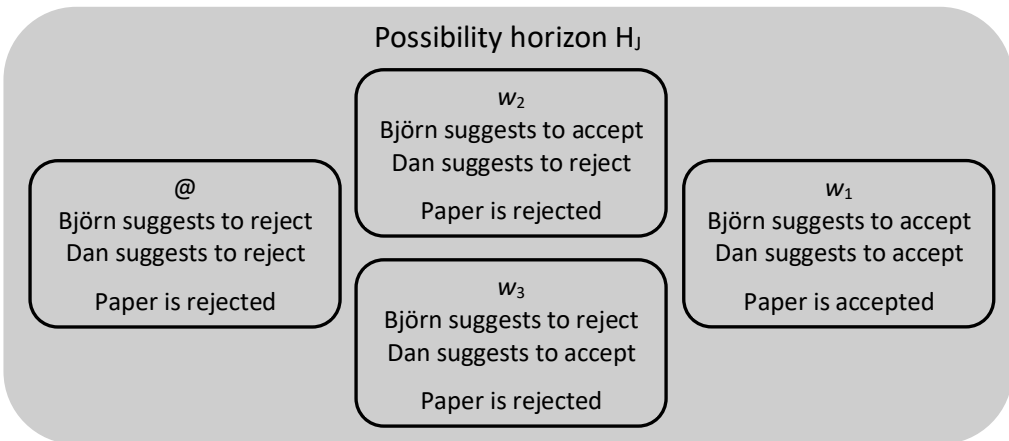
Still, it is far from clear that Parfit and Feit are correct that PH implies that the set or plurality consisting of Björn's and Dan's recommendations harmed me. As Petersson (2004, 2018) argues, when we are deciding whether someone would have been better off had a certain set of events not occurred, we should consider what happens in the closest possible world where this set does not occur. In a case like *Journal of Over-Determination* in which the relevant actions are counterfactually independent of each other, the closest possible world where the set consisting of Björn's and Dan's recommendations does not occur is a world where one of these recommendations still occurs. It might be the world where Björn but not Dan

[2] Feit (2015) does not make entirely clear what a plurality is. For discussion, see Petersson (2018: 846-7).

[3] Parfit (1984) and Feit (2015) place a further restriction on what it is for a set (or plurality) of events to harm someone: A set of events harms A if and only if that set is *the smallest set* such that, if none of the events had occurred, A would have been better off. They include the extra restriction to avoid the contra-intuitive result that completely unrelated events, such as Fred Astaire's dancing in the distance, also belong to a set that harmed me. Here, I will set aside this complication.

recommends rejection, or the world where Dan but not Björn recommends rejection. Hence, if the set consisting of Björn's and Dan's recommendations had not occurred, one of these recommendations would still have occurred, and I would not have been better off than I would have been if both had recommended rejection. So, while it might seem that PH entails that Björn's and Dan's recommendations together harmed me, this verdict is upon closer reflection mistaken.[4]

This point might need some elucidation. To get any counterfactual analysis to deliver plausible results, we need to decide which possible worlds are relevant. As Norcross (2005), Klocksiem (2012) and others point out, the relevant worlds are often provided by context. In the case under consideration, it is not made explicit what would happen in the closest possible world where Björn does not recommend rejection. It is assumed from context, I take it, that he would recommend accepting the paper. Still, upon reflection, he might do other things. He might for example contact Dan and persuade him that the paper deserves to be published, with the result that they both recommend the journal to accept the paper. In that case, CCA and PH would not have the implications we thought they had. They would entail that Björn's recommendation to reject harmed me: I am better off in the closest possible world where he does not recommend rejection (the world where he contacts Dan). This open-endedness will not do. To be able to evaluate any counterfactual analysis, we need to decide which possible worlds are relevant. Petersson's argument (and Parfit's and Feit's arguments, for that matter) presupposes that we have done so already. In *Journal of Over-Determination,* for instance, it is implicitly assumed that there are four relevant possible worlds: the actual world (@) where both Björn and Dan recommends rejecting the paper, the possible world where none of them recommends rejection (w_1), and the possible worlds where one of them but not the other suggest rejection (w_2 and w_3). This is the *possibility horizon* relevant for the case as Touborg (2018) would say, inspired by what Mackie (1974) calls a "causal field".



[4] Like Parfit and Feit, Jackson (1997) also mistakenly concludes that the relevant set of actions is harmful in cases of overdetermined harm.

Note that this possibility horizon does not include possibilities not alluded to in the original description of the case. It does for instance not include the possible world where Björn contacts Dan and persuades him to recommend the journal to accept my paper.

Petersson's argument implicitly assumes a possibility horizon like H_J, and it can be reconstructed as follows: Since Björn's and Dan's recommendations are counterfactually independent of each other, the relevant possible world closest to @ is not w_1, but w_2 or w_3, depending on whether Björn or Dan was closer to recommending the journal to accept the paper. Therefore, my paper had been rejected even if not both had suggested to reject it. It had been rejected since one of them still had recommended rejection. We can then conclude that PH applied to this possibility horizon yields the counter-intuitive result that the set consisting of Björn's and Dan's recommendations did not harm me. We can also conclude that while PH can explain the intuition that Björn's recommendation did not harm me since my paper would have been rejected anyway (and that an analogous thing may be said about Dan's recommendation to reject), it fails to explain the intuition that their recommendations harmed me – just like CCA does.

A defender of the idea that the set or plurality consisting of Björn's and Dan's recommendations harms me might argue that the relevant comparison is not the one between the actual world where both of them recommend rejection (@) and the world in which only one of them does (w_2 or w_3), but the one between the actual world and the world in which none of them does (w_1). This is the position that Parfit and Feit take. Parfit states that we should consider what happens "if they had all acted differently" (1984: 71) and Feit likewise says that we should consider what happens "if none of [the relevant events] had occurred" (2015: 371).

This does not help much. First, as Petersson (2018) points out, "the relevant counterfactual comparison is given by the case-description" (846), and if the principle we use requires that we make another counterfactual comparison than the one given by the case, it is not applicable. In *Journal of Over-Determination,* it is given by the case-description that in the closest relevant possible world where Björn does not recommended rejection, Dan still does (and vice versa), which entails that in the closest relevant possible world where the set consisting of two recommenddations to reject does not occur, one of the reviewers still recommends rejection. The requirement that we take the world where they both act differently as the relevant one makes PH inapplicable to this case, or at least – we might add –insists that we ignore essential features of the case.

Second, even if we agree with Parfit and Feit that the relevant possible world for comparison is the one where neither Björn nor Dan recommend rejection, and that this principle is applicable to the cases like *Journal of Over-Determination,* we still end up in an unattractive position. While we get the result that the set consisting of their recommendations harmed me, we also get the result that neither Björn's nor Dan's recommendation harmed me. Neither Parfit nor Feit take this to be an unattractive position. Parfit says that it might still be wrong of me to perform some

act if this act is part of a harmful set of acts even though my act harms no one. Feit says (considering structurally similar cases) that while we cannot say that Björn's recommendation or Dan's recommendation harmed me, we can still say that each of them was *involved* in harming me. Still, it seems attractive to be able to say that Björn's recommendation was one the acts that harmed me, and likewise for Dan's recommendation. After all, each recommendation was sufficient for bringing this outcome about.[5]

In a way, Petersson (2018) takes one step further than Parfit and Feit do. He argues that we should hold on to CCA and conclude that neither Björn's recommendation, Dan's recommendation, nor the set of events consisting of both recommendations harmed me.[6] This conclusion seems counterintuitive, and I think we can do better than that.

A Better Account

If we consider *Journal of Over-Determination* more closely, we see that my paper is closer to getting accepted in the possible worlds w_2 and w_3 where either Björn or Dan suggests accepting my paper than in the actual world @ where both of them suggest rejecting it. I am not better off in these other worlds than in the actual one – the consolation of having one positive review is balanced out by the frustration of almost getting the paper published – but there is a difference in terms of closeness to the world where I am better off. We can say that the rejection of my paper is more *secure* when both Björn and Dan recommend rejection, than it is when only one of them does. Similarly, the acceptance of my paper is more secure – it is closer to happening – in the worlds where only one of them recommends rejection than it is in the world where both do. We can use this insight to construct a more accurate analysis of what it is to harm someone:

> *The security comparative account of harming (SCA):* An event *C* harms a person in H if and only if *S*'s being better off in this respect would have been more secure in H had *C* not occurred.[7]

Here, H stands for "possibility horizon". In relativising what it is for an event to harm someone to H, this principle makes explicit the idea that we must specify the relevant possible worlds before we apply our analysis.

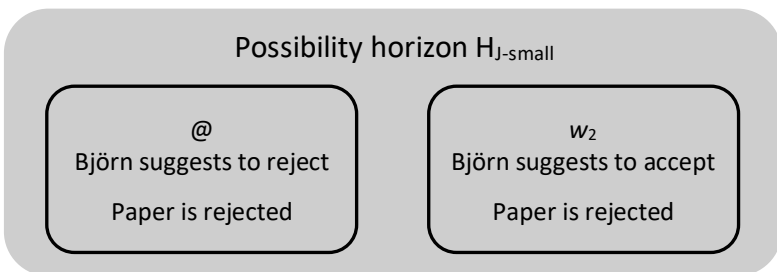
[5] PH faces other problems as well. See e.g. Johansson and Risberg (2019) and Johansson (this volume), especially his discussion on premise 1.

[6] Petersson (2018: 848-9) gives some reasons why this seemingly counterintuitive position might not be so counterintuitive after all. For brevity, I omit discussing these reasons.

[7] In this paper, SCA is meant to evaluate whether an event *C* is *pro tanto* harmful for *S.* However, the arguments I give here could be amended to apply to a reading of SCA in terms of overall harm.

SCA entails that Björn's recommendation to reject the paper harmed me, given H_J. Even though I would not have been better off in the closest-to-@ world where Björn did not recommend rejecting it, my being better off is more secure had he not recommended rejection. Likewise, given H_J, SCA entails that Dan's recommendation to reject harmed me. Moreover, if we allow for SCA to be applied to sets (or pluralities) of events, it also entails that the set consisting of Björn's and Dan's rejection harmed me. Had this set not occurred, I would not have been better off, but my being better off had been closer to occurring. Even without allowing for SCA to be applied to sets (or pluralities), we might still say that there is a set (or plurality) of events that harms me: each of Björn's and Dan's recommendations harms me, and together they constitute a set of events where the each event in that set individually harms me.

SCA can also explain our shifting intuitions about whether Björn and Dan harmed me. As argued above, if we assume that H_J captures the relevant possibilities in the case, SCA entails that each of the recommendations harmed me, and possibly that the set of recommendations harmed me. However, there is another way of understanding the case. We might think that I would not have been better off had Björn decided to suggest the journal to accept my paper since Dan decided to reject it anyway. That is, given that Dan decided to reject the paper, Björn's decision does not matter for whether the paper is rejected or not. There is a natural reading of this idea in terms of which possibilities that are relevant. When we say things like "given that Dan decided to reject the paper", or "Dan decided to reject it", it seems that we are not treating it as an open possibility that Dan could have acted otherwise. That is, we seem to treat Dan's decision as a background condition rather than a potential cause of harm. If we do, we do not really treat the case as involving four relevant possibilities, but only two: that Björn either decides to suggest rejection or decides to suggest the journal to accept the paper. As a result, we get a much smaller possibility horizon, as follows:



When applied to this smaller possibility horizon, SCA entails that Björn's recommendation did not harm me. The rejection of the paper is as secure in w_2 as it is in @. This point might need some elaboration. We can think about the degree of

security in terms of distances between possible worlds. The security of an outcome in a certain possible world is the distance between this world and the closest possible world where this outcome does not occur. To decide the security of the outcome that the paper is rejected in @, we thus have to look at the distance between this world and the closest-to-@ world in $H_{J\text{-small}}$ where this outcome does not occur. However, there is no such world in $H_{J\text{-small}}$. If we start out from @ and travel out into the modal space containing only relevant possible worlds, we will never encounter a world where the paper is not rejected. Therefore, we might say that the distance in question is infinite, and that the outcome that the paper is rejected in @ is infinitely secure. The same goes for the security of the outcome that the paper is rejected in w_2. Hence, the security of the outcome that the paper is rejected is the same in @ as it is in w_2, namely infinitely secure.

We can thus understand the two intuitions as stemming from different possibility horizons. The intuition that Björn and Dan harmed me originates in the larger possibility horizon where we treat it as an open possibility that one or both of them could have recommended otherwise, and the intuition that neither of them harmed me originates in smaller possibility horizons such as $H_{J\text{-small}}$. If we take it as a given that Dan will recommend rejecting my paper, it seems that Björn's recommendation did not matter for whether I was harmed or not. Likewise, if we treat Björn's recommendation as a background condition, Dan's recommendation seems to make no difference for whether I am harmed or not.[8]

This helps us see where CCA goes wrong (which it does, *pace* Petersson 2018). On a standard understanding, CCA tells us to compare the actual world with the closest-to-@ world where *C* does not occur. In effect, it tells us to only take into consideration possibility horizons such as $H_{J\text{-small}}$, making us blind to the fact that there are worlds further away that might be relevant in the evaluation. Granted, CCA (on its standard interpretation) might allow for considering larger possibility horizons such as H_J. It does not exclude the idea that there is a possibility that for example both Björn and Dan would recommend accepting my paper. Still, even if we take such larger possibility horizons into consideration, CCA promptly tells us that the comparison relevant for determining whether *C* harms *S* is the one between the actual world @ and the closest-to-@ world where *C* does not occur. Thereby, possible worlds further away will never matter in the evaluation. Yet, if we allow them to matter contrary to the recommendations of CCA, we find plausible ways of explaining our intuitions in cases like *Journal of Over-Determination.* Similar remarks apply to PH.

[8] Norcross (2005) makes a similar suggestion. He argues that the conversational context indicates the relevant possible world for evaluating whether an event harms someone, and so that an event might be correctly described as harming in one conversational context, and correctly described as benefitting in another.

Pre-emption Cases

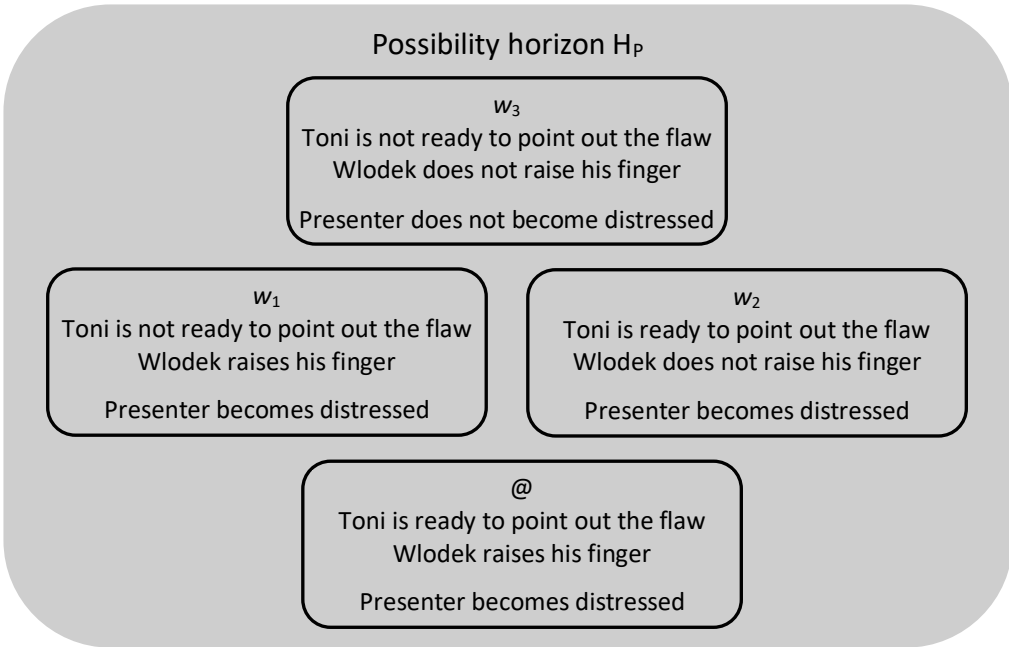
While SCA gives intuitively correct verdicts about harming in some notoriously tricky cases, it cannot be the accurate account of harming we are looking for. It gives patently erroneous verdicts in pre-emption cases, such as the following:

> *Presentation on pre-empted harm:* Toni and Wlodek are listening to a presentation on the topic pre-empted harm at the Higher Seminar. At the same time, they spot a crucial flaw in the argument. Immediately, Wlodek raises his index finger to signal to the chair that he wants to say something. Wlodek gets the word and explains the crucial flaw in the argument with the result that the presenter becomes quite distressed. Had Wlodek not raised his finger, Toni would have raised his a few moments later and explained the flaw in the argument with the result that the presenter had been just as distressed. However, when seeing Wlodek raising his finger, Toni decided not to raise his. He thought that one interruption to the presentation was enough.

Here, it seems clear that Wlodek but not Toni harmed the presenter. You might object that pointing out crucial flaws in someone's argument is not harmful but rather beneficial for this person. Doing so helps her to improve her arguments, to discard mistaken theses, and to make philosophical progress. This objection is relevant, but rests on a misunderstanding. SCA is meant to capture what it is for an event to be *pro tanto* harmful for someone, or in other words, to be harmful *in some respect.* Bearing this in mind, it seems plausible to say that Wlodek (but not Toni) harmed the presenter in one respect: Wlodek's comment made the presenter distressed during the presentation. We could agree that Wlodek's comment also benefitted the presenter in several ways, we could even agree that Wlodek's comment all-things-considered was beneficial for the presenter, and still hold on to the idea that Wlodek's comment harmed the presenter in this one way. In fact, I think this is precisely how we should understand the situation.

Now, it is clear that Wlodek but not Toni harmed the presenter in a way. Toni would have harmed the presenter had Wlodek not beaten him to it, but as things turns out, he did not. SCA, however, entails that both Wlodek and Toni harmed the presenter. To see this properly, we first have to settle the relevant possibility horizon. Consider the time at which Wlodek raises his index finger. At this time, there are four relevant possibilities, as indicated in the possibility horizon H_P (depicted on the next page).

Applied to H_P, SCA wrongly entails that Toni harmed the presenter. In the closest-to-@ world where Toni is not ready to point out the crucial flaw in the argument (w_1), the presenter still becomes distressed. However, this outcome is less secure than in the actual world. In w_1, the only thing that needs to change for the presenter not to become distressed is Wlodek raising his finger, wanting to say



something, whereas in @, Wlodek raising his finger *and* Toni's readiness to point out the flaw need to change in order for the presenter not to become distressed.
For comparison, CCA also gives inaccurate verdicts in pre-emption cases like this. While SCA entails that both Toni and Wlodek harmed the presenter, CCA entails that neither Toni nor Wlodek harmed the presenter. Toni does not harm the presenter since, even if he had not been ready to point out the flaw, Wlodek would have pointed out the flaw anyway. Wlodek, in turn, did not harm the presenter according to CCA since, had he not raised his finger and pointed out the flaw, Toni would have done so instead, and the presenter had become just as distressed anyway.[9]

Non-comparative Accounts of Harm

Seeing that comparative accounts of harming like CCA, PH and SCA run into trouble, we might be tempted to turn to non-comparative accounts, such as the following:

[9] This problem has been pointed out and discussed by e.g. Bradley (2012) and Johansson and Risberg (2019).

> *Non-comparative account of harming (NCA):* An event harms someone if it causes the person to be in an intrinsically bad state.
>
> (see e.g. Shiffrin 1999; Harman 2009)

If we take what it is for an event to cause another at face value, NCA seems to give the right verdicts in the cases we have considered so far. It is Wlodek and not Toni who causes the presenter to be distressed, and to be distressed is an intrinsically bad state. So, NCA entails that Wlodek but not Toni harms the presenter (in this respect). Further, both Björn's and Dan's recommendations caused my article to be rejected, and by extension they caused my sadness and my impoverished chances of getting tenure (two intrinsically bad states, we might assume). Therefore, NCA entails that both Björn's and Dan's recommendations harmed me.

We might consider what kind of account of causation that would fit NCA. For a start, a simple counterfactual account of causation – sometimes called the But-For test for causation – reinvites trouble. According to this account, *C* causes *E* if and only if had *C* not occurred, *E* had not occurred.[10] It entails that neither Björn's nor Dan's recommendation caused my article to be rejected in *Journal of Over-Determination*, and that neither Wlodek nor Toni caused the presenter distress in *Presentation on pre-empted harm.* As a result, the simple counterfactual account of causation does not yield the verdicts on causation needed for NCA to give the right verdicts about harming. The problem, I take it, is that the account reinvites counterfactual comparisons to an allegedly non-comparative account of harming. If we want to keep NCA truly non-comparative, we need to couple it with some non-comparative analysis of causation.

Accounts that build on the idea of minimal sufficiency seems a better fit for NCA. An elementary version of such accounts can be stated as follows:

> *Elementary minimal sufficiency: C* causes *E* if and only if
> (i) *C* belongs to a set of actual antecedent events that guarantees, given the relevant laws, that *E* will occur, and
> (ii) if you remove *C* from the set, the set no longer guarantees that *E* will occur.[11]

Given this account of causation, NCA gives the right verdict in the cases we have considered so far. To begin with, it entails that Björn's and Dan's recommendations harmed me. Take Björn's recommendation for instance. It belonged to a set of actual antecedent events that included his recommendation but not Dan's. This set guaranteed, given the rules of the journal, that I would be sad and have few chances of getting tenure, and it did so regardless of whether Dan recommended rejection.

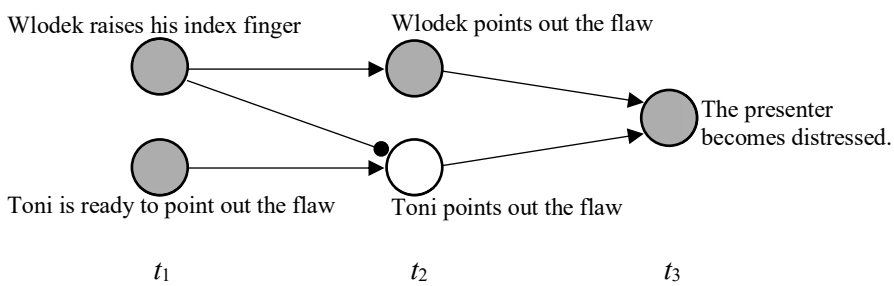
[10] Lewis (1973, 2004) presents more elaborated versions of this account.

[11] Mackie's (1974) INUS condition for causation and Wright's (1985) NESS condition for causation provides examples of such accounts.

Still, the set had not guaranteed this outcome if we had removed Björn's recommendation from it. Björn's recommendation is necessary for the sufficiency of that set. Therefore, Björn's recommendation caused me to be in an intrinsically bad state (sad, and with few chances of getting tenure), and hence NCA entails that Björn's recommendation harmed me. A parallel argument shows that Dan's recommendation harmed me.

However, *Elementary minimal sufficiency* needs to be elaborated to give the right verdict in *Presentation on pre-empted harm.* As this account is formulated now, it entails that both Wlodek and Toni caused the presenter's becoming distressed. At the time when Wlodek raised his index finger, Toni's readiness to point out the crucial flaw in the presenter's argument was minimally sufficient for the presenter's becoming distressed. It belonged to a set (not including Wlodek raising his index finger) of actual events that guaranteed, given the relevant laws, that the presenter would become distressed, and if we removed Toni's readiness from this set, it would no longer so guarantee. Still, it is obviously false that Toni caused the presenter's distress. As things turned out, Toni's readiness to point out the flaw in the argument was pre-empted before by Wlodek raising his index finger. It only guaranteed the outcome, but did not cause it.

The situation can be illustrated using a neuron diagram, where circles (or "neurons") represent events, and arrows represent causal connections. A shaded circle indicates that the neuron fires; that is, that the event occurs. The temporal reading is from left to right. If a neuron fires, it sends a signal through the connecting arrow to the right. If a neuron receives such a signal from its left, it will fire. A line ending in a black dot represents an inhibitory signal, hindering the neuron at its end point from firing.



As the diagram makes clear, events relevant for the causal evaluation of the situation occur in the intermediate time after Toni was ready to point out the flaw but before the presenter becomes distressed. At t_2, Toni decides not to raise his finger when he sees that Wlodek raises his. However, *Elementary minimal sufficiency* only takes into account what happens at time t_1 and t_3 when evaluating whether Toni's

readiness caused the presenter's distress. Following Caroline Touborg (2018), we can elaborate a more accurate account of minimal sufficiency that lets us capture what happens also at intermediate times, as follows:

Let us say that there is an *apparent process* from *C* to *E* when there is a chain of relations of minimal sufficiency. That is, when *C* belongs to a set of simultaneous events that is minimally sufficient a later event for D_1, D_1 belongs to a set of simultaneous events that is minimally sufficient for a later event D_2, … , and D_n belongs to a set of simultaneous events that is minimally sufficient for *E*. When we consider the line of events more closely, and consider more intermediate events between *C* and *E*, we might sometimes find that the apparent process was not genuine. That is, we might find intermediate times when the chain is broken. In such cases, we should conclude that *C* is not a cause of *E.* Conversely, if the chain is not broken when we consider more and more intermediate times, there is a genuine process connecting *C* to *E,* and we should conclude that *C* causes *E.* Call this account of causation *Elaborated minimal sufficiency.*

In *Presentation on pre-empted harm,* we find that the apparent process is not genuine. When we bring t_2 into the consideration, we find that the chain of relations of minimal sufficiency connecting Toni's readiness to point out the flaw to the presenter's becoming distressed is broken. As evidenced by the fact that Toni did not point out the flaw even though he was ready to do so, Toni's readiness to point out the flaw does not belong to a set of actually occurring events at t_1 that guaranteed the occurrence of this intermediate event at t_2.. Thus, *Elaborated minimal sufficiency* correctly entails that Toni did not cause the presenter to become distressed, and so we can use it together with NCA to show that Toni did not harm the presenter.

Still, *Elaborated minimal sufficiency* entails that Wlodek raising his index finger was a cause of the presenter's becoming distressed (via his pointing out the flaw), so together with NCA it correctly entails that Wlodek harmed the presenter.

Have we found an accurate account of harming? Unfortunately, we have not. Still, our exploration of NCA and minimal sufficiency has not been in vain. We can use *Elaborated minimal sufficiency* together with SCA to construct an accurate account of harming. For what remains of this paper, I will first go through a few reasons why NCA coupled with *Elaborated minimal sufficiency* fails as an account of harming, and then go on to suggest a more accurate account.

Why NCA Fails

NCA together with *Elaborated minimal sufficiency* fails as an account of harming for a number of reasons. First, it gives counterintuitive verdicts in so called switching cases, such as the following:

> *The assistant's choice:* The assistant of the editor of the *Journal of Over-Determination* has the task of notifying me that my paper has been rejected. He can do this in two ways: either he can send the message himself by email, or he can wait until the journal's application system sends me the notification automatically. Either way, I will get notified, whereby I will get sad. As it turns out, the assistant sends the notification himself by email.

Here, it seems that the assistant does not harm me by sending the message via email.[12] I would get the message anyway, and there was nothing he could do to avoid me getting sad. However, this is not what *Elaborated minimal sufficiency* coupled with NCA entails. Remember that we have to consider intermediate events between the point in time when the assistant sends the email and the point in time when I receive the message. At one such point in time, the email arrives at my computer. We then find that the assistant's sending the message was minimally sufficient for the message's arriving at my computer, and that the message's arriving at my computer was minimally sufficient for me reading it and getting sad. So *Elaborated minimal sufficiency* entails that the assistant caused me getting sad. There is a genuine process going from the assistant's sending the email and my getting sad. NCA, in turn, then counterintuitively entails that the assistant harmed me.

Second, NCA combined with *Elaborated minimal sufficiency* cannot explain our shifting intuitions in *Journal of Over-Determination.* Remember that it seems on the one hand that Björn did not harm me since, given that Dan recommended rejection, I would become sad and have few chances of getting tenure regardless of whether Björn recommended rejection. Similarly, given Björn's recommendation to reject, it seems that Dan's recommendation did not harm me. On the other hand, it seems that Björn's and Dan's recommendations harmed me since, if they had not recommended rejection, I would not have become sad and I would have better chances of getting tenure. However, NCA combined with *Elaborated minimal sufficiency* entails that Björn's recommendation did harm me. It belongs to a set of simultaneous events that was minimally sufficient for me getting sad and with few chances of getting tenure. And the chain of relations of minimal sufficiency remains when we consider more and more intermediate times. The same goes for Dan's recommendation. So, rather than explaining our shifting intuitions, the account of harm under consideration brutely entails that one of the intuitions is correct.

Third, as Michael Rabenberg (2014), Molly Gardner (2017) and others point out, NCA gives counterintuitive verdicts in some cases regardless of which account of causation we use. Consider for instance the following case:

[12] Switching cases are common in literature on causation and on moral responsibility. See e.g. Thomson (1976) and Paul and Hall (2013). Standardly, the agent in such cases is not considered to cause the outcome, to be blameworthy for the outcome or to have control over the outcome. In a similar vein, it seems to me that what the agent in does in switching cases does not stand in right relation to the person *S* suffering harm for it to be correct to say that what the agent does harms *S*. In switching cases, unlike in pre-emption cases, there is no relevant possibility that the harm does not occur.

> *Dan's phone call*: Björn is really sad because he and his best friend Dan are not neighbours anymore when Dan calls him on the phone just to say "hi". After the call, Björn is still sad, but much less so.

Here, it does not seem that Dan's call harmed Björn. On the contrary, Dan's call cheered him up. However, this is not what NCA entails. Assuming that Dan's call caused Björn to be less sad (which any plausible account of causation would imply), NCA entails that Dan's call harmed Björn. Being a bit sad is an intrinsically bad state (we might assume), so Dan's call caused Björn to be in an intrinsically bad state.

The first two objections indicate that *Elaborated minimal sufficiency* is an inadequate account of causation. The third objection indicates that NCA's insistence on taking intrinsically bad states into account is problematic. My suggestion is that we go back to considering comparative accounts of harming like SCA to avoid the last problem, and that we use *Elaborated minimal sufficiency* to correct for the problems SCA runs into.

Harming

SCA seems to give a necessary but not sufficient condition for harming. It gives intuitively correct verdicts in some cases, like *Journal of Over-Determination*. However, in other cases, like *Presentation on pre-empted harm*, there seems to be something missing. While SCA correctly entails that Wlodek harmed the presenter (in a way), it fails to pinpoint the reason why Toni did not.

Elaborated minimal sufficiency, in turn, seems to capture a necessary but not sufficient condition for causation. It gives the intuitively correct verdict that Wlodek but not Toni caused the presenter to become distressed in *Presentation on pre-empted harm*, but it fails to pinpoint a reason why the assistant's email was not a cause of my being sad in *The assistant's choice.* It also fails to explain why it seems that, given that Dan recommended to reject my paper, Björn's recommendation to reject my paper does not seem to be a cause of my getting and having few chances of getting tenure.

This suggests that we might combine the conditions given in SCA and *Elaborated minimal sufficiency* to get a more accurate account of harming with two necessary conditions, as follows:

> *Harming:* An event *C* harms a person *S* in H if and only if
> (a) There is a genuine process connecting *C* to *S'*s being worse off, and
> (b) *S'*s being better off would have been more secure in H had *C* not occurred.[13]

[13] This account of harming is inspired by Touborg's (2018) account of causation and is in many aspects similar to the accounts of teleological reasons and blameworthiness Touborg and I have developed elsewhere.

This account gives the right verdict in all the cases we have considered. To begin with, it can explain our shifting intuitions in *Journal of Over-Determination.* Consider first the larger possibility horizon H_J according to which there are four relevant possible worlds: Björn and Dan recommends rejection, Björn but not Dan recommends rejection, Dan but not Björn recommends rejection, and neither Björn nor Dan recommends rejection. Given this possibility horizon, *Harming* entails that both Björn and Dan individually harmed me (as opposed to, for instance, merely being *involved* in harming me). As we have already seen, (a) there is a genuine process connecting Björn's recommendation to my being sad and having few chances of getting tenure, and (b) my not being sad and having few chances of getting tenure would be more secure in H_J if Björn had not recommended rejection. A similar argument shows that Dan's recommendation to reject my paper harmed me, given H_J.

Further, given the smaller possibility horizon $H_{J\text{-small}}$, where we do not treat it as a relevant possibility that Dan had recommended otherwise, *Harming* instead entails that Björn's recommendation to reject my paper did not harm me. Even though there still is a genuine process connecting his recommendation to my being sad and having few chances of getting tenure, his recommendation does not make this outcome more secure. In fact, there is no relevant possibility in this possibility horizon that I would have been better off. That is, while condition (a) still is satisfied, condition (b) is not. A parallel argument shows that Dan's recommendation did not harm me, given an alternative $H_{J\text{-small}}$ that holds fixed Björn's recommendation to reject.

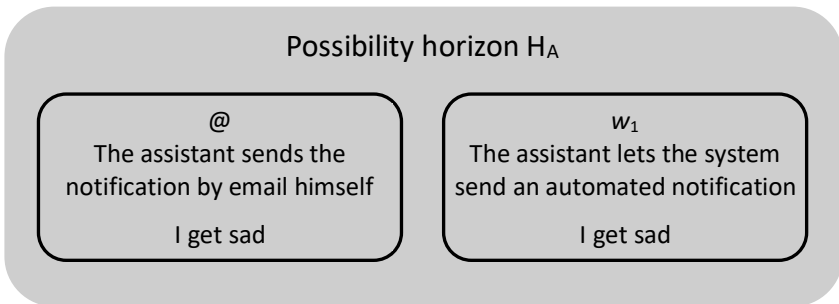
This way, *Harming* allows us to explain the shifting intuitions as stemming from different possibility horizons, or from different ways of understanding the situation. This raises the further question of which way to understand the situation that is the more accurate one. *Harming* does not help us out here. It does not say anything about which possibility horizon we should use. However, there are pragmatic reasons for thinking that the larger possibility horizon is the more accurate one. For one thing, it seems arbitrary to say that while there is a relevant possibility that Björn would have recommended otherwise, there is no relevant possibility that Dan would have done so (and vice versa). If we treat it as on open possibility that one of them could have recommended otherwise, there is a pressure to accept that both of them could have recommended otherwise.[14]

Harming also gives the right verdict that Wlodek but not Toni harmed the presenter (in a way) in *Presentation on pre-empted harm.* As we have seen, (b) both Wlodek raising his finger and Toni's readiness to point out the crucial flaw increased the security of the presenter's becoming distressed in H_P. This is why SCA wrongly indicates that Toni caused the presenter to become distressed. However, (a) there is a genuine process connecting Wlodek raising his finger (via his getting

[14] For more arguments why the larger possibility horizon typically is the more accurate one, see e.g. Gunnemyr (2021: 237-42).

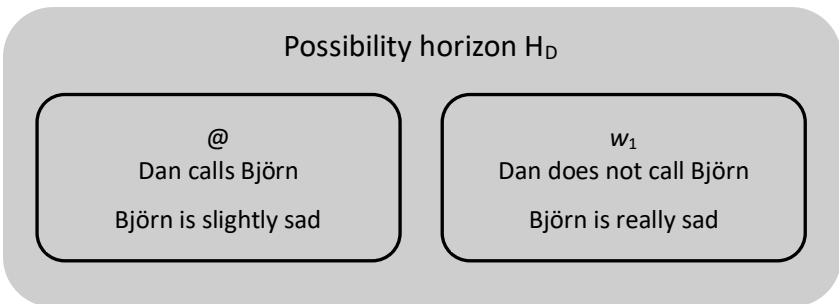
the word from the chair and pointing out the crucial flaw in the presenter's argument) to the presenter's becoming distressed, but no similar process connecting Toni's readiness to point out the flaw in the argument to this outcome. Therefore, *Harming* entails that Wlodek but not Toni harmed the presenter.

Next, *Harming* correctly entails that the assistant did not harm me by sending me an email himself rather than letting the system send me an automated notification in *The assistant's choice*. As explained earlier, (a) his sending the email is connected to my being sad via a genuine process. (This was why *Elaborated minimal sufficiency* gave the wrong verdict about the case.) Still, (b) his sending the notification by email did not increase the security of my sadness. To see this, we have to settle the relevant possibility horizon. In this case, there are two relevant possibilities, as indicated in the following possibility horizon:



There is no relevant possible world where I am better off; I get sad in every relevant possible world. Therefore, the security of the outcome that I get sad is just as secure in all relevant possible worlds. So, the assistant's sending the email does not make it less secure that I am better off, which means that condition (b) is not satisfied.

In addition, *Harming* yields the intuitively correct verdict that Dan did not harm Björn in *Dan's phone call.* While there is a genuine process connecting Dan's call to Björn's being (only) slightly sad, Björn's being better off had not been more secure if Dan had not called him. On the contrary, Björn would have been worse off had Dan not called him. To see this clearly, consider the relevant possibility horizon:



As you see, Björn is better off in @ than in w_1. Moreover, Björn's being better off would not have been more secure had Dan not called him. Indeed, had Dan not called him, Björn would have been worse off. He would have been really sad rather than just slightly sad.

Conclusions and Further Questions

To sum up, *Harming* gives intuitively correct verdicts about when an event harms someone in a wide range of cases. It gives the right verdict in overdetermination cases like *Journal of Over-Determination,* in pre-emption cases like *Presentation on pre-empted harm,* switching cases like *The assistant's choice,* and cases like *Dan's phone call* where someone's harm is relieved but not fully so. *Harming* can also explain the shifting intuitions we might have in cases like *Journal of Over-determination*: the different intuitions stem from different possibility horizons – or less formally, from different ways of understanding the situation at hand.

There is still work to do. As it stands, *Harming* will deliver counterintuitive verdicts in late pre-emption cases, in cases where it matters which possible world we take to be the relevant contrast, and in non-threshold cases (i.e. collective harm cases without a threshold). To get an idea of how *Harming* could be modified to cover such cases as well, see Gunnemyr (2021: chs 5, 6, 11 and 12) To work out the details will have to be work for another day.[15]

References

Bradley, Ben (2009) *Well-being and death*. Oxford: Clarendon Press.

Bradley, Ben (2012) "Doing away with harm". *Philosophy and Phenomenological Research, 85*(2): 390-412.

Feinberg, Joel (1984) *Harm to others*. New York: Oxford University Press.

Feit, Neil (2015) "Plural harm". *Philosophy and Phenomenological Research, 90*(2): 361-88.

Feit, Neil (2016) "Comparative harm, creation and death". *Utilitas, 28*(2): 136-63.

Gardner, Molly (2017) "On the strength of the reason against harming". *Journal of Moral Philosophy, 14*(1): 73-87.

Gunnemyr, Mattias (2021) *Reasons, blame, and collective harms* (PhD thesis). Lund University.

[15] I want to thank Andrés G. Garcia, Jens Johansson, and Jakob Werkmäster for constructive input on an earlier version of the paper.

Harman, Elizabeth (2009) "Harming as causing harm" in M. A. Roberts & D. T. Wasserman (Eds.) *Harming future persons* (137-54). Dordrecht: Springer Netherlands.

Jackson, Frank (1997) "Which effects" in J. Dancy (Ed.), *Reading Parfit* (42-53). Oxford: Blackwell.

Johansson, Jens (2023) "Petersson on plural harm" in A. G. Garcia, M. Gunnemyr, & J. Werkmäster (Eds.) *Value, morality & social reality: Essays dedicated to dan egonsson, björn petersson & toni rønnow-rasmussen.* Lund: Department of Philosophy.

Johansson, Jens, & Olle Risberg (2019) "The preemption problem". *Philosophical Studies, 176*(2): 351-65.

Klocksiem, Justin (2012) "A defense of the counterfactual comparative account of harm". *American Philosophical Quarterly, 49*(4): 285-300.

Lewis, David (1973) "Causation". *The Journal of Philosophy, 70*(17): 556-67.

Lewis, David (2004) "Causation as influence (extended)" in J. Collins, N. Hall, & L. A. Paul (Eds.) *Causation and counterfactuals* (75–106). Cambridge, Mass.: MIT Press.

Mackie, John L. (1974) *The cement of the universe: A study of causation*. Oxford: Clarendon.

Norcross, Alastair (2005) "Harming in context". *Philosophical Studies, 123*(1/2): 149-73.

Parfit, Derek (1984) *Reasons and persons*. Oxford: Clarendon Press.

Paul, L. A., & Edward J. Hall (2013) *Causation: A user's guide*. Oxford: Oxford University Press.

Petersson, Björn (2004) "The second mistake in moral mathematics is not about the worth of mere participation". *Utilitas, 16*(03): 288-315.

Petersson, Björn (2018) "Over-determined harms and harmless pluralities". *Ethical Theory and Moral Practice, 21*(4): 841-50.

Rabenberg, Michael (2014) "Harm". *Journal of Ethics and Social Philosophy, 8*(3): 1-32.

Shiffrin, Seana Valentine (1999) "Wrongful life, procreative responsibility, and the significance of harm". *Legal Theory, 5*(2): 117-48.

Thomson, Judith Jarvis (1976) "Killing, letting die, and the trolley problem". *The Monist, 59*(2): 204-17.

Touborg, Caroline Torpe (2018) *The dual nature of causation: Two necessary and jointly sufficient conditions* (Doctoral thesis). University of St Andrews.

Wright, Richard W. (1985) "Causation in tort law". *California Law Review, 73*(6): 1735-828.

Socratic Provocation in Art

Frits Gåvertsson

Abstract. In his 'Provocation in Philosophy and Art' Dan Egonsson argues that provocation is integral to Socrates' way of doing philosophy both when aiming for (the interlocutor's) personal moral development and as an instrument for societal change, and that provocation in art differs significantly from its Socratic counterpart. Morally dubious provocation in art can, however, Egonsson argues be justified on the grounds of its aesthetic qualities. In this response I discuss a number of aspects of Egonsson's insightful and thought-provoking treatment of the Socratic method and artistic provocation, and argue that Socratic provocation can have an important role to play in art that is structurally similar to its role in philosophy since provocative features of a work of art can be what grounds, or makes experientially available, the aesthetic qualities of the work.

Introduction

What follows is a commentary on Dan Egonsson's thought-provoking and perceptive 'Provocation in Philosophy and Art' (2015), where Egonsson argues that the role of provocation differs between philosophy and art due to the divergent *telē* of the disciplines—philosophy aims at discovering 'fundamental truths', whereas art, if it even has anything describable as a well-defined telos, aims to 'create aesthetic values' construed in terms of beauty and creativity (Egonsson 2015: 31)—and that morally dubious aesthetic provocation can nevertheless be justified on the grounds of the aesthetic qualities of the artistic provocation. In what follows, I argue that certain kinds of aesthetic provocation are such that they to a significant degree overlap with—or are perhaps identical to—Socratic provocation in philosophy.

The essay, meant to honour both Egonsson's scholarly efforts and remarkably encouraging Socratic teaching-style, proceeds as follows: After some stage-setting, I first give a brief characterisation of the concept of provocation before turning to its Socratic form. After that I consider comparable forms of Socratic provocation in the arts. The final section offers up some concluding remarks.

Setting the Stage

In what follows we shall not be mainly concerned with, but will touch upon, issues having to do with conceptual, axiological, or metaphysical connections between art and morality (on these issues see *e.g.,* Schellekens-Dammann 2020, 2008; Gaut 2007; Hämäläinen 2016; Kieran 2002; Stecker 2019, 2005; for an overview see Caroll 2000). Rather, at the forefront of the current text lies a structural issue, namely, 'Can artistic provocation, and provocation in the arts, be structurally similar (or perhaps identical) to its philosophical, *i.e.,* Socratic, counterpart?'.[1]

In saying that there is such structural similarity with regards to how provocation functions in the two domains, I do not wish to commit myself here to any further structural similarities. It is nevertheless true that what follows, in a seemingly paradoxical move, aligns both with John McDowell's (1983) manner of arguing for moral objectivity on aesthetic grounds and Philippa Foot's (1970) manner of arguing that morality, just as aesthetics, is subjective since both of these authors put structural issues (in terms of parallels between the moral and aesthetic domains) front and centre.[2] For similar structural reasons I do not wish to entangle myself in the fascinating issue of possible orderings (lexical or otherwise) of the two domains (on this see Egonsson 2015: 32-33; Wolf 2015).[3] Nor do I wish to take a definitive

[1] Consequently, nothing of what follows depends upon us accepting any form of, analytic connection between the aesthetic and the ethical, or some axiological 'interaction theory' (Schellekens-Dammann 2020) of the value of art according to which aesthetic and moral value interact in important ways, or metaphysical connection between the two domains. Sure enough, accepting some version of these views will probably make it easier to argue for structural similarities between artistic and philosophical provocation, and I tend to think that axiological interaction seems rather plausible, at least with regard to certain artworks and artforms, but nothing in what follows hinges on that being so. As was pointed out by an anonymous reviewer the notion of 'structural similarity' employed here—and the precise determination of what kind of things can stand in such similarity relations—is difficult to fully explicate but all that is needed here, I think, is the idea that the moral and aesthetic domains can be more or less similar (with me seeing more such similarity in regards to the role of provocation than does Egonsson).

[2] Foot would ultimately change her mind on this issue (*cf. e.g.,* Foot 2001; on this see Hacker-Wright 2013 and the contributions to Hacker-Wright 2018).

[3] Egonsson (2015:32) argues that the ordering between the two domains (and the issue of which type of reason is construed as overriding) in all probability comes down to a form of soft relativism (*i.e.* the different domains normally retain some sensitivity to one another) rendering the choice between the two an existential one. I have serious qualms regarding such existential choices, but will not pursue the matter here (although see *e.g.,* Murdoch 1956, 1970).

stand here on the issue of interaction between aesthetic and ethical value. Nothing in what follows prevents us from saying either that the final value of a work of art such as *e.g.,* Umberto Boccioni's 1913 futurist bronze sculpture *Unique Forms of Continuity in Space* [*Forme uniche della continuità nello spazio*] is more or less valuable because of the purportedly (proto-)fascist overtones of the work, or for that matter, that such moral values are neither here nor there as far as the value of the work as a work of art is concerned.[4]

Finally, another clarificatory concession must be made here. What follows is, arguably, dependent upon us granting that even if we agree with Egonsson (2015: 31) that art's functional *telos* is the production of aesthetic value (which Egonsson construes in terms of beauty and creativity (2015: 31)), it might still be the case that the grounding of such value, at least sometimes, in turn depends upon cognitive elements. Although this seemingly excludes formalist approaches (*i.e.* approaches, such as those of Bell (1913) and Fry (1920), that take aesthetic experiences, qualities, properties, and values to be formal in the sense of being accessible by direct sensation), it does not restrict the argument to what we—following Kant (1790: §16) but without feeling obliged to engage with the considerable interpretative difficulties associated with the notion—might call 'dependent' [*anhängend*] beauty, since we might still construe aesthetic experiences as free play unbounded by the conceptual and therefore direct, even if we allow for the possibility that such experiences are grounded in, or necessarily preceded by, cognition.[5]

The Concept of Provocation

I agree with Egonsson (2015: 29) that, for current purposes, we need not concern ourselves with a delineation of the concept of 'provocation' in terms of necessary and sufficient conditions since a rough description will do. Going along with Egonsson (2015: 29-30) we can say that a provocation is an intentional or unintentional incitement of an *active* moral negative response that can be ascribed to 'the *content* of a message (that is statement or work), the *manner* in which it is

[4] Egonsson (2015: 32) distances himself from 'ethicism' (Gaut 2007: 10), i.e., the thesis that an aesthetically relevant ethical flaw is also necessarily an aesthetic flaw (the beauty of the brief love affair depicted in Clint Eastwood's 1995 romantic drama film *The Bridges of Madison County* based on the Robert James Wallers' novel being Egonsson's rather clever counter-example (see also *e.g.,* Stecker 2005; Kieran 2002; Schellekens-Dammann 2020)). I agree that strong ethicism of this kind seems rather implausible and think that the first option listed above renders the correct verdict in the Boccioni case, but will not pursue the matter further here.

[5] For influential criticisms of formalist approaches see Walton 1970; Danto 1981: 94-95; for an overview see Dowling 2022). For influential construals of the Kantian distinction between 'free' [*freie*] and 'dependent' beauty see Allison 2001; Crawford 1974; Guyer 1997. We will get back to the possibility of accommodating Socratic provocation in a formalist framework towards the end of the present text.

delivered and the way in which it is *received*' (2015: 29, emphasis in original).[6] Furthermore we can say that instances of provocation need not be intentional (in any sense over and above what is required for intentional action), but that paradigmatic instances of provocation both in philosophy and art are.

In order to move on from simply trying to understand the notion of provocation at work here, and without committing ourselves to any kind of overtly Wittgensteinian methodology (although see *e.g., PI* §§ 19, 23, 241), we might perhaps say that provocation only makes sense against a backdrop of human life as a whole and that perhaps provocation in philosophy, as well as in the arts, can only perform its most interesting function as part of an investigation into 'how we ought to live' (Pl. *Rep.* 352^{d}5-6; *cf.* Pl. *Grg.* 487^{e}7-488^{a}2) in quite general terms.[7] Most importantly, this approach, or other relevantly similar approaches, would suggest that the value, or function (see Egonsson 2015: 31ff.), of provocation in art as well as in philosophy that we ought to look for here—*i.e.,* instances where such value, or such functional characteristics, converge across the two domains—stands to be found in a broad engagement with human life and meaning-making.

If I am right in thinking that there are, or at least can be, instances of provocation in art as well as in philosophy that come together in this intersection between art, philosophy, and life in general then it would seem that while Egonsson is surely right in claiming that in many cases a morally dubious aesthetic provocation can (only) be justified on the grounds of the aesthetic qualities of the work (2015: 31-33) and that aesthetic provocation often is quite different from its Socratic-philosophical counterpart, we should also be open to the possibility that the two can come together in such a way that at least some aesthetic or artistic provocations function in a way that is structurally, or functionally, similar to Socratic provocation in philosophy. In order to see how, we need to say something more about the nature and function of Socratic provocation.

---

[6] Egonsson (2015: 30) insightfully compares the structure of provocation, thus understood, to blasphemy. The etymology of blasphemy (gr. *blasphēmía*; from *blaptō* ('to hurt') and *phēmē* ('speech, talk, utterance' but also 'fame' and 'reputation')), as Yvonne Sherwood (2021: 2; *cf.* Burnes Coleman 2011) notes, brings 'blasphemy' rather close not only to 'provocation', but also to our notion of 'hate speech', both in terms of being offensive and having a social dimension. (See also *e.g.,* Perret 1987; Fisher & Ramsay 2000). As was pointed out by an anonymous reviewer, Egonson's construal of 'provocation' casts the net rather widely and while we might quarrel with this it is important to keep in mind that such a wide construal also has its benefits, chief among them in the present context being that it allows for a work of art to be provocative even in the absence of any corresponding intention on the part of the artist (even if it might well be the case that typical instances of provocation are such that someone intentionally seeks to elicit a negative response to oneself or one's actions).

[7] An alternative way of putting forth the same point would be to argue, with *e.g.,* Michael Thompson (2008: 25-82) and Philippa Foot (2001: 25-37), that 'life' is a logical concept (see also Midgley 1973, 1979; on this see Lipscomb 2016). Another—I think fruitful—alternative, suggested to me by an anonymous reviewer is to utilize an Austinian (1962) 'speech-act' analysis of 'provocation'. It seems to me that such an Austinian approach can be fruitfully combined with the aforementioned Wittgensteinian approach. For readings of Wittgenstein and Austin along such combinatory lines see *e.g.,* Cavell (1979); Forsberg (2022).

Socratic Provocation

Even if Plato's Socrates is marked by paradox (Vlastos 1971) it is safe to say that he clearly envisages provocation as central to not only his dialectics but to his societal rôle even in the early, or 'Socratic' dialogues[8]:

> For if you put me to death, you will not easily find another, who, to use a rather absurd figure, attaches himself to the city as a gadfly to a horse, which, though large and well bred, is sluggish on account of his size and needs to be aroused by stinging. I think the god fastened me upon the city in some such capacity, and I go about arousing and urging and reproaching each one of you, constantly alighting upon you everywhere the whole day long. Such another is not likely to come to you, gentlemen; but if you take my advice, you will spare me. But you, perhaps, might be angry, like people awakened from a nap, and might slap me, as Anytus advises, and easily kill me; then you would pass the rest of your lives in slumber, unless God, in his care for you, should send someone else to sting you. And that I am, as I say, a kind of gift from the god (Pl. *Ap.* 30^{e}-31^{a}, trans. Harold North Fowler).

As Egonsson (2015: 27-29) shows, this self-image persists through the occasionally overbearingly preachy 'middle' dialogues (*cf. e.g.,* Pl. *Rep.* 331^{d} ff., 357^{a} ff., 449^{a} ff., 471^{c} ff.; on this see Miller 1985)—where the very idea of philosophical provocation is at one point criticised, albeit briefly (Pl. *Gorg.* 482^{c} ff.)—and is still very much present in what is probably later works such as the *Symposium*.[9] The

[8] On the ordering of the dialogues into early (or 'Socratic' *cf. e.g.,* Arist. *Metaphysics* 987^{b}1, *Sophistical Refutations* 183^{b}7), middle, and late see Gregory Vlastos (1991: ch. 2-3; see also Vlastos 1994: 135). This tripartite division is obviously not without its problems (*Timaios, Theaitetos* and the first book of the *Republic* are difficult to place, for instance), but it gives us a handy way of organising the *corpus platonicum* that gives us a sense of development over time.

[9] The *Symposium* itself could arguably be seen as an example of the kind of Socratically provocative art that we are here concerned with since it makes those who engage with it contemplate the passion of personal love (in the sense of both *erôs* and *philia*), a central aspect of human life, in new ways and challenges social conventions. Like in the other dialogues that relate to these issues (the *Lysis* and the *Phaedrus*) the central character of Socrates plays the part of the quintessential philosopher (*i.e.,* lover of wisdom and (elenchtic) discussion) and subverter of erotic norms since both of these aspects bring Socrates into conflict with the *paiderastia*—a set of social norms that regulate the intercourse between an older male and a teenage boy where the latter is supposed to learn virtue from the former. There are, naturally, a range of problematic aspects to this social practice that should be evident to the modern-day reader. But those issues, having to do with, among other things, the power dynamics involved are not Socrates' (or Plato's) immediate concern in the *Symposium* (even though the discussion at Pl. *Symp.* 204^{d}-209^{e} seems to address and criticise the sexism and exclusively male perspective of the previous speakers and some issues, such as *e.g.,* the classist practises involved in determining the 'guest list' for a symposium of this sort are addressed in other dialogues). What Plato is arguing here, I believe, is that the ideology of the *paiderasteia*—that love is a combination of a love of the soul (virtue) and love of the body (sexual gratification; cf. Pl. *Symp.* 180^{c-d}) where the latter must masquerade as the former—involves an inherent risk of us succumbing to fantasy and illusion. We might be led to believe that the fruits of love of the soul (real wisdom and virtue) can be gained as easily and quickly as sexual gratification. In other words, Plato is warning us that focusing too much

provocative element is thus a constant through the changes in method that seemingly transpires in the dialogues. Socrates' provocations are there to jump-start the *standard elenchus*[10] of the early dialogues and they serve to retain the *pro forma* interlocutor's interest through the maieutic middle period[11], and Socrates' provocativeness is equally as prominent in the 'erotic' dialogues[12] of the middle period. The drunk Alcibiades's likening of Socrates' arguments (*logoi*) to the songs of the satyr Marsyas gives us insight into the function of their provocative elements:

> Whether they are played by the greatest flautist or the meanest flute-girl, his melodies have in themselves the power to possess and so reveal those people who are ready for the god and his mysteries (Pl. *Symp.* 215^{c}, trans. A. Nehamas and P. Woodruff; *cf.* Miller 1985: 163).

Marsyas is a notoriously double-edged figure in the mythic tradition. On the one hand we have common stories of the hubristic satyr that rebels against the gods by picking up Athena's discarded aulos and challenging Apollo to a music contest, but on the other hand we are also on occasion (*e.g.,* Diodorus Siculus *Library of History* III.59) met with the wise Marsyas marked by intelligence (*sunesis*) and self-control (*sophrosune*), which is the side presumably alluded to by Alcibiades. I believe that Plato, by having Alcibiades invoke the satyr's likeness to Socrates in the *Symposium*, is drawing on both sides of the character, thus inviting us as readers to

---

on the pleasures of sexual intercourse and thus not committing fully to the process of self-transformation might make us forget that self-knowledge is hard to gain.

[10] *I.e.* a process where one of Socrates' interlocutors presents a thesis, p, (usually a suggested definition of an ethical concept) and Socrates goes on to show how its negation ($\neg p$) follows from some other propositions (q, r) which the interlocutor (and usually Socrates) take to be true, thus showing that the conjunction of the original thesis and these $\{p \,\&\, q \,\&\, r\}$ is false due to inconsistency (on this see Vlastos 1994: 1-28). Ex. *Charmides* 160^{e}2-161^{b}4: (p): temperence (*sōphrosunē*) is a sense of shame (*aidōs*)(160^{e}2-5); (q) temperence is fine (*kalon*) and good (160^{e}13); (r) Homer was right to say that a sense of shame is not always a good thing (161^{a}2-4).

[11] In these dialogues the theses under investigation are introduced, argued for, examined, and amended by Socrates himself in a didactic style, with the interlocutor reduced to a yes-man that might occasionally raise objections but never puts up sustained resistance. On this see (Vlastos 1994: 29-37). The underlying epistemic assumption here, it is commonly assumed, is that what appears to be new knowledge is really reminded or recollected. A, to my mind, more promising suggestion that I think is in line with Egonsson's general argument is to treat Plato's Theory of Recollection as touching upon a similar problem as that which concerns Wittgenstein in the sections (§§ 201 ff.) on rule-following in the *Philosophical Investigations—i.e.,* how can someone who is being shown part of a pattern know how to go on? On the standard picture Plato's answer is that we are reminded of innate knowledge, but what if we instead take the Wittgensteinian route (which seems compatible with a Platonic concern for our human nature) and argue that the pattern is part of human life? On this see (Anscombe 1993; see also McDowell 1984; Mac Cumhaill and Wiseman 2022: 254-256)

[12] In these dialogues the idea seems to be that through love of what is beautiful (*cf.* Pl. *Symp.*) and Good (*cf.* Pl. *Rep.*) the soul can come to gaze on the Forms. It is telling that it is in these dialogues that Plato's use of myth and metaphor is at its most prominent.

evaluate the different sides to the paradoxical Socrates of the dialogue. One of the greatest virtues of Egonsson's 'Provocation in Philosophy and Art' is the way that this tension is brought out: just as there is a thin line between righteousness and self-righteousness, there is a thin line between engagingly productive provocation and provocation that results in utter breakdown of the discussion. Socratic provocation, then, is a means to "provoke and irritate in order to awake" (Egonsson 2015: 28) those susceptible to go deeper, at the expense of what is familiar, and so structures the interlocutors' own self-initiation into the mystery that is philosophy. Naturally, there is an extra-textual level here as well: through Socrates' exchange with his interlocutors Plato challenges *us*, as readers, as we are invited to show, or at least ponder, our own aptitude (Miller 1985: 165-166). Or, as Iris Murdoch puts it:

> Plato pictures human life as a pilgrimage from appearance to reality. The intelligence, seeking satisfaction, moves from uncritical acceptance of sense experience and of conduct, to a more sophisticated and morally enlightened understanding (Murdoch 1977: 2).

In order for this pilgrimage to be set in motion the reader must be provoked into taking the first stumbling steps. The deepest function of Socratic provocation, then, is to provoke us, as readers, into philosophical reflection that goes beyond the (explicit) content of the dialogue and manifests as our own philosophical insight.

Egonsson's discussion of these matters makes it abundantly clear not just how painful this process can be—since it requires us to cast aside not only convention but also our neurotic self-obsession in order to see reality as it really is (*cf.* Pl. *Rep.* 515 ff.; Murdoch 1958b: 268; Holland 2012)—but also how *fragile* Socratic provocation is. Socrates's provocations in the *Symposium* aim both at questioning societal convention (*e.g.,* the exclusion of women from symposia and the practise of *paiderastia*) and neurotic self-obsession (the main criticism of the other symposiasts' speeches is that they are both too particular, *i.e.* are concerned with particular loves rather than love itself, and all too tied up with the speech-giver's own field of expertise and perspective (*cf.* Agathon's remarks at Pl. *Symp.* 195^{a} and Socrates' rebuttal in the form of an *elenchus* at Pl. *Symp.* 198^{a}-201^{c})). Plato often has the discussion come dangerously close to deteriorating, thus illustrating the fragility of provocation as a philosophical strategy.[13] The reason for the fragility of provocation mirrors, and is in many cases parasitic upon, the painfulness of the process of philosophical realization since it is the demand to cast aside convention

[13] Interlocutors such as Protagoras (Pl. *Prot.* 333^{e}, 360^{d}), Callicles (Pl. *Gorg.* 489^{b-c}), and Anytus (Pl. *Men.* 94^{e}) all at various places voice their irritation with Socrates' provocative conversation style. In addition, the *Euthydemus* would not work as a special case in which Socrates' ironic modesty is played for comic effect (thus constituting a self-reflexive critique of ironic technique itself), which in turn is transformed into a subtle mockery of rhetoric, were it not put forth in contrast to Socrates' usual provocative style (on this see Michelini 2000).

and neurosis that most reliably irritate the interlocutors.[14] In order for Socratic provocation to fulfil its function of combating both the interlocutors' neurotic self-obsession and societal convention, as Egonsson (2015: 29) observes, '[t]he spectators are to be amused whereas the interlocutor is to be kept in good humour'.[15] With this understanding of Socratic provocation as a painful and fragile incitement to move beyond both our own neurotic tendencies and societal convention in hand, let us move on to investigate whether structurally or functionally similar forms of provocation can be conceived of as central to (at least some instances, genres, and types of) Art.[16]

Provocation in Art

Provocative art is arguably as old as art itself, and yet far from all such provocative art can be said to be reliant on 'Socratic provocation' in the sense sketched above. As Egonsson (2015: 30) points out, provocative aspects of art were increasingly emphasised starting with romanticism and continuing into the historical avant-garde and beyond. Much of this historical development is regrettably tangled up in values, concerns, and ideals that are less than flattering. It is indeed inviting to view the resulting artistic provocation as simply either (a) provocation for provocation's sake (*e.g.,* Hugo Ball's 1916 'sound-poem' *Karawane*), (b) as primarily directed towards

[14] Perhaps this is at its most obvious in Socrates' exchange with the headstrongly naïve Polus in the *Meno* (Pl. *Men*. 461^{b}-481^{b}), even though the matter is further complicated by the fact that the point of that particular exchange apart from showing off Socrates' dialectical skills, as far as I understand it, is also to illustrate the difficulty of discerning a genuine philosophical victory (fairly won and free of fallacious reasoning) from mere rhetorical subjugation (on this see *e.g.,* Vlastos 1967; Johnson 1989).

[15] Naturally, as Ann Michelini observes, irony and humility can sometimes be self-defeating: "In *Euthydemus*, both the narrator, Socrates, and his secondary audience, Crito, suggest that the primary audience in the Lyceum may have seen Socrates' fake adoration of the brothers at face value, as demeaning him and elevating them" (2000: 516).

[16] It might be thought that, as one anonymous reviewer objected, that the dialogical character of Socratic provocation—where Socrates' provocations are part of a process where Socrates follows up on the provocation and guides the interlocutor to a better understanding of the topic at hand—differentiates it from, at least paradigmatic, provocation in art as we are, at least typically, left to our own devices in sorting out our responses to art (or are at least not given the opportunity of continued dialogue with the artist). While there is something to this line of criticism, I would like to point to two things that I think lessen its appeal. Firstly, the Platonic dialogues themselves are literary artworks that we, as readers, are meant to wrestle with in solitude (that is, the dialectics are not as open-ended as they might at first glance appear and there was never meant to be anything like a continuous conversation with the author). Secondly, it doesn't seem to me that we are as solitary in coming to grips with our responses to art as the objection assumes. Rather, even if we do not necessarily subscribe to anything as strong as a Gadamerian (*e.g.,* Gadamer 1986) understanding of art as interlocutor, art is still plausibly something that we engage with as a community, otherwise, what would be the point of evaluative criticism (on this see Gilmore 2013).

art itself and its various media ((see Greenberg 1961; Brüger 1984) *e.g.,* Virginia Woolf's stream-of-consciousness novels, Cézanne's landscape paintings[17], Mallarmé's poems), (c) the institution of the artworld and its critics (*e.g.,* the *Salon des Refusés*[18], Duchamp's 'ready-mades', Maurizio Cattelan's 2019 *Comedian*), or even (d) the public at large (*e.g.,* Andres Serrano's 1987 *Piss Christ*, Tracy Emin's 1998 sculpture *My Bed*, Cattelan's *The Ninth Hour* (on the latter see Schellekens 2007: 82)).

While Egonsson (2015: 30) might be right that much modern art is "anti-philistine but pro-critic", still, as should be evidenced by the list of works just given, elements of this modernist avant-gardist drive towards provocation (seen by Lyotard (1984) as driven by a search for the sublime) touch upon philosophical concerns that arguably render its provocativeness Socratic in the sense we are here interested in. While much of the impetus behind the Decadent poets' rallying cry of 'Épater la bourgeoisie' was undoubtedly pure-shock value it does still imply defying convention in a sense that corresponds to the societal aspect of Socrates' provocative practices. There are a number of candidate artworks and art forms—from the Dada and Fluxus movements, via Allan Kaprow's 'Happenings', to the *Esthétique Rélationnel* of the 1990's (see Bourriaud 1997 [2002]; on this see Bishop 2004)—that would seem to rely in some way or other upon something close to Socratic provocation since these artforms and artworks in various ways seek to interactively engage with its audience, and such engagement may well occasionally require a provocative impetus to kick-start said engagement. Granted, proto-philosophical concerns of this kind may not be enough for us to consider such artistic provocations Socratic in the full sense, but it does at least open up for the possibility of genuine Socratic provocation in the arts as a possible grounding of aesthetic qualities and values.

If we instead take a birds-eye view of this development and construe the 'ancient quarrel' (Pl. *Rep.* 607$^{b\text{-}c}$) between 'poetry' and 'philosophy' in Western culture as a series of responses to Plato's challenge directed at mimetic art—*i.e.,* art hinders philosophical awakening by directing our attention (on this see Murdoch 1959, 1970, 1977; on this see Wolf 2015: 163-180; Forsberg 2013: 138-150; Bolton 2019; Gåvertsson 2018: 61-85) away from what is real and towards a simulacrum, thus

[17] At least, I gather, this is how Maurice Merleau-Ponty (1964a, 1964b) understands Cézanne's post-impressionist paintings. Given such an understanding the artist's work involves significant philosophical elements pertaining primarily to phenomenology and the philosophy of perception.

[18] Arguably, many works associated with the Salon des Refusés, such as *e.g*., Édouard Manet's *Le Dejeuner sur l'Herbe* (1863), scandalised the general public as much as the artworld and can be seen as aiming at defying convention in a manner that would render their provocativeness at least proto-Socratic. Even the provocative use of light and darkness in *Le Déjeuner sur l'Herbe* can, I think, be seen as a not-so-subtle form of social commentary and the hard to pin down gaze of the central figure seems like a simultaneous invitation to see the Other and a commentary on the moral metaphor of 'vision', but, then again, the difficulty in fixating the meaning of the work remains part of its provocative appeal (on this see *e.g.,* Læssøe 2005).

fostering our own self-protective illusions (and so serving to fuel our neurotic tendency of turning towards the ego or resort to conventional thinking) rather than directing our gaze towards a transcendent reality—we can, following Beardsley (1958: 558-571) and Gaut (2007: 3-5), talk in terms of three general types of response. Two of these, it seems, directly depend upon something like Socratic provocation as an essential part of their response to Plato's challenge.

Firstly, there are those—often called 'humanists', including figures such as *e.g.,* Ruskin, Henry James, Tolstoy, and others—that seek to go against Plato's challenge head on by trying to emphasise the moral value of art. Any such strategy seemingly draws on the first aspect of Socratic provocative potential in art, to put it in Platonic terms, to combat the interlocutor-spectator's neurotic self-obsession and in doing so facilitate their pilgrimage from appearance to reality.

Secondly, another type of response—often called 'transgressionalist' and associated with figures such as Manet and Stravinsky, but also, perhaps more prominently with de Sade, Maplethorpe, and Easton Ellis—argues against Plato's challenge that art can be good because it transgresses, and so invites us to challenge, conventional morals, assumptions, and attitudes. In a manner similar to the humanist, this response draws on the second aspect of Socratic provocative potential in art by challenging societal convention.

A third type of response—often called 'aestheticism' and primarily associated with prominent figures during the latter half of the nineteenth century and onwards such as Whistler, Bell, Fry and Beardsley—argue that Plato's challenge is misdirected since moral evaluation of art rests on a category mistake, or that subjecting art to moral or cognitive evaluation somehow abases it.

It should be plain to see that it is accommodating Socratic provocation in art given an aestheticist outlook that is the most pressing challenge here since such an approach does not seem to draw on the cognitive or moral potential in Socratic provocation in the same direct manner as the other two. Nevertheless, it seems possible, even given an aestheticist stance to admit that Socratic provocative elements in a work of art can be either a (perhaps necessary) causal or metaphysical ground for the aesthetic interest of the work (which could be cashed out in terms of value, qualities, properties, judgments, attitudes, and so on). That is, provocation in art, and aesthetic provocation, can be a prerequisite, or ground, for a work of art succeeding as a work of art, and whatever Socratic provocation adds to such a ground need not be considered what Beardsley (1958: 558) terms "side effects" but may well be a prerequisite for, or constituent part of, the ground of that produces the inherent immediate effect. If this is an open possibility, then Egonsson's (tentative) assertion that philosophy and art have divergent goals does not preclude the possibility of Socratic provocation having a structurally similar function in art.

It might turn out that much modern art in general, and perhaps conceptual art in particular since much of it is put forth in a provocative spirit (see Young 2001; on this see Schellekens [forthcoming]), even if it seeks to provoke in fact is of little interest since much of what it says and is turns out rather morally trivial, cognitively

commonplace, and aesthetically unchallenging or uninteresting. Still, it seems to me that most provocative works that retain their aesthetic interest over time do so in virtue of their provocative elements being Socratic. The reason these works merit continual discussion, and why such artworks play a major part in our lives and our understanding of ourselves and our relation to the world is precisely because they challenge us to engage with *e.g.,* our own neurotic tendencies and social conventions in order to truly see the world and the Other. Thus, it might be that much art might fail to attain an interesting degree of Socratic provocation, but that '[g]ood art, on the other hand, provides work for the spirit' (Murdoch 1977: 77). It seems plausible to me that proper engagement with, say, Cormac McCarthy's Western novel *Blood Meridian* (1985) and the aesthetic achievement of its prose requires attending to and being provoked by the abhorrent violence depicted. That is, it seems to me that our attention wouldn't be directed at the rhythm, sound, and cadence of the words were it not for the terror it invokes. This way of looking at Socratic provocation in art also provides us with additional ways to analyse less successful attempts at artistic provocation since we have opened up for failure not only, as Egonsson (2015: 34) suggests through a lack of artistry and originality, but also through a failure to properly see and attend to the Other and to reality, or failing to ground, or make experientially available, the aesthetic qualities of the work.

Nothing in what was said above necessarily implies anti-formalism. To see why, let us return to Boccioni's *Unique Forms of Continuity in Space*. Nothing prevents us from saying that the fluidly forceful dynamic qualities of the work are accessible by direct sensation whilst still saying that these qualities are grounded in, or made experientially available by, the Socratically provocative nature of the work. This also allows us to say that the provocative (proto-)fascist undertones of the work are aesthetically important not because they somehow deepen our moral perspective or transgress societal convention, but because they ground, or make experientially available, the dynamic qualities of the work.

If I am right in what has been said here it might be that Socratic provocation has a role to play in art that is structurally similar to its function in philosophy, even if the cognitive, the moral, or what have you, should be properly and steadfastly differentiated from the aesthetic since the relevant aesthetic qualities might be grounded in the provocative elements of the work. If this is right, then Egonsson's assertion that artistic provocation, in order to be aesthetically successful, must be intrinsically aesthetically fascinating, must be modified somewhat since it might be the case that Socratic provocation in art can also function as grounds for, or means for making accessible, the aesthetic qualities of the work. In these latter cases it would appear that Socratic provocation in art and philosophy function in a structurally similar manner.

Conclusion

What has been said here about how I believe that Socratic provocation can have an important, albeit perhaps rather limited, role to play in art constitute but some initial remarks on provocation and its place in art and morality. In short, I have argued that Socratic provocation—understood as a painful and often fragile incitement to move beyond both our own neurotic tendencies and societal convention—can have an important role to play in art that is structurally similar to its role in philosophy. These remarks should not be taken as indicating my endorsement of a thoroughly cognitivistic or moralistic understanding of art, artistic or aesthetic value, and the like. Quite to the contrary, perhaps cognitive values are being emphasised a tad too much in the cultural and intellectual climate at the moment (at the expense of investigations focusing on more purely experiential matters); we shouldn't forget the purer enjoyments of aesthetic pleasure, even if it sometimes takes something provocative to ground or evoke such unusually delightfully pleasant experiences, and this appears to me a conclusion befitting the Dan Egonsson I know.[19]

References

Allison, Henry E. (2001) *Kant's Theory of Taste* Cambridge: Cambridge University Press.

Anscombe, G. E. M. (2011 [1993]) "The Origin of Plato's Theory of Forms" 1-9 in Mary Geach and Luke Gormally *From Plato to Wittgenstein: Essays by G. E. M. Anscombe* St. Andrews Studies in Philosophy and Public Affairs, Exeter: Imprint Academic.

Austin, John Lansgshaw (1962) *How To Do Things With Words* Oxford: Clarendon Press.

Bell, Clive (1913) *Art* New York: Capricorn Books.

Bolton, Lucy (2019) *Contemporary Cinema and the Philosophy of Iris Murdoch* Edinburgh: Edinburgh University Press.

Bourriaud, Nicholas (2002 [1997]) *Relational Aesthetics* Dijon: Les Presses du Réel.

Brüger, Peter (1984) *Theory of the Avant-Garde* Minneapolis: University of Minnesota Press.

Burnes Coleman, Elizabeth (2011) "The Offences of Blasphemy: Messages in and through Art" *Journal of Value Enquiry* 45: 67-84.

Caroll, Nöel (2000) "Art and Ethical Criticism: An Overview of Recent Research" *Ethics* 110(2): 350-387.

[19] This work was supported by a grant [P19-0937:1] from *The Bank of Sweden Tercentenary Foundation*. I want to thank Kristiina Savin, Anna Karlsson, Jonas Hansson, Disa Runeby, Max Minden Ribeiro, Martin Sjöberg, Niklas Dahl, Alexander Velichkov, and two anonymous reviewers for their valuable comments on earlier drafts of this paper.

Cavell, Stanley. (1979) *The Claim of Reason: Wittgenstein, Skepticism, Morality, and Tragedy* New York: Oxford University Press.

Crawford, Donald W. (1974) *Kant's Aesthetic Theory* Madison: University of Wisconsin Press.

Danto Arthur (1981) *The Transfiguration of the Commonplace* Cambridge, Ma.: Harvard University Press.

Dowling, Christopher (2022) "Aesthetic Formalism" *The Internet Encyclopedia of Philosophy*, ISSN 2161-0002, https://iep.utm.edu/aesthetic-formalism/, 2022-04-26.

Egonsson, Dan (2015) "Provocation in Philosophy and Art". *The International Journal of Social, Political, and Community Agendas in the Arts*, 10(3): 27-35.

Fisher, Anthony & Ramsay, Harden (2000) "Of Art and Blasphemy" *Ethical Theory and Moral Practice* 3: 137-167.

Foot, Philippa (1970) "Morality and Art" *Proceedings of the British Academy* 56: 131-144.

Foot, Philippa (2001) *Natural Goodness* Oxford: Oxford University Press.

Forsberg, Niklas (2013) *Language Lost and Found: On Iris Murdoch and the Limits of Philosophical Discourse* London: Bloomsbury.

Forsberg, Niklas (2022) *Lectures on a Philosophy Less Ordinary: Language and Morality in J. L. Austin's Philosophy* New York: Routledge.

Fry, Roger (1920) *Vision and Design* London: Chatto & Windus.

Gadamer, Hans-Georg (1986) *The Relevance of the Beautiful* Cambridge: Cambridge University Press.

Gaut, Berys (2007) *Art, Emotion and Ethics* Oxford: Oxford University Press.

Gåvertsson, Frits (2018) *Perfection and Fiction: A Study in Iris Murdoch's Moral Philosophy* Lund: Media Tryck.

Gilmore, Jonathan (2013) 'Criticism' 375-383 in Gaut, Berys & McIver Lopes, Dominic (eds.) *The Routledge Companion to Aesthetics* London: Routledge.

Greenberg, Clement (1961 [1936]) "Avant-garde and Kitsch" in *Art and Culture: Critical Essays*: 3-21. Boston, Mass.: Beacon Press.

Guyer, Paul (1997) *Kant and the Claims of Taste* 2nd edition Cambridge: Cambridge University Press.

Hacker-Wright, John (2013) *Philippa Foot's Moral Thought* London: Bloomsbury.

Hacker-Wright, John (ed.) (2018) *Philippa Foot on Goodness and Virtue* Palgrave Macmillan.

Hämäläinen, Nora (2016) *Literature and Moral Theory* London: Bloomsbury.

Holland, Margaret (2012) Social Convention and Neurosis as Obstacles to Moral Freedom' 255-273 in Broackes, Justin (ed.) *Iris Murdoch, Philosopher* Oxford: Oxford University Press.

Johnson, Curtis N. (1989) "Socrates' Encounter with Polus in Plato's *Gorgias*" *Phoenix* 43(3): 196–216.

Kieran, Matthew (2002) "Forbidden Knowledge: The Challenge of Cognitive Immoralism" in Gardener S. and Bermudez, J. (eds), *Art and Morality*: 56–73. London: Routledge.

Læssøe, Rolf (2005) "Édouard Manet's 'Le Déjeuner sur l'herbe' as a Veiled Allegory of Painting" *Artibus et Historiae* 26(51): 195–220.

Lipscomb, Benjamin (2016) "'Slipping Out Over the Wall': Midgley, Anscombe, Foot and Murdoch" in Kidd J. & McKinnell L. *Science and the Self: Animals, Evolution, and Ethics: Essays in Honour of Mary Midgley*: 207-223. London: Routledge.

Lyotard, Jean-François (1984) "The Sublime and the Avant-garde" *Art Forum* 22(8): 36-43.

Mac Cumhaill, Clare and Wiseman, Rachel (2022) *Metaphysical Animals: How Four Women Brought Philosophy Back to Life* London: Chatto & Windus.

McDowell, John (1983) "Aesthetic Value, Objectivity, and the Fabric of the World" in Schaper, Eva (ed.) *Pleasure, Preference, and Value* Cambridge: Cambridge University Press.

McDowell, John (1984) "Wittgenstein on Following a Rule" *Synthese*, 58(3), 325–363.

Merleau-Ponty, Maurice (1964a) "Cézanne's Doubt" in *Sense and Non-Sense*: 1-25 Northwestern University Press.

Merleau-Ponty, Maurice (1964b) "Eye and Mind" in James Edie (ed.) *The Primacy of Perception*: 159–190. Evanston: Northwestern University Press.

Michelini, Ann N. (2000) "Socrates Plays the Buffoon: Cautionary Protreptic in *Euthydemus*" *The American Journal of Philology* 121(4): 509–535.

Midgley, Mary (1973) "The Concept of Beastliness" *Philosophy* 48: 111-135.

Midgley, Mary (1979) *Beast and Man: The Roots of Human Nature* London: Methuen.

Miller, Mitchell (1985) "Platonic Provocations: Reflections on the Soul and the Good in the Republic" in Dominic O'Meara (ed.), *Platonic Investigations*. Catholic University of America Press, 163-193.

Murdoch, Iris (1956) "Vision and Choice in Morality" *Proceedings of the Aristotelian Society, Supplementary Volumes* 30: 14-58.

Murdoch, Iris (1959) "The Sublime and the Beautiful Revisited" *Yale Review* December: 247-271 reprinted in Conradi, Peter & Steiner, George (1999) *Existentialists and Mystics: Writings on Philosophy and Literature*: 261-286 London: Penguin Books.

Murdoch, Iris (1970) *The Sovereignty of Good* London: Routledge & Kegan Paul.

Murdoch, Iris (1977) *The Fire and the Sun: Why Plato Banished the Artists* Oxford: Oxford University press.

Schellekens-Dammann, Elisabeth (2008) *Aesthetics and Morality* London: Bloomsbury Aesthetics.

Schellekens-Dammann, Elisabeth (2020) "Evaluating Art Morally" *Theoria* 86(6): 843-858.

Schellekens-Dammann, Elisabeth (forthcoming) "Conceptual Art", *The Stanford Encyclopedia of Philosophy* (Summer 2022 Edition), Edward N. Zalta (ed.), forthcoming URL = <https://plato.stanford.edu/archives/sum2022/entries/conceptual-art/>

Sherwood, Yvonne (2021) *Blasphemy: A Very Short Introduction* Oxford: Oxford University Press.

Stecker, Robert (2005) "The Interaction of Ethical and Aesthetic Value" *The British Journal of Aesthetics* 45(2): 138–150.

Stecker, Robert (2019) *Intersections of Value: Art, Nature and the Everyday* Oxford: Oxford University Press.

Thompson, Michael (2008) *Life and Action: Elementary Structures of Practice and Practical Thought* Cambridge, MA.: Harvard University Press.

Vlastos, Gregory (1967) "Was Polus Refuted?" *The American Journal of Philology* 88(4): 454–60

Vlastos, Gregory (1971) "The Paradox of Socrates" in Vlastos, Gregory (ed.) *The Philosophy of Socrates: A Collection of Critical Essays* Garden City, New York: Anchor Books, 1-21.

Vlastos, Gregory (1991) *Socrates, Ironist and Moral Philosopher* Ithaca: New York: Cornell University Press.

Vlastos, Gregory (1994) *Socratic Studies* Cambridge: Cambridge University Press.

Walton, Kendall, (1970) "Categories of Art"*The Philosophical Review* 79(3): 334–367.

Wittgenstein, Ludwig (1953) *Philosophical Investigations* (*PI*) G.E.M. Anscombe and R. Rhees (eds.), G.E.M. Anscombe (trans.), Oxford: Blackwell.

Wolf, Susan (2015) *The Variety of Values: Essays on Morality, Meaning, & Love* Oxford: Oxford University Press.

Young, James O. (2001) *Art and Knowledge* London: Routledge.

Human Rights and Human Dignity

Lena Halldenius

Introduction

Let me take the liberty of opening this essay with a personal reflection and an anecdote. I came to philosophy having just finished my law degree. Uncertain of what to do now – all I knew was that I did not want to be a practicing lawyer – I took a course in moral philosophy out of a vague notion that it would be more intellectually stimulating than my law studies had been. And it was. But it was also frustrating in a way that prompted me to laboriously forge my own way through it.

I had chosen law because I was interested in social and political issues. My incentive to study philosophy was the same and has remained so. My philosophical starting point was and is the fact that we live together in political societies, that societies differ, change over time, and are always non-ideal. Norms we subscribe to or fight over and to what we ascribe value cannot be understood in isolation from this fact. At any rate, I am not interested in trying to understand such things in isolation from this fact. For me, there is no distinction to be made between philosophy and politics. Philosophy and philosophers are part of the messy social world we live in. We can only do our best to try and understand it and, particularly if you do moral or political philosophy, what the problems are in the ways we live in societies and how we can make them better. Philosophy is not its own precinct. It gives us tools and methods, not a licence to discount knowledge and experiences that are not philosophical, or that are philosophical only not produced by philosophers but by historians, disability scholars, feminist activists, or even lawyers. If you think I am selling philosophy short, that's fine. We do not need to agree on this.

How to do philosophy while still being deeply committed politically took me some time to figure out, and it started already at that first moral philosophy course. I first felt that I was expected to leave everything behind, that nothing I already knew mattered. We read the history of ethics, not as in thinking about ethical

questions in different historical circumstances but as a parade of men, selected – as the preface to the textbook put it – for their "greatness". I was puzzled. It seemed to me that these philosophers from the past were quite legitimately addressing ethical questions emanating out of their own times, challenges faced by their own societies, but that was not supposed to matter. What makes the great men great? Needless to say, there were no women in the parade. There were no women philosophers at all on any reading list, but plenty of great men declaring that women are intellectually less able than men. We were, in effect, taught that women make bad philosophers. One woman thinker, however, turned out to be indirectly present, which brings me to my anecdote.

We were reading J. S. Mill's *On Liberty*, a text written very much from within a political experience, and I loved it. The lecturer drew our attention to Mill's opening dedication, in which he laments the loss of his intellectual companion and wife, acknowledging her co-authorship of this work as well as all others written since they started working together. Our lecturer was apologetic, saying that he usually did not mention personal details of a philosopher's life (why not? I thought), but that Harriet Taylor Mill's contribution begged to be noted since it had been suppressed, both at the time and after, even though Mill himself recognized it. That lecturer was Dan Egonsson. Without knowing it, he got me thinking about how politically fraught and precarious the writing and publishing of philosophy have always been, and about that undercurrent of voices that are trivialized, ignored, or actively written out of the collective well of philosophical thought.

The voices that are not considered great, are they discounted because of their views or because of their perceived lack of fitness for philosophical thought? The intellectual capacities deemed required for philosophy – reason, intellectual autonomy and discipline, imagination, and judgement – were the same capacities that supposedly set humanity apart from the rest of the sentient world, the capacities that ground the moral dignity of man, but also the same capacities that women, according to most of the great men we read, do not possess. Given the incurable "inferiority and infirmities" of women and the natural "sovereignty of the male"[1], are women human?

Against this background, I will now go on to consider the philosophical construction of the dignity of the human and then to a critical but hopefully constructive engagement with Dan Egonsson's reflections on social attitudes about the dignity of humanity and how and why moral philosophers ought to account for them. I agree that philosophers should account for actual social attitudes, but my discussion will tend towards the conclusion that "accounting for" should be a consciously political and critical exercise, lest philosophers risk reproducing prejudice and bias.

[1] David Hume, quoted in Battersby, 1981.

History, Rights, and the Dignity of "Man"

History – including philosophical history – is contested terrain; it is also inevitably a construction. As E. H. Carr famously said in *What is History?* in 1961, "facts speak only when the historian calls on them". What becomes a *historical* fact rather than just stuff that happened is constructed in the process of writing history – also philosophical history. Carr's point is that which facts, ideas, or voices that are given the floor in the writing of history – also philosophical history – is a decision, not a discovery. And it's a decision that will be shaped by whatever it is that the historian – or philosopher – wants to show and finds to be significant and worthy of attention.

Harriet Taylor Mill's role in writing *On Liberty* is still debated and the works she co-wrote are still reissued under J. S. Mill's name alone.[2] Why? Because acknowledging her contribution would diminish his greatness? But if we know that a text about political liberty is co-written by someone whose person in the eyes of the law was "suspended" during marriage, "incorporated" into the person of her husband (Blackstone, 1765, 442), does that not add a certain urgency to the philosophy? It seemed to me that philosophy written by those who do not fit the bill of greatness – not then and not now – is a kind of performative act, an acting out of intellectual liberty denied, a claiming of rights not recognized, also against the philosophy of the great who, with few exceptions, have had quite a lot to say about the intellectual and moral inferiority of women and people of colour. We were not taught those bits, but this is a struggle worthy of our attention.

When I started studying early-modern philosophy properly – particularly Mary Wollstonecraft – it was even clearer to me that this is a struggle not only over who is included in the philosophical canon but over humanity itself. True to the Enlightenment project, Wollstonecraft says in her *A Vindication of the Rights of Men* that capacity for reason and moral improvement puts us above "brute creation" (Wollstonecraft 1995 [1790], 33), with "us" meaning us humans. Her feminist project is to analyse and disclose the denial to women of exactly those capacities that the philosophy to which she subscribes regards as distinctive of human nature. Generally accepted norms – that women are formed by nature to be "domestic brutes" as she put it in *A Vindication of the Rights of Woman* (1995 [1792], 88) and thought to be the rightful property of men (Wollstonecraft, 2009 [1798], 109) – were validated by the philosophy of the great. How do you write yourself into an intellectual world where contingent political hierarchies and widely accepted moral attitudes are elevated to the natural order of things and the analytical categories designed to exclude you?

In *The Politics of the Human* (2015), Anne Phillips notes that the language of the human is ostensibly one of inclusion, yet any definition of the human serves to exclude, and will inevitably be a matter of history and politics rather than objective

[2] On the controversy over Harriet Taylor Mill's co-authorship of *On Liberty*, see McCabe, 2021, 252-254.

observation. Any distinction found to be crucial will be contingent upon attitudes that could have – should have? – been different but which, once established, shape the way we think and what we take for granted.

Some evidence suggests that the idea that there is something particularly morally noble or dignified about being human *per se*, rather than being noble among the human, is a product of Renaissance humanism (Phillips, 2015, 23). This alleged nobility of being "human" seems always to have been tied to moral and intellectual capacities ostensibly associated with humanity while leaving a crack open for the question if all humans are human in that sense. The one category whose human dignity has never been in doubt in the minds of the great, is white men of independent means. When Wollstonecraft writes in 1792, in the preface to *A Vindication of the Rights of Woman*, that she pleads for her sex, not for herself, she means that she pleads for the humanity of women, for their equal status as moral subjects and equal capacity for virtue. The struggle was over humanity, since the human rights men claimed for themselves as humans were founded on those same mental capacities that women were denied. "Woman" thus turns into a mongrel concept: a human being but somehow less so, more brutish. Immanuel Kant famously left women "in an unresolved middle position" between moral agent and thing (Halldenius, 2011), "an anomalous kind of human being whose moral predisposition never fully develops" (Kleingeld, 1999, 64). Deliberations over who qualifies into the moral domain of the human is a philosophical staple in the works of the great. There is in that sense a direct line between Kant in the late 18th century and James Griffin, who in 2008 relegates "mental defectives" to the margins of rights bearing humanity (2008, 44).

Since I work in the field of human rights philosophy – both of today and its early-modern incarnations – I have had ample reason to grapple with the political implications of the association between "human" and certain mental capacities as foundation for subjectivity and rights. Humans have rights "simply in virtue of their humanity", as the saying goes (Cruft & Liao, 2015, 4f), but that "humanity" is morally loaded.

Philosophers working on the concept of human rights rather than related value concepts, like freedom or justice, are more prone to invoking historical precedence for their own ideas. It is not obvious what it is about human rights that prompts history to make itself felt in this way; I just note that historical precedence is made to serve in argument for various understandings of human rights in a way that has to do more with the "human" than with the "rights". James Griffin is an example of this. He claims that "our" concept of human rights (that is, the concept he himself favours and defends) is a product of eighteenth-century Enlightenment philosophy (2008) and has not changed since then. It is ours through inheritance. I have critically analysed Griffin's account of human rights elsewhere (Halldenius 2016). Here I merely let him be an example of the claim that there is in history an account to be observed or discovered – *our* account no less – about the moral rights-bearing status of the human such that "human rights" refers to whatever claimable political

and social arrangement this particular account of the moral human validates. According to Griffin, the historically received moral status of the human is the human person's capacity to form and pursue a worthwhile life – normative agency – and this is what grounds human dignity and rights. "Human rights" is the name we give to the requisite set of protections of this human capacity (Griffin 2008, 13), which seems simultaneously to be a feature of humanity as such and a practical skill that some humans have while some do not. This means that even though human rights are held "simply in virtue of being human", some humans have no human rights. Is this the philosophy of human rights handed down to us by the Enlightenment tradition, Wollstonecraft's tradition?

Remember Carr: "facts speak only when the historian calls on them". Selectively putting history to use in service to one's own preferred conception about morality, the message is sold as if it could not be in any other way. When Griffin claims historical precedence for a particular account of personal dignity by drawing a line from ancient natural law to "us", he skirts several alternative sets of historical facts.[3] In Rome "dignitas" referred to a man's social distinction and power, for a Medieval thinker dignity is a kind of intellectual nobility, a striving for glory through truth (Robiglio, 2006), while for Enlightenment radical philosophers, like Wollstonecraft and Thomas Paine, dignity marked the pride of someone who is just and virtuous but also independent of arbitrary and unaccountable political rule.[4]

It is certainly true that in early-modern accounts of natural rights from John Locke to Thomas Paine and Mary Wollstonecraft it is as given that the capacity for reason distinguishes the rights-bearing subject. It is precisely because of this that these moral capacities associated with humanity feature so prominently in early-modern debates about equality. But if reason is intrinsic to human nature, reason can never legitimately serve in arguments to privilege some humans over others, be it men over women, or great men over "idiots".[5] As Wollstonecraft puts it: "Who made man the exclusive judge, if woman partake with him the gift of reason?" (1995 [1792], 69). It is a rhetorical question and a political challenge rolled into one. If

[3] Griffin's claim is that the human person's moral capacity to form and pursue a worthwhile life has grounded dignity and rights since "Greek and Roman antiquity" (9), via Medieval theological accounts of man's innate disposition to reason (176), to early-modern secular ideas of natural rights and the Rights of Man. Since the seventeenth and eighteenth centuries, there 'has been no theoretical development of the idea itself' (13).

[4] Here is Thomas Paine in *Common Sense*, attacking Sir John Dalrymple: "he who can calmly hear and digest such doctrine [American obligation to the King of England], hath forfeited his claim to rationality—an apostate from the order of manhood—and ought to be considered as one who hath not only given up the proper dignity of man, but sunk himself beneath the rank of animals, and contemptibly crawls through the world like a worm." (Paine, 1776).

[5] See Simon Jarrett's fascinating historical analysis of the idea of the disabled mind, and how a largely tolerant social inclusion of "idiots" was replaced over the course of the nineteenth century, and as the medical professions claims to "expertise" grew, by loathing, contempt, and isolation from society (Jarrett, 2020).

reason is the moral mark of humanity, then it should be a foregone conclusion that women have the same moral standing as men. Given that women instead were not granted the same moral standing as men – and were denied political rights on that same ground – there were only two possible logical conclusions: either women are subjected to illegitimate oppression by having their natural and equal rights denied, or women are not human. Which is it? That question still resonates in human rights theory and moral philosophy alike.

For Griffin, "humanity" is a shorthand for the "dignity of the human person". Personhood or normative agency is what grounds rights (Griffin 2008, 152). Jeremy Waldron prefers to think of dignity as a status rather than a value-concept but this status is a normative one, a moral rank or bearing that does not ground rights exactly, but rather instantiates them (Waldron 2015, 2012). The dignified humanity is a kind of moral aristocracy with only one (high) rank, or a caste society with only one (high) caste (2015, 34). If there is a standard attitude in human rights philosophy, it includes a variation of the twin idea that human beings have rights by virtue of their humanity, where "humanity" is this moral construction.

As these reflections indicate, the moral standing of the human has political and legal implications way outside the bounds of moral philosophy. If having or not having human rights is predicated on having or not having – or being believed to have or not have – certain mental capacities, then we are inevitably faced with an uncomfortable challenge. Those people whose rights bearing capacity – and by implication their political and legal status – is placed in doubt in these discussions are invariably people whose lives in society are already disproportionately vulnerable to risk, poverty, discrimination, and prejudice, and who are in disproportionate need of the protection and security provided by democratic and social institutions. For most philosophers, the very notion of human rights seems to bring with it, or imply, or be premised upon a certain moral *je ne sais quoi* of the human. So what is it?

A Discussion of Egonsson's *Standard Attitude* to Human Dignity

In his book *Dimensions of Dignity. The Moral Importance of Being Human* (1998), Dan Egonsson proceeds from the intuition that there is "something morally special about being human" (1998, 54) or, in the words of Roger Wertheimer: "being human has *moral cachet*" (quoted in Egonsson 1998, 33). In the remainder of this essay, I will reflect on this "something morally special" – human dignity – with Egonsson as my starting and focal point. I will end up being partially critical of Egonsson's account, but I wish to stress my appreciation of his reflective, meandering way of going about it, inviting discussion rather than closing it down through assertiveness.

The main question for a moral philosopher, one might think, is whether it is morally justified to give precedence to humans and, if so, on what grounds, but Dan Egonsson wants us first to consider what attitudes people actually hold. He proceeds from what he refers to as the Standard Attitude (SA) to human dignity, which includes the intuition I just mentioned, that there is "something morally special about being human" (Egonsson 1998, 54). Egonsson believes this attitude to be widely shared and that moral philosophers should take it into account precisely for that reason.

The Standard Attitude, then, is not a normative principle; it does not dictate what we should or should not do and one is not blameworthy for not sharing it. It is an observation (I will return to the empirical soundness of it) informing the hypothesis that most people – not only philosophers – do think that there is something morally special about being human and that this status justifies giving priority to humans over non-human animals. The further claim is that the attitude held by people is that being human *per se* has this moral quality; it is part of what it is to be a human being, intrinsic to the human species (Ibid., 127). We will have reason to return to that as well.

Egonsson's SA is partly inferred from behaviour: the claim is that people tend to act as if they make a moral distinction between humans and non-human animals even if, when asked, they might not readily accept the distinction. Unpacking SA, we find that it features *centrally* in our minds (in the sense that it affects many of our moral opinions), that it entails regarding human beings as *inviolable* (once they exist), as *irreplaceable* (one human cannot be swapped for another without moral loss), and as *equal* in the fundamental sense that whatever value one gives to human life is given to all humans (Ibid., 91-103).

This looks like a package deal, but Egonsson quickly dismisses equality as an integral aspect of SA, for the reason that it does not seem to tally with behaviour after all: "In what sense can we be said to live as if we believe that all human beings are *equally valuable*?" (Ibid., 103, my emphasis). Fair question, but let's pause here for a bit. Egonsson emphasizes early on that SA in the "Western tradition" is derived from Christianity rather than Aristotle (Ibid., 4) and that it importantly cuts two ways: human beings are more valuable than non-human animals *and* no human being is more valuable than another human being. The prince and the slave are equally created in the image of God, as it were.

So, SA is supposed to be an attitude not only about human beings compared to non-human beings, but human beings compared to each other. The aspects of SA mentioned above – inviolability, irreplaceability, and equality – stand in different relations to these two comparisons. Inviolability and irreplaceability indicate a difference between our attitudes to humans compared to non-human animals. Animals are typically not inviolable to us, at least not if they are categorized as food, vermin, or a nuisance to human interests. Animals can be inviolable for religious reasons (like the sacred status of cows in Hinduism) or for reasons of conservation (like endangered species which, ironically, are endangered because of what humans have done to their habitat) but, and in tension with the main message of SA, animals

can take on the standing of both inviolability *and* irreplaceability in the minds of human beings. The personal bond between a dog or cat and its owner can have an emotional quality indistinguishable from relationships between humans. This should be familiar to us.

In a moving essay about grief, the author V. S. Naipaul writes about the death of his father, brother, and cat Augustus. He also writes about his sister's cat, who was killed by wild dogs. "Grief for that particular cat, whose ways she knew so well, almost like the ways of a person, never left her." (Naipaul, 2020). Naipaul's sister was not grieving for *a* cat but for that particular individual cat, who was indeed irreplaceable to her, as Augustus also came to be for Naipaul. But there remains a difference of some importance, I think. The moral status of the beloved dog or cat does not generalize to the category of dogs or cats. The fact that some dogs are loved just like children or friends, does not seem to impart inviolability or irreplaceability on the stray dog that no one cares about. We might be ready to impart inviolability and irreplaceability on particular animals but not on all, and not equality of standing between them. This is a further reason for supposing that the specificity of SA as something separating humans from non-humans needs the criterion of equality in order to not collapse.

Equality is the one aspect of SA that refers directly to the comparison *between* humans. If we dismiss equality as integral to SA, then SA does no longer cut in the double way that Egonsson claims for it. I agree that people in general do not live as if they believe that all human beings are *equally valuable* but SA, as I understand it, does not require that we regard all human beings as equally valuable, only that we regard them as *equally human* in the fundamental sense that all humans have dignity, or are above the threshold of dignity, or however we'd like to put it. Equality here refers to being equally reckoned among beings with dignity and is, it seems to me, indispensable for SA. If SA is supposed to be an attitude about the moral worth of being human *per se* then it is incompatible with allowing that some human beings are less (valuable as) human than others.

I suspect there is something else lurking here: the attitude that humans have a "moral cachet" that non-humans do not have is perhaps more consistent and stable than the attitude that all humans are equally valuable as humans. As the slaves of this world have continually pointed out (often at high cost to themselves), the abstract doctrine that all humans are equal in human dignity has never tallied very well with actual human behaviour nor with how societies have actually functioned, but the same can be said of the other aspects of SA. There is scant reason to find "equality" to be more of a behavioural outlier than, say, inviolability. Are we really comfortable in concluding that humans on the whole act as if they believe that all human beings (once they exist) are inviolable? To exemplify: the backlash against abortion rights in the United States – exemplified by the threat to the legal precedence of Roe v. Wade – while children being gunned down in schools are treated as an acceptable cost for keeping up the right to carry assault weapons indicates substantial support for the attitude that human beings are inviolable only *before* they exist.

If we as philosophers want to account for generally held attitudes in our normative theorising, then the question arises what counts as evidence of a generally held attitude. Philosophers' own intuitions are probably a poor guide. Another challenge is that there very likely is a disconnect between attitudes that people express when they are explicitly asked for them, say in questionnaires, and actual behaviour, particularly in social and political situations where there is peer pressure, prejudice, and conflicting interests. If SA is to be inferred from behaviour, then what is relevant behaviour and how do we know? Here is an example.

As we have seen, Egonsson dismisses equality as an integral aspect of SA because it does not fit with how we live. I have claimed that if equality is not part of SA, then SA is no longer an attitude regarding the value of being human *per se* and cannot serve in intra-human comparisons. But dismissing equality also jars against Egonsson's own defence of SA against a "serious objection" (1998, 86f). The objection is this: let's accept that philosophers should account for SA in their moral theory, on the ground that SA is widely accepted.[6] But on the same argument, should philosophers not also account for other attitudes and preferences that are less palatable, like racist and sexist attitudes? If most people prefer humans to non-human animals, but also prefer some humans (say male, able-bodied, and white humans) to others, then why should moral philosophers account for the first and not the second?

Egonsson does not think there is need to worry, mainly because "we have to remember that nowadays there is an almost world-wide and strong opinion against racism" (1998, 87). There are two problems with this, though. First, if this faith in the unbiased attitudes of people in general really were true, it would dismantle Egonsson's argument for dismissing equality as part of SA, since it would then indeed be the case that "all human beings are equally valuable", at least in terms of race. But, and second, there is no evidence for believing that it is true, at least not if we are to infer attitudes from behaviour. The state of our world – migration policies that favour white and rich over black and poor, misogynist violence against women, the fact that people of colour are poorer and their political liberties more precarious – suggests that we have to remember the exact opposite.

There is no stability to be had for SA if it is supposed to be inferred from actual behaviour. Alternatives are that SA can be inferred from doctrinal beliefs (the association to the Christian tradition suggests as much) or from beliefs that people report when they are explicitly asked to report them, like in questionnaires or controlled psychological studies. So let's look at that, in a roundabout kind of way, remembering that SA is the combined attitude that human life has more value than non-human life and that no human life is more valuable – as human life – than another.

[6] "Account for" will manifest differently depending on one's moral philosophy. For a Kantian, accounting for SA could be to use it in support of a deontological principle that it is always right in itself to favour humans over non-humans. For a utilitarian, the argument could be that not accounting for SA in moral theory would harm people's preference for it.

Dignity and Personalism

In SA "human beings" refers to biological human life (Egonsson 1998, 34): being a human life *per se* has moral cachet. But in real life it is not easy to disentangle attitudes towards the value of something from attitudes towards the value of things that are customarily associated with it. Is it the property of being biologically human that endows a being with dignity, or characteristics typically associated with human life, like such intellectual and emotional capacities of reason and self-awareness that feature in conceptions of human personhood? Egonsson discusses moral theories that attribute dignity to human life indirectly, via various capacities or religiously motivated notions about the human spirit or soul. He hypothesizes that a belief that being human has dignity in itself can be a shorthand, whereby people come to associate the value of a property of a thing with the thing itself (35). That is indeed quite possible, but it is not obvious how it affects SA. SA is not a moral principle; it is a claim about moral attitudes held by "most people" combined with an argument that moral philosophers should account for widely held attitudes simply because they are widely held, not because they are widely held *and* correct. SA is the attitude that human life has dignity *per se*, regardless of what reasons (mistaken or not), if any, that people have for holding it (if they hold it). SA presumably is agnostic regarding what the *per se* refers to: either the bareness of biological human life or certain morally relevant capacities or properties regarded as intrinsic to human life. In the first case, the moral valuation of human life is extrinsic to it, or ascribed to it, for reasons that could be anything. In the second case, the moral valuation is intrinsic; the moral value of human life is, as it were, found in it. There is something peculiar here that warrants a quick dip into the cultural history of moral attitudes.

Egonsson explicitly links SA to "the West", and the West to a largely Christian cultural and moral sphere. The influence of Christianity on our notions of morality in "the West" is certainly undeniable – even for committed atheists – and this holds particularly regarding the dignity of the human. But in the Christian moral-cultural-religious context, human dignity is very much a matter of personhood, not species membership. I venture to say that one theistic notion that has shaped philosophical thinking in the West more than most others is personalism.[7] Personalism in its modern form expresses a form of human exceptionalism: a binary distinction, not a gradual one, between humans and non-humans. Humans have higher value than other beings, not only in their own estimation but *tout court*. Only humans have full subjectivity, only humans are some*one* rather than somet*hing*. This worth, standing, or subjectivity – call it dignity – is to do with man's capacity for reason, self-awareness, and self-determination. Importantly, in classical personalism these dignity-inducing capacities are not characteristics that are checked one individual at

[7] On "personalism" in philosophy, see Williams and Bengtsson, 2022. On the influence of personalism on human rights doctrines in the 20th century, see Lindkvist 2017, and Moyn 2015.

the time or inferred from what actual individuals are like; they are believed to be inherent to human nature. They are an aspect of the metaphysics of humanity, not skills or functions that some human individuals have while others do not. In other words, all human beings have dignity because all human beings are, by virtue of their intrinsic nature, persons. This is the second case referred to: the moral value of human life is *found* in it. All humans are inviolable and irreplaceable by virtue of their personhood, while individuated objects and non-persons are not. (Needless to say, this theist belief in the equal dignity of all humans has always clashed rather brutally with the realities of the world and continues to do so. This harks back to my hunch that the dignity of the human is inferred from doctrinal belief rather than behaviour.)

In modern secular moral thinking, this metaphysical idea has been half-heartedly discarded, and a distinction introduced between human being and human person, such that personhood is treated as an added extra on top of being human. Personhood is now a practical skillset that most but not all individuated human beings develop (remember Griffin's "mental defectives"), and which can be lost again, for example through dementia. On this position, all persons are human beings, but not all human beings are persons. Personhood becomes an empirical question, thus generating debates over the moral standing of specific groups of people, like babies and people with severe cognitive disabilities. (I acknowledge but will not go into the ongoing discussions about non-human animals having personhood by virtue of their cognitive capacities.) One point worth emphasizing is that it is only on this latter position that it makes sense to distinguish between attributing dignity directly or indirectly. On the personalist position, making the distinction is the mistake.

It is hard to deny that the moral notion that human dignity is grounded in personhood is thoroughly dispersed in Western Christian and secular culture. To the extent that SA insists that the moral standing of dignity – the inviolability, irreplaceability, and equality of human individuals – is tied to biological human existence, rather than to attribution of personhood to that existence, it looks rather queer in this context and cannot be inferred from doctrinal belief either. But this crisp moral distinction between human being and human person commonly made by moral philosophers and theologians might not be made by people in general. What do we know about that?

The Empirical and Political Reality of Attitudes About Dignity

If SA cannot be inferred from behaviour and not from doctrinal belief, can it be inferred from what people report when they are explicitly asked? When Egonsson's book was published in the late 1990s there was not much reliable data on that, but there is now. A recently published study investigated, through a series of controlled

tests, the bases of "moral anthropocentrism" (the view that humans have moral priority over other beings) as a psychological phenomenon (Caviola et. al., 2022). The study was designed to capture specifically whether the moral priority of humans is explained by a valuation of mental capacities (personhood) or by sheer speciesism. The conclusion is that people do give priority to humans over non-humans and that there is a mix of reasons why. Mere species-membership is one of them, but some moral weight is also given to mental capacity (Caviola et. al., 2022, 11). All else being equal, respondents give more weight to individuals with higher mental faculties, suggesting that belief in humans' higher mental capacities explains the speciesism. The results suggest that giving more moral weight to human life *per se* – as in SA – is in itself a thick judgement, infused with assumptions about what humans are capable of. This holds at least when people are explicitly asked to make these judgements in comparisons between humans and non-human rather than between different human beings. The latter complicates matters.

In a larger context of prioritizing and favouritism, the authors of the study note that there is ample evidence of people prioritizing members of their own human ingroup, based on nationality, religion, political views, etc., compared to humans associated with other groups. They also conclude that humans giving priority to humans when compared to non-human animals is an extended form of such ingroup favouritism (Caviola et. al., 2022, 16). Confronted by a choice between humans and non-humans, humans favour humans because they identify with humans. When the choice is instead between human individuals associated with different groups, the same process of ingroup favouritism will favour humans associated with a group to which one's affiliation is stronger. Swedes favouring Swedes over Syrians, Christians favouring Christians over Muslims, and men favouring men over women are, if this interpretation holds up, psychologically not much different from humans favouring humans over non-humans. A difference would be that there is a socially acceptable moral justification in the latter case, and since people like to be seen as morally upright, they are more likely to report biases that are socially accepted, like favouring humans over animals. But these are not stable moral categories; they are culturally and historically fluid and open to contestation, for good and bad. The more socially acceptable racist and sexist attitudes are, the less is the moral cost of signalling them. It should be remembered that racist and sexist attitudes typically entail assumptions of lesser mental capacities in disfavoured groups. Assumptions of lesser mental capacities – less dignity – in one group can therefore be explained by already existing hostility against that group, rather than the other way around.

If, as the research by Caviola et. al. suggests, humans favouring humans is an exercise of ingroup bias, not a universalized corrective to ingroup bias, then SA stands on shakier ground against "the serious objection" (Egonsson 1998, 86) that accommodating widely shared pro-human attitudes commits the philosopher to also accommodating racist, nationalist, sexist and similar attitudes if they are also widely shared, which they are. If we want to say that pro-human attitudes are morally

relevant while racist and sexist ones are not (Ibid., 87), we are already making normative statements.

As Goodhart points out, attitudes about dignity are socially constructed standards that are conditioned by dominant norms and understandings of appropriate practices (Goodhart 2018, 405). Socioeconomic and political inequalities shape social norms and peoples' attitudes to themselves and others and mark some people as less worthy of concern and respect, as well as of a voice and a decent wage. The social reality of dignity is shaped by power and convention but as such is also malleable by critical counter-discourses.

I agree with Dan Egonsson that moral (and political) philosophers should engage with social attitudes that people actually hold. As Wollstonecraft put it: "we must mix in the throng, and […], attain a knowledge of others" (1995 [1792], 196). But accounting for moral attitudes must also be to critically account for the non-ideal political reality in which they are formed, including how the precarity of vulnerable and subordinated groups feature in the legitimation of attitudes that mark them as less dignified, less human.

References

Battersby, Christine (1981) "An Enquiry concerning the Humean Woman". *Philosophy*, 56(217): 303-312.

Blackstone, William (1765) *Commentaries on the Laws of England. Book the First.* Oxford: Clarendon Press.

Carr, E H (2018 [1961]) *What is History?* London: Penguin Books.

Caviola, Lucius, Stefan Schubert, Guy Kahane, Nadira S. Faber (2022) "Humans first: Why people value animals less than humans". *Cognition*, 225(105139): 1-17.

Cruft, Rowan, S. Matthew Liao, & Massimo Renzo (Eds.) (2015), *Philosophical Foundations of Human Rights*. Oxford: Oxford University Press

Egonsson, Dan (1998) *Dimensions of Dignity. The Moral Importance of Being Huma*n. Dordrecht: Kluwer Academic Publishers.

Goodhart, Michael (2018) "Constructing dignity: Human rights as a praxis of egalitarian freedom". *Journal of Human Rights*, 17(4): 403-417.

Griffin, James (2008) *On Human Rights*. Oxford: Oxford University Press

Halldenius, Lena (2011) "Kant on Freedom and Obligation under Law". *Constellations. An International Journal of Critical and Democratic Theory*, 18(2): 170-189

Halldenius, Lena (2016) "On the Use and Abuse of History in Philosophy of Human Rights" in J. Ross Kjærgard & K-M Simonsen (Eds.) *Discursive Framings of Human Rights: Negotiating Agency and Victimhood*. Oxford: Routledge.

Jarrett, Simon (2020) *Those They Called Idiots: the Idea of the Disabled Mind from 1700 to the Present Day*. London: Reaktion Books.

Kleingeld, Pauline (1999) "Kant, History, and the Idea of Moral Development". *History of Philosophy Quarterly*, 16(1): 59-80

Lindkvist, Linde (2017) *Religious Freedom and the Universal Declaration of Human Rights*. Cambridge: Cambridge University Press.

McCabe, Helen (2021) *John Stuart Mill, Socialist*. Montreal: McGill-Queen's University Press

Mill, John Stuart (1982 [1859]) *On Liberty.* Gertrude Himmelfarb (Ed.). London: Penguin Classics.

Moyn, Samuel (2015) *Christian Human Rights*. Philadelphia: University of Pennsylvania Press.

Naipaul, V. S. (2020) "The Strangeness of Grief". *The New Yorker*, January 6, 2020 Issue.

Paine, Thomas (1776) *Common Sense*. Philadelphia: R. Bell.

Phillips, Anne (2015) *The Politics of the Human*. Cambridge: Cambridge University Press.

Robiglio, Andrea (2006) "The Thinker as a Noble Man (bene natus) and Preliminary Remarks on the Medieval Concepts of Nobility". Vivarium, 44(2-3): 205-247.

Waldron, Jeremy & Meir Dan-Cohen (2015) *Dignity, Rank, and Rights*. Oxford: Oxford University Press

Williams, Thomas D. and Jan Olof Bengtsson (2022). "Personalism". The Stanford Encyclopedia of Philosophy. Edward N. Zalta (Ed.), URL https://plato.stanford.edu/entries/personalism/

Wollstonecraft, Mary (1995 [1790, 1792]) *A Vindication of the Rights of Men* with *A Vindication of the Rights of Woman, and Hints*. Sylvana Tomaselli (Ed.). Cambridge: Cambridge University Press.

Wollstonecraft, Mary (2009 [1788, 1798]) *Mary* and *The Wrongs of Woman*. Gary Kelly (Ed.) Oxford: Oxford University Press. Revised Edition.

Petersson on Plural Harm

Jens Johansson

Abstract. The counterfactual comparative account of harm has counterintuitive implications in cases involving overdetermination and preemption. A popular strategy for dealing with these problems appeals to *plural harm*—several events being *jointly* harmful. Björn Petersson criticizes this strategy on the grounds that it conflicts with a strong intuition that helps to motivate the counterfactual comparative account, namely, that harming someone essentially involves making a difference for the worse for her. In this paper, I argue that Petersson's argument is unconvincing.

1. Introduction

Björn Petersson presents the following case:

> *Recommendations:* I just had this paper rejected by the *Journal of Overdetermination Studies*. According to that journal's strict policy, manuscripts are rejected when one of the two reviewers recommends rejection, regardless of what the other reviewer says. In this case, both reviewers recommended rejection. (Petersson, 2018: 841; wording slightly modified; name of case added.)

To avoid confusion, I am going to use 'Björn' to refer to Petersson *qua* character in *Recommendations*, and 'Petersson' to refer to Petersson *qua* actual philosopher. Petersson adds some further details to the case: first, Björn would have been better off had the paper not been rejected; second, since "the comforting effect of one positive review would have been outbalanced by the frustration created by being

rejected in spite of such a review" (2018: 842), he would not have been better off if only one reviewer—Reviewer #1 or Reviewer #2—had recommended rejection; and third, it holds for each reviewer that she would have recommended rejection even if the other one had not. It is clear from the context that Petersson also assumes, fourth, that if neither Reviewer #1 nor Reviewer #2 had recommended rejection, then the paper would not have been rejected.

As Petersson says, *Recommendations* is an instance of the widely discussed *overdetermination* problem for the *counterfactual comparative account* of the nature of harm (CCA). This account can be formulated as follows:

> CCA An event harms a person if and only if she would have been better off if it had not occurred.[1]

Because neither Reviewer #1's nor Reviewer #2's action (recommending rejection) leaves Björn worse off than he would have been had it not been performed, CCA implies that neither action harms Björn. But, Petersson suggests, intuitively each of the reviewers' actions does harm Björn. To make this more clearly intuitive, let us add to the case the further detail that the rejection caused Björn disappointment and sadness.

CCA also faces the same kind of problem in various cases involving *preemption*. Consider this case, in which one rejection preempts another:

> *Desk Rejection:* Dan just had this paper desk rejected by the Editor-in-Chief of the *Journal of Preemption Studies*. Having the paper desk rejected, in spite of its high quality, brings Dan disappointment and sadness. If the Editor-in-Chief had not desk rejected the paper, then it would instead have been desk rejected by the Associate Editor, which would have left Dan no better off than he actually is.

Intuitively—or so many would say—the Editor-in-Chief's action harms Dan. But CCA implies that it does not.

Much of the debate has focused on a more physically dramatic preemption case, in which Bobby Knight, a basketball coach notorious for his rage, attacks a philosopher (Norcross, 2005; see also, e.g., Boonin, 2014: 62–63; Bradley, 2012; Feit, 2015, forthcoming; Hanna, 2016; Immerman, 2022; Jedenheim Edling, 2022; Johansson and Risberg, 2019, 2022). Consider this version:

[1] CCA is defended in, e.g., Boonin, 2014; Bradley, 2009; Feit, 2015, 2019, 2021, forthcoming; Klocksiem, 2012; Parfit, 1984: 69; Petersson, 2018; Timmerman, 2019. CCA is intended as an account of *overall* harm (harmfulness all things considered), as opposed to *pro tanto* harm (harmfulness to some extent, or in some respect). Like Petersson, I will be concerned with overall harm only. Correspondingly, both in the formulation of CCA and throughout our discussion, 'better off' is shorthand for 'better off *overall*'—and 'worse off' is, of course, shorthand for 'worse off *overall*'. Being (worse) off overall simply amounts to having a higher (lower) lifetime well-being level.

> *Choking:* Bobby Knight interprets one of Toni's arguments, involving an evil demon, as an attempt to make fun of Knight's own character. One day he meets Toni and chokes him. If he hadn't choked Toni, he would have dismembered him.

Intuitively—or so many would say—the choking harms Toni. However, CCA entails that it does not. Indeed, assuming a parallel account of harm and benefit, CCA implies that the choking even *benefits* Toni, as he would have been even worse off without it.

A popular strategy for dealing with the overdetermination and preemption problems for CCA is to appeal to *plural harm*. This strategy is inspired by a suggestion made by Derek Parfit (1984: 70–72), but has primarily been developed by Neil Feit (2015, 2022, forthcoming; see also, e.g., Jedenheim Edling, 2022; Timmerman, 2019: 244, fn. 6). The basic idea is that while CCA, which concerns a *singular* event's harming someone, is entirely correct, we should add to it, roughly, that a *plurality* of several events harms a person—in other words, that several events *jointly* harm a person—insofar as she would have been better off had *none* of those events occurred. This approach, its proponents argue, yields reasonable results in the relevant overdetermination and preemption cases. For instance, while this approach does not allow us to say that each reviewer's action harms Björn in *Recommendations*, it does imply something in the vicinity—namely, that each reviewer's action belongs to a plurality that harms Björn.

Petersson (2018) argues, however, that the plural harm approach abandons part of the main motivation for CCA, namely, the intuition that *making a difference for the worse* for someone is essential to harming her. According to Petersson, we should therefore be content with CCA alone and reject the proposed addition about pluralities. The implication that many overdetermination and preemption cases involve much less harm than they initially appear to do, he argues, is in the end acceptable.

In my opinion, Petersson is right that the plural harm approach does not adequately deal with the overdetermination and preemption problems for CCA (Carlson, Johansson, and Risberg, 2023b; Johansson and Risberg, 2019). His argument for this conclusion, however, is unconvincing. That will be my main point. But I shall also briefly suggest that, ironically, CCA *itself* fails to respect the idea that making a difference for the worse for someone is essential to harming her.

2. Plural Harm

CCA proponents have advanced various different versions of the plural harm approach, but since Petersson's criticism is applicable to all of them he focuses on a fairly simple version. I shall call that version (only marginally modified) the *simple plural harm account*, or *SPH*, and formulate it as follows:

SPH A plurality of events P harms a person if and only if (a) she would have been better off had no event in P occurred, and (b) there is no proper subplurality P* of P such that she would have been better off had no event in P* occurred.[2]

Feit and others have explicated the main elements of SPH and similar principles as follows.

First, a proper subplurality of a plurality of events P is a plurality that contains only, but not all, events in P (e.g., Feit, 2015: 376).

Second, talk of a "plurality" of several events should not here be taken to carry any ontological commitment to some entity that somehow has those events as constituents—for instance, a singular compound event with those events as parts (e.g., Feit, 2015: 370). Instead, it should simply be understood as a way of speaking of *those events*; saying that a plurality of several events harms someone is just a way of saying that *they* harm her.

Third, a plurality can consist of a single event—and speaking of a plurality of a single event is simply a way of speaking of *that event* (e.g., Feit, 2015: 371). Since SPH should be understood as covering pluralities of any size, and any one-event plurality trivially satisfies (b), SPH entails CCA (though not vice versa).

Fourth, talk of harmful pluralities of several events should here be given a "non-distributive" reading: saying that a plurality of several events harms someone—that *they* harm her—is to say that they harm her *together*, not that *each* of them (or even that at least *one* of them) harms her (e.g., Feit, 2015: 370). Compare: neither Dan Egonsson's 2007 book, *Preference and Information* nor Toni Rønnow-Rasmussen's 2021 book, *The Value Gap* is 416 pages long, but together they are 416 pages long.[3] A distributive reading of the harmful pluralities talk would render SPH inconsistent with, and thus unable to assist, CCA. Suppose, for example, that P is a plurality of two singular events, *e*1 and *e*2, and that P satisfies (a) and (b). On a distributive reading of the harmful pluralities talk, it follows that *e*1 as well as *e*2 is itself harmful on SPH. But since P satisfies (b), neither *e*1 nor *e*2 can be harmful on CCA. Indeed, a distributive reading of the harmful pluralities talk would render SPH incoherent. For in addition to rendering SPH inconsistent with CCA, it would leave untouched

[2] Petersson (2018: 842) formulates the principle on which he focuses as follows: "A plurality of events harms A if and only if that plurality is the smallest plurality of events such that, if none of them had occurred, A would have been better off." Unlike SPH, this principle has the disadvantage of ruling out that each of several pluralities that are "tied for smallest" harms the person. See further Feit, 2015: 374–375; Jedenheim Edling, 2022: 1856–1857; Norcross, 2005: 170.

[3] Petersson says that the "term 'together' carries with it a flavour of togetherness or collectivity, suggesting intentional co-ordination, planning, we-thinking or interdependence" (2018, 846–847). In my view, that flavor is rather mild, as the book example illustrates. See also Feit, forthcoming: sect. 5.2. In any case, SPH emphatically should not be understood as invoking any kind of "intentional co-ordination, planning, we-thinking or interdependence."

the fact that SPH entails CCA (see the third clarificatory remark above, about one-event pluralities).

Fifth, the motivation for condition (b) is that an event that has no effect at all, however indirect, on a person's well-being is not plausibly involved in harming her (e.g., Feit, 2015: 370). For example, suppose that in *Recommendations*, a bear is sleeping far away, and that this event in no way affects Björn's well-being. Although Björn would have been better off had no event in the plurality consisting of the reviewers' two actions and the bear's sleeping occurred, this three-event plurality plausibly does not harm him—the bear's sleeping seems to be not even involved in harming him. SPH accommodates this judgment, as the plurality does not satisfy (b); it has a proper subplurality, consisting solely of the reviewers' actions, such that Björn would have been better off had no event in *it* occurred.

For reasons already indicated, SPH implies that in *Recommendations*, the two reviewers' actions together harm Björn: whereas neither of them leaves him worse off than he would have been had *it* not occurred, he is worse off than he would have been had *neither* of them occurred. As for *Choking*, Feit (2015: 381) argues that there must be some mental events in Bobby Knight's mind, such as certain feelings of rage, which explain why he would have dismembered the victim had he not choked him. The plurality consisting of those events and the choking, Feit claims, leaves the victim worse off than he would have been had none of them occurred (though each of its proper subpluralities leaves him no worse off than he would have been had no event in *it* occurred).[4] If this is right, this plurality harms Toni on SPH. Similarly, SPH proponents can say that in *Desk Rejection*, there must be some events in the Associate Editor's mind (such as an intention to desk reject the paper if given the chance) that explain why she would have desk rejected the paper if the Editor-in-Chief had not. Arguably, whereas the plurality consisting of those mental events does not leave Dan worse off than he would have been had none of *them* occurred (since the Editor-in-Chief would still have desk rejected the paper), and the Editor-in-Chief's desk rejecting the paper does not leave Dan worse off than he would have been had *it* not occurred (since the Associate Editor would then have desk rejected the paper), Dan is worse off than he would have been had no event in the plurality consisting of those mental events *and* the Editor-in-Chief's desk rejecting the paper occurred. If this is right, the latter plurality harms Dan on SPH.

Of course, none of this blocks CCA's—and thereby SPH's—implication that neither Reviewer #1's nor Reviewer #2's action harms Björn in *Recommendations*, that the Editor-in-Chief's action does not harm Dan in *Desk Rejection*, and that Bobby Knight's action does not harm Toni in *Choking*. However, SPH proponents contend that their view delivers a result that is good enough—namely, that each of those actions is at least *involved* in harming the respective victim, by belonging to a plurality that harms him.

[4] For reasons to doubt this claim, see Johansson and Risberg, 2019: 358–360. See also Jedenheim Edling, 2022: 1869–1871.

3. Petersson's Criticism

Again, SPH is supposed to be a way of saving CCA from the overdetermination and preemption problems. According to Petersson, however, SPH is inconsistent with part of the main motivation for CCA—namely, the intuition that harming someone essentially involves *making a difference for the worse* for her. As that intuition is a strong one, Petersson suggests, we should stick to CCA alone and deny SPH. He argues, moreover (though I shall not consider this particular move further), that the claim that there is little or no harm in the pertinent overdetermination and preemption cases has less radical consequences than one might think. For instance, he suggests, we can still claim in many such cases that the relevant agents act morally wrongly. (Petersson refrains, however, from accusing Björn's reviewers of wrongdoing in *Recommendations*.)

The reason that SPH is inconsistent with the idea that making a difference for the worse for someone is essential to harming her, Petersson argues, is that many pluralities that are harmful for someone on SPH make no difference for worse for her. Indeed, Petersson says, this is exemplified in precisely those kinds of overdetermination and preemption cases that SPH is primarily designed to handle. A plurality makes a difference for the worse for someone, Petersson apparently assumes, only if she would have been better off if the plurality had not occurred. However, according to Petersson, in the relevant overdetermination and preemption cases the pluralities that are harmful on SPH do not satisfy this condition. For any such plurality, Petersson contends, is such that if it had not occurred, then sufficiently many of the events in it would still have occurred, leaving the person no better off than she actually is.[5]

For instance, in *Recommendations*, as we have seen, the plurality consisting of the reviewers' acts of recommending rejection harms Björn on SPH. According to Petersson, however, if this plurality had not occurred, then one of the two actions in it would still have occurred, in which case Björn's paper would still have been rejected and he would have been no better off than he actually is. After all, a stipulation of the case is that it holds for each reviewer that she would have recommended rejection even if the other one had not. (Without this stipulation, CCA would not imply that each reviewer's action is harmless.) Hence, Petersson concludes, the plurality of the reviewers' acts makes no difference for the worse for Björn.

While Petersson does not discuss *Desk Rejection* or *Choking*, similar remarks seem to apply to them. Again, in *Desk Rejection*, the supposedly harmful plurality consists of the Editor-in-Chief's desk rejecting Dan's paper and the mental events that explain why the paper would have otherwise been desk rejected by the

[5] For a closely related argument, whose target is Parfit's view of a group of *agents* harming someone, see Petersson, 2004: 297–300. Cf. Gunnemyr, 2019: 408.

Associate Editor. Judging from his reasoning concerning *Recommendations*, Petersson would deny that this plurality makes any difference for the worse for Dan. Had this plurality not occurred, the argument would go, then either the Editor-in-Chief's desk rejecting the paper or the relevant mental events would still have occurred, leaving Dan no better off than he actually is. In particular, in the absence of the Editor-in-Chief's action, the relevant mental events would have resulted in the Associate Editor's desk rejecting the paper. In *Choking*, the allegedly harmful plurality consists of Bobby Knight's choking Toni and the mental events that explain why Knight would have otherwise dismembered Toni. Presumably, Petersson would claim that this plurality makes no difference for the worse for Toni, on the grounds that if it had not occurred, then either the choking or the relevant mental events would still have occurred, leaving him no better off than he actually is. In particular, in the absence of the choking, the relevant mental events would have resulted in Knight's dismembering Toni, leaving him even worse off than he actually is.

4. Response to Petersson

As I understand it, Petersson's argument can be reconstructed as follows (focusing, as he does, on *Recommendations*):

(1) A plurality of one or several events makes a difference for the worse for a person only if she would have been better off if it had not occurred. (premise)

(2) Björn would not have been better off if the plurality consisting of the reviewers' two actions had not occurred. (premise)

(3) The plurality consisting of the reviewers' two actions does not make a difference for the worse for Björn. (from 1, 2)

(4) If SPH is true, the plurality consisting of the reviewers' two actions harms Björn. (premise)

(5) If SPH is true, then a plurality can harm someone without making a difference for the worse for her. (from 3, 4)

(6) No adequate response on behalf of CCA to the overdetermination and preemption problems is such that if it is true, then a plurality can harm someone without making a difference for the worse for her. (premise)

(7) SPH is not an adequate response on behalf of CCA to the overdetermination and preemption problems. (from 5, 6)

This valid argument has four premises: (1), (2), (4), and (6). While (4) is highly plausible—and is, of course, precisely what SPH proponents want to highlight regarding *Recommendations*—the other three are questionable, at least when taken together.

4.1 Premise (1)

Contrary to (1), to begin with, there are many cases in which some plurality of one or several events leaves someone no worse off than she would have been without it, but in which it is nevertheless perfectly natural to say that it makes a difference for the worse for her. In perhaps the clearest such cases, the plurality in question is a singular action and there was some alternative action, which the agent *could* but *would* not have performed instead of the actual one, and which would have left the person better off (cf. Johansson and Risberg, 2019, forthcoming). To find suitable examples, we need look no further than our already familiar overdetermination and preemption cases. Consider the one-event plurality of Bobby Knight's choking Toni in *Choking*. Again, Toni would not have been better off if Knight had not choked him, as he would then have been dismembered. But given the additional stipulation that a third alternative available to Knight was to simply leave Toni alone, it seems entirely sensible to say that the choking makes a difference for the worse for Toni.[6]

While *Choking* might illustrate the present point especially forcefully, it is worth noting that similar remarks also apply to *Recommendations*, the very case on which Petersson focuses. Consider the one-event plurality of one reviewer's—say, Reviewer #1's—recommending rejection. Let us add to the case that one of Reviewer #1's available alternatives was to contact Reviewer #2 and persuade her that the paper deserves to be published, in which case the journal would have accepted the paper. While this is not what Reviewer #1 *would* have done had she not recommended rejection, its being something that she *could* have done renders it perfectly natural, it seems to me, to say that her recommending rejection makes a difference for the worse for Björn. Furthermore, similar remarks apply also to our other preemption case, *Desk Rejection*—just add the detail that the Editor-in-Chief could easily have convinced the Associate Editor that Dan's paper should be published.[7]

---

[6] Relatedly, in the debate on so-called *collective impact cases*, one important view—call it *Difference-Making*, or *DM*—is that if *O* is a morally significant outcome, then there is an *O*-based moral reason against an act only if the act would make a difference to whether *O* obtains (see, e.g., Nefsky, 2017: 2744). Even opponents of DM (such as Nefsky) regard it as highly respectable. But if (1) is true, surely it is also true that an act makes a difference to whether *O* obtains only if *O* would not have obtained had the act not been performed. If so, then *Choking* provides a simple counterexample to DM. Clearly, if Knight could have left Toni alone, and *O* = Toni's being hurt, then there is an *O*-based moral reason against the choking. It cannot be this easy to refute DM.

[7] Some might claim that one or several of *Choking*, *Recommendations*, and *Desk Rejection* show (1) to be mistaken even without the supposition of an alternative that would have left the person better off.

I have criticized (1)—one premise in Petersson's argument for SPH's not being an adequate response on behalf of CCA to the overdetermination and preemption problems. In a way, however, my criticism of (1) creates trouble for SPH. I have argued that in the cases at hand, the relevant action makes a difference for the worse for the victim, although it does not satisfy CCA's condition. A natural view is that making a difference for the worse for someone is sufficient for harming her. My criticism of (1) thus provides additional ammunition against CCA—additional, that is, to the counterintuitiveness of CCA's implication that the relevant actions are harmless—and thereby also against SPH (which, again, entails CCA). On the other hand, my criticism of (1) might also accentuate CCA proponents' need for something like SPH. For, plausibly, the more unappealing it is to deny that an action harms a person, the more a theorist who is committed to such a denial should want to be able to say that the action is at least *involved* in harming her, by belonging to plurality that harms her. In any case, as I have already emphasized (section 1), my aim is not to defend SPH (or CCA), but to criticize Petersson's argument.

4.2 Premise (2)

Premise (2) says that Björn would not have been better off had the plurality consisting of the reviewers' two actions not occurred. It is clear that if not both events in this plurality had occurred—that is, if it had not been the case that each of them occurs—then one of them would still have occurred, leaving Björn no better off than he actually is. But (2) is nonetheless questionable.[8]

Recall, to begin with, that speaking of the plurality of the reviewers' two actions is just a way of speaking about *them* (section 2). Thus, asking whether Björn would have been better off had this plurality not occurred is just a way of asking whether he would have been better off had *Reviewer #1's action and Reviewer #2's action* not occurred—that is, whether he is better off in the nearest possible world, *w*, in which *Reviewer #1's action and Reviewer #2's action* do not occur. And there is a very natural way of understanding that question on which the answer is, contrary to (2), Yes. For it seems clear that there is a perfectly natural reading of 'in *w*, Reviewer #1's action and Reviewer #2's action do not occur' on which it is true just in case *both* events *fail* to occur in *w*—that is, just in case *neither* of them occurs in *w*. Of course, such a reading would be irrelevant if it had to be distributive—as explained in section 2, the relevant pluralities talk should be understood non-distributively. In

I take no stand on this stronger claim. (One bad reason to accept it is that regardless of the agent's alternatives, the action makes a difference for the worse for the victim in the sense of making him worse off *afterwards* than he was *before*. This does not concern *lifetime* well-being—see footnote 1.) In any case, the stronger claim would of course only strengthen the criticism of (1).

[8] My criticism of (2) is closely related to, though much less detailed than, Feit's response to Petersson (Feit, forthcoming: sect. 5.2).

order to make sure, then, that our reading is non-distributive, let us focus on this formulation:

> (a) Reviewer #1's action and Reviewer #2's action are an interesting pair of actions, and do not occur in *w*.

Since neither Reviewer #1's action nor Reviewer #2's action is an interesting *pair* of actions, it is clear that (a) should be understood non-distributively. And surely there is a very natural reading of (a) on which it implies that *neither* action occurs in *w*. Indeed, it seems rather unnatural to read (a) in a way that allows it to be true even if one of the actions occurs in *w*.

The general point here has nothing in particular to do with events and their non-occurrence in non-actual possible worlds. To see this, return to one of our earlier examples (section 2). Suppose someone knows that Egonsson's *Preference and Information* is 176 pages and Rønnow-Rasmussen's *The Value Gap* is 240 pages, and says the following:

> (b) *Preference and Information* and *The Value Gap* are 416 pages long, and do not disappoint.

Like (a), (b) should clearly be understood non-distributively—neither book is 416 pages. Now, surely there is a very natural reading of (b) on which the part about non-disappointment is true just in case *neither* book disappoints. By contrast, it is rather unnatural to read (b) in a way that allows the part about non-disappointment to be true even if it is merely the case that *not both* books disappoint.

In short, even when we make sure that no distributive (and hence irrelevant) reading is being presupposed, the most natural thing to say is that contrary to (2), Björn would have been better off had the plurality of Reviewer #1's action and Reviewer #2's action not occurred. Nothing in this criticism of (2) is in conflict, or even tension, with the stipulation that if *not both* actions had been performed, then one of them would still have been performed.

4.3 A Possible Reply

I have criticized both (1) and (2). Of course, Petersson's argument fails even if I am partially mistaken and have only managed to show that one of (1) and (2) is false. However, Petersson might try to show that I am not only partially but wholly mistaken. In response to my criticism of (1) and (2), he might offer the following speech:

> Contributing to a *Festschrift* only to provide uncharitable interpretations of one of its recipients is, to be honest, to give with one hand and take away with the other. Let me try to set things straight.

> I can happily accept that there is a reading of (1) on which it is false. In particular, I have nothing against the suggestion that Reviewer #1's action, in the expanded version of *Recommendations* proposed by my critic—the version in which Reviewer #1 could have convinced Reviewer #2 to accept the paper—"makes a difference for the worse" for Björn, in some legitimate sense of that phrase. (After all, Reviewer #1's action leaves Björn worse off than one alternative action would have done.) However, there is also a reading of (1) on which it is true—and inconveniently for my critic, that also happens to the reading that I intend. All I mean by saying that a plurality "makes a difference for the worse" for someone is that she would have been better off if it had not occurred. This renders (1) trivially true.
>
> Similarly, I can happily accept that there is a natural (and non-distributive) reading of (2) on which it is false. In particular, I can grant that there is a natural (and non-distributive) reading on which a formulation like 'in *w*, Reviewer #1's action and Reviewer #2's action do not occur' is true just in case neither of those events occurs in *w*. However, there is also a (non-distributive) reading of (2) on which this premise is true—something that even my critic apparently acknowledges, despite his complaints about such a reading being "rather unnatural." And once again, that happens to be the reading that I intend. All I mean by saying that someone would have been better off if a given plurality had not occurred is that she would have been better off if *not all* events in the plurality had occurred. (Admittedly, in the case of a one-event plurality, this might be an awkward way of putting it. But that is of no real importance; after all, it is obviously true that a one-event plurality's, that is, a singular event's, leaving someone worse off than she would have been if *not all* events in that plurality had occurred is necessary and sufficient for it to leave her worse off than she would have been if *the event* had not occurred.) On my intended reading, (2) is clearly true—again, if not both of the reviewers' actions had occurred, then one of them would still have occurred, leaving Björn no better off than he actually is.

On this line of response, then, (1) and (2) should be understood as (1*) and (2*), respectively:

> (1*) A plurality of one or several events is such that a person would have been better off if it had not occurred only if she would have been better off if it had not occurred.
>
> (2*) Björn would not have been better off if *not all* events in the plurality consisting of the reviewers' two actions had occurred.

Since (1*) and (2*) are undeniably true, Petersson's imaginary speech is indeed a way of rescuing both (1) and (2).

4.4 Premise (6)

However, the above line of response seems to make a difference for the worse, so to speak, for the plausibility of the argument's final premise:

(6) No adequate response on behalf of CCA to the overdetermination and preemption problems is such that if it is true, then a plurality can harm someone without making a difference for the worse for her.

This premise reflects Petersson's claim that the intuition that harming someone essentially involves making a difference for the worse for her is part of the main motivation for adopting CCA in the first place, and thus something that a proper defense of CCA against overdetermination and preemption worries needs to respect. Interpreted not overly narrowly, (6) is plausible. Obviously, for example, CCA proponents would have no use for a view that entails that a plurality's harmfulness has nothing to do with whether there is some reasonably nearby possible world in which the person is better off.

However, recall that Petersson's imaginary speech would have us understand (1) as the trivial (1*), and (2) as (2*). Given this, all we can infer from these premises, conjoined with (4)—the unproblematic premise that if SPH is true, then the plurality consisting of the reviewers' two actions harms Björn—is the following:

(5*) If SPH is true, then a plurality can harm someone even if she would not have been better off if *not all* events in it had occurred.

And then, in order to yield (7)—the conclusion that SPH is not an adequate response on behalf of CCA to the overdetermination and preemption problems—premise (6) must be understood in the following, narrow way:

(6*) No adequate response on behalf of CCA to the overdetermination and preemption problems is such that if it is true, then a plurality can harm someone even if she would not have been better off if *not all* events in it had occurred.

But (6*) seems to me to lack support. Maybe Petersson is right that one main intuition underlying CCA is that harming someone essentially involves making a difference for the worse for her. However, it is difficult to believe that that intuition is fine-grained enough to distinguish between a plurality's leaving someone worse off than she would have been had *not all* events in it occurred, on the one hand, and its leaving her worse off than she would have been had *no* event in it occurred, on the other. In particular, with regard to a one-event plurality—which, after all, is what CCA is about—the distinction is, if present at all, subtle in the extreme. Clearly, the only way for *not all* events in a one-event plurality to occur is for *no* event in it to occur, and vice versa. As far as the alleged intuitive basis for CCA is concerned, then, it is hard to see why an appeal to the "no event" factor, as opposed to the "not all events" factor, should disqualify a view from being an adequate defense of CCA against overdetermination and preemption objections.

5. Concluding Remarks: CCA and Its Supposed Motivation

I want to conclude by briefly considering—without being able to go into any detail—two other cases, which illustrate problems for CCA that are not of the overdetermination or preemption kind. One reason these cases are interesting is that they present potential counterexamples to CCA. More important in the present context, however, is that these cases also suggest that CCA is *itself* incompatible with the idea that Petersson takes to be part of its intuitive foundation—again, that harming someone essentially involves making a difference for the worse for her.

The first case illustrates the so-called *failure to benefit* problem for CCA (Bradley, 2012: 397; Feit, 2019; Hanna, 2016; Johansson and Risberg, 2020, forthcoming; Klocksiem, 2022; Purves, 2019):

> *No Clubs:* Jörn contemplates giving a set of golf clubs to Peter, but eventually decides to keep them for himself. If Jörn had not decided to keep the clubs, he would have given them to Peter, which would have made Peter better off than he actually is. Peter never knows about any of this.

CCA implies that Jörn's decision to keep the clubs harms Peter. Intuitively, however, it merely fails to benefit Peter; it does not harm him. So, CCA has a counterintuitive implication here.

In addition, *No Clubs* suggests that CCA fails to respect what Petersson regards as part of its main motivation. Duncan Purves (2019; see also Klocksiem, 2022) argues that the reason CCA goes wrong in cases like this is that it fails to take seriously the distinction between *making* an upshot happen and *allowing* it to happen. In *No Clubs*, Purves would say that although Peter would have had a higher well-being level without Jörn's decision, the decision is still harmless to Peter, as it merely *allows* him to occupy—and does not *make* him occupy—his actual, lower well-being level. Whether or not Purves's proposal is right in its details, intuitively it does seem rather attractive to say that Jörn's decision does not *make* a difference for the worse for Peter. If this is right, an event can be harmful on CCA without making any difference for the worse for the person.

The second case illustrates the *mere indicators* problem—the problem that CCA apparently entails that some mere indicators of harm are themselves harmful (Carlson, Johansson, and Risberg, 2022: 422; Johansson and Risberg, forthcoming):

> *Omniscience:* Pete feels intense pain. As a result, Örn, who is an essentially omniscient being, forms the belief that Pete feels intense pain. If Örn had not formed that belief, that would have been because Pete didn't feel intense pain.

CCA implies that Örn's forming the belief that Pete feels intense pain harms Pete. Intuitively, however, this event does not harm Pete; it merely indicates that

something else does. In this case as well, then, CCA has a counterintuitive implication.

Importantly, moreover, it seems wrong to say that Örn's forming the belief *makes* a difference for the worse for Pete. After all, it is only the pain, and whatever has led up to it, that plausibly *affects* or *influences* Pete's well-being adversely. Once again, then, CCA appears to conflict with the idea that harming someone essentially involves making a difference for the worse for her.

In response to this latter charge—as well as to the corresponding charge based on *No Clubs*—Petersson might protest that in his vocabulary, an event's "making" a difference for the worse for someone does not involve anything other than the person's being worse off than she would have been had the event not occurred. (Cf. Petersson's imaginary speech in section 4.3.) Hence, Petersson might say that in his vocabulary, Jörn's decision and Örn's forming the belief actually do make a difference for the worse for Peter and Pete, respectively. If so, *No Clubs* and *Omniscience* fail to show that if CCA is true, there can be harming without negative difference-making. That 'making' can also be used to refer to something ontologically heavier, Petersson might claim, is irrelevant.

Of course, this response does not alter the fact that CCA implies, counterintuitively, that Jörn's decision and Örn's forming the belief are harmful. So that problem remains. In the present context, however, a related but more important problem is that it is simply an independently appealing idea that harming involves negative difference-making in some ontologically fairly heavy sense of 'making'—ontologically heavier, at least, than the sense suggested in Petersson's possible response.[9] Not only is this idea appealing when considered in isolation, it also gets support from cases like *No Clubs* and *Omniscience*. For, intuitively, Jörn's decision and Örn's forming the belief are harmless precisely *because* they do not really *affect* or *influence* Peter's or Pete's well-being negatively—they do not *make* a difference for the worse for Peter or Pete (again, in some ontologically fairly heavy sense). Hence, CCA's apparent inability to respect the idea that harming someone essentially involves making a difference for the worse for her, in some ontologically fairly heavy sense, is evidence against CCA.

In my opinion, this is only one of several mutually independent reasons to regard CCA as a seriously flawed view. (For other independent reasons, see, e.g., Carlson, 2019, 2020; Carlson, Johansson, and Risberg, 2021, 2022, 2023a; Johansson and Risberg, 2019, 2020, forthcoming.) If this is right, then CCA's being a seriously flawed view seems to be a suitable topic for the *Journal of Overdetermination Studies*.[10]

---

[9] For a defense of the closely related view that for an event to harm someone is for it to affect her well-being adversely, see Johansson and Risberg, forthcoming.

[10] For very helpful comments I am grateful to Mattias Gunnemyr, Magnus Jedenheim Edling, and Caroline Torpe Touborg. Work on this paper was supported by Grant 2018-01361 from Vetenskapsrådet and Grant P21-0462 from Riksbankens Jubileumsfond.

References

Boonin, David (2014) *The non-identity problem and the ethics of future people*. New York: Oxford University Press.

Bradley, Ben (2009) *Well-being and death*. New York: Oxford University Press.

Bradley, Ben (2012) "Doing away with harm". *Philosophy and Phenomenological Research,* 85(2): 390–412.

Carlson, Erik (2019) "More problems for the counterfactual comparative account of harm and benefit". *Ethical Theory and Moral Practice*, 22(4): 795–807.

Carlson, Erik (2020) "Reply to Klocksiem on the counterfactual comparative account of harm". *Ethical Theory and Moral Practice*, 23(2): 407–413.

Carlson, Erik, Jens Johansson, and Olle Risberg (2021) "Well-being counterfactualist accounts of harm and benefit". *Australasian Journal of Philosophy*, 99(1): 164–174.

Carlson, Erik, Jens Johansson, and Olle Risberg (2022) "Causal accounts of harming". *Pacific Philosophical Quarterly*, 103(2): 420–445.

Carlson, Erik, Jens Johansson, and Olle Risberg (2023a) "Benefits are better than harms: a reply to Feit". *Australasian Journal of Philosophy*. Published online ahead of print.

Carlson, Erik, Jens Johansson, and Olle Risberg (2023b) "Plural harm: plural problems". *Philosophical Studies*, 180(2): 553–565.

Feit, Neil (2015) "Plural harm". *Philosophy and Phenomenological Research*, 90(2): 361–388.

Feit, Neil (2019) "Harming by failing to benefit". *Ethical Theory and Moral Practice*, 22(4): 809–823.

Feit, Neil (2022) "How harms can be better than benefits: reply to Carlson, Johansson, and Risberg". *Australasian Journal of Philosophy,* 100(3): 628–633.

Feit, Neil (forthcoming) *Bad things: on the nature and normative role of harm.* New York: Oxford University Press.

Gunnemyr, Mattias (2019) "Causing global warming". *Ethical Theory and Moral Practice*, 22(2): 399–424.

Hanna, Nathan (2016) "Harm: omission, preemption, freedom". *Philosophy and Phenomenological Research*, 93(2): 251–273.

Immerman, Daniel (2022) "The worse than nothing account of harm and the preemption problem". *Journal of Moral Philosophy*, 19(1): 25–48.

Jedenheim Edling, Magnus (2022) "A new principle of plural harm". *Philosophical Studies*, 179(6): 1853–1872.

Johansson, Jens and Olle Risberg (2019) "The preemption problem". *Philosophical Studies*, 176(2): 351–365.

Johansson, Jens and Olle Risberg (2020) "Harming and failing to benefit: a reply to Purves". *Philosophical Studies,* 177(6): 1539–1548.

Johansson, Jens and Olle Risberg (2022) "Against the worse than nothing account of harm: a reply to Immerman". *Journal of Moral Philosophy*. Published online ahead of print.

Johansson, Jens and Olle Risberg (forthcoming) “A simple analysis of harm”. *Ergo*.

Klocksiem, Justin (2012) “A defense of the counterfactual comparative account of harm”. *American Philosophical Quarterly*, 49(4): 285–300.

Klocksiem, Justin (2022) “Harm, failing to benefit, and the counterfactual comparative account”. *Utilitas*. Published online ahead of print.

Nefsky, Julia (2017) “How you can help, without making a difference”. *Philosophical Studies*, 174(11): 2743–2767.

Norcross, Alastair (2005) “Harming in context”. *Philosophical Studies*, 123(1–2): 149–173.

Parfit, Derek (1984) *Reasons and persons*. Oxford: Oxford University Press.

Petersson, Björn (2004) “The second mistake in moral mathematics is not about the worth of mere participation”. *Utilitas*, 16(3): 288–315.

Petersson, Björn (2018) “Over-determined harms and harmless pluralities”. *Ethical Theory and Moral Practice*, 21(4): 841–850.

Purves, Duncan (2019) “Harming as making worse off”. *Philosophical Studies*, 176(10): 2629–2656.

Timmerman, Travis (2019) “A dilemma for Epicureanism”. *Philosophical Studies*, 176(1): 241–257.

Egalitarian Justice as a Challenge for the Value-Based Theory of Practical Reasons

Benjamin Kiesewetter

Abstract. In this essay, I argue that the objections that have been raised against the view that equality is intrinsically valuable also provide objections to the view that all practical reasons can be explained in terms of value. Plausible egalitarian principles entail that under certain conditions people have claims to an equal share. These claims entail reasons to distribute goods equally that cannot be explained by value if equality has no intrinsic value.

The relation between reasons and value has attracted a lot of attention in the recent meta-ethical literature. Some philosophers – Toni Rønnow-Rasmussen is one of their most prominent proponents – have explored the idea that value can be analyzed in terms of reasons.[1] Others have suggested that reasons must itself be explained in terms of value. While these views appear to be in tension at first sight, the most popular versions of them turn out to be consistent with one another: there is no contradiction in holding that all reasons *for action* are to be explained in terms of value, while at the same time maintaining that value is to be analyzed in terms of reasons *for attitudes.*

Some of the arguments in favour of a value-based view of practical reasons have to do with the alleged attraction of a much more general Value-First Approach to

[1] See esp. Rabinowicz and Rønnow-Rasmussen (2004), Rønnow-Rasmussen (2011) and Rønnow-Rasmussen (2022, Pt. II).

normativity.[2] But there are also arguments related to the nature of *action* in particular, which are therefore neutral on the question whether value or reasons *for attitudes* are explanatorily more fundamental. For example, it looks like the value-based theory of practical reasons harmonizes well with the so-called Guise of the Good Thesis, according to which intentional action always aims at some good.[3] For it seems plausible to think that acting intentionally involves acting for reasons, and that acting for reasons involves taking oneself to have a reason to act. The value-based theory of practical reasons suggests that taking something to be a reason for action involves taking the action to be good in some way or other. Together, these assumptions entail that acting intentionally involves taking the action to be good, thereby explaining (a version of) the Guise of the Good Thesis. But since this argument for the value-based theory appeals to an assumption that is specifically concerned with the nature of intentional action, it is neutral on the question of whether value is also prior to reasons for attitudes other than intentions and thus neutral on fitting attitude accounts of value.

Those who think that value is to be explained in terms of reasons for attitudes are, however, committed to a value-independent notion of a reason and should therefore be open to the possibility that some practical reasons are among those reasons that cannot be explained in terms of value. In this paper, I will argue that the considerations that have been brought forward against the view that equality is intrinsically valuable provide good reasons to reject the value-based theory of practical reasons: there are reasons of egalitarian justice that are not value-based. The argument complements structurally similar points I made in an earlier article that can be considered a companion of the present paper (Kiesewetter 2022). In this companion article, I argued that reasons created by the exercise of a normative power to obligate oneself or others, in particular reasons to keep one's (valid) promises and reasons to obey (legitimate) authorities, cannot plausibly be explained in terms of the value of the actions that they support. The normative assumptions on which this argument is build are, in my view, part of what is sometimes called 'commonsense morality', but they are clearly controversial among moral philosophers. The aim of the present paper is to strengthen the case against the value-based theory of practical reasons by providing another counterexample, which is independent of the assumption that we can create reasons by exercising normative powers. Needless to say, this argument also relies on normative assumptions that are not uncontroversial. However, if it can be shown that the value-based theory conflicts with a number of assumptions of commonsense morality that are independent of each other, this strengthens the case against it.

I shall start with introducing the value-based theory of reasons (§1) and rehearsing the challenge it faces with normative powers (§2). I will then turn to distributive

[2] See e.g. Maguire (2016) and Wedgwood (2017, Ch. 4). The Value-First Approach seems also in the background of many epistemic teleologists, such as Foley (1987) and Goldman (2001).

[3] See Anscombe (1957, 70–78) for a famous defence of the Guise of the Good Thesis.

justice, arguing that the value-based theory fails to accommodate reasons to distribute goods equally in certain cases (§3). I conclude by reflecting on the question of whether the counterexamples to the value-based theory can be unified, and by briefly addressing the implications of the argument for the Guise of the Good Thesis (§4).

1. The Value-Based Theory of Practical Reasons

The value-based theory that I am concerned with in this paper can be stated as the following claim:

> *The value-based theory of practical reasons (VBT)*: For all agents A, and all actions ϕ that A can perform: A has a reason to ϕ if, only if, and because ϕ-ing has value.

Some remarks are in order. As I understand it, VBT is neutral on whether the value of actions is instrumental or final. It includes theories according to which actions can be finally valuable, but also consequentialist theories according to which actions always derive their value from the fact that they are conducive to a finally valuable outcome. In line with the former view, Joseph Raz claims that "the only reason for any action is that the action, in itself or in its consequences, has good-making properties"[4]. The latter view is taken by Barry Maguire, who holds that "to be a reason for an option is to be a fact about that option's promoting some state of affairs, on the condition that the state of affairs is valuable".[5] Following Rønnow-Rasmussen (2002), one might deny that the property of being instrumentally valuable is a genuine value property (rather than simply the property of being conducive to a value). On this view, VBT should be understood as the view that all reasons for action are explained by the fact that the actions they support instantiate or promote a value.

Secondly, VBT is also neutral on whether the relevant value is personal or impersonal value, or whether it can be either. Roger Crisp provides an example for the former view, when he claims that "any ultimate reason for action must be grounded in well-being"[6]. Following Rønnow-Rasmussen's recent arguments for the mutual irreducibility of personal and impersonal value (2022), however, proponents of VBT seem well-advised to allow both kinds of value as grounds for practical reasons. Moreover, it seems to follow from his view that VBT is a less unfied theory than it appears on first sight.

Thirdly, as it is understood here, VBT claims that practical reasons are to be explained *directly* in terms of the value that complying with the reason instantiates

[4] Raz (2001, 2). See also Wedgwood (2009).

[5] Maguire (2016, 237).

[6] Crisp (2006, 37).

or promotes. On other views, at least some reasons are explained *indirectly* in terms of value, by appeal not to the value of compliance, but, for example, the value of having a general rule that requires compliance, or the value of having a general disposition to comply (to mention just two possibilities). Such views are beyond the scope of this paper.

2. The Argument from Choice-Based Reasons

Consider the following two principles:

> *The promising principle:* If A validly promises B to ϕ, then A has an obligation, and thus a reason, to ϕ.[7]
>
> *The authority principle*: If A has legitimate authority over B, and A validly commands B to ϕ, then B has an obligation, and thus a reason, to ϕ.[8]

The notion of an obligation figuring in these principles is meant to be a normative notion, which entails a reason for compliance. It is, moreover, meant to be a contributory rather than an overall normative notion.[9]

As indicated already, I take these principles to be elements of moral commonsense. Conscientious promisors have to believe, as such, that they have a reason to keep their valid promises and conscientious subordinates have to believe, as such, that they have a reason to obey the valid command of an authority they consider legitimate. Those who sincerely participate in the practices of promising or authority relationships are thus committed to accepting the mentioned principles.

Both of these principles entail that if certain conditions of validity and legitimacy are satisfied, persons have the power to create a reason for action by choosing to do so (namely, by choosing to promise or command an action), which is why we can call the resulting reasons *choice-dependent*. Such reasons contrast with *content-dependent* reasons, which are provided by features of the action they support rather than by the choice of a person who has the power to create it.

While it is clear how content-dependent reasons can be value-based, the existence of choice-dependent reasons is hard to reconcile with VBT. Since promises and commands can be valid even if the promised or commanded action is valueless, the

[7] Compare Raz's "promising principle", which is more general: "If a person communicates an intention to undertake by that very act of communication a certain obligation then he has that obligation" (Raz 1986, 173). As he makes clear, Raz takes his principle to entail that "we are obligated to perform action X, if we promised to perform X" (*ibid*.).

[8] Compare Raz (1986, 60): "What is validly required by a legitimate authority is one's duty".

[9] See Kiesewetter (forthcoming) for a defence of the view that obligations can be contributory.

principles entail that people can choose to create reasons for antecedently valueless actions. The only way for the proponent of VBT to accommodate such reasons is to argue that keeping a promise or obeying an authority is valuable *as such*. But it is difficult to see why these acts would be valuable as such if not because they are ways of discharging an obligation. That is, in order to establish that keeping one's promises and obeying a legitimate authority are valuable, we already have to assume that these acts are obligatory, and hence we already have to assume that we have a reason to perform these acts. Consequently, we cannot appeal to this value in order to explain the reason. This is, in a nutshell, the argument that promissory reasons and reasons to obey, are counterexamples to the value-based theory of reasons.[10] It is natural to think that these points generalize to all choice-dependent reasons.

However, while the assumption of choice-dependent reasons can be supported by reference to our pretheoretical conception of morality, it is also notoriously controversial among moral philosophers. The case against the value-based theory of practical reasons would therefore be stronger if it did not rely on it. In the following section, I shall present a new counterexample against VBT, which is independent of the existence of normative powers and choice-dependent reasons.

3. The Argument from Egalitarian Justice

The argument I wish to defend is based on the following principle:

> *The equal distribution principle:* If a number of persons are the only ones that have a claim to a share of some divisible good, and none of them has a claim to a greater share than any other, then each has a claim to an equal share, and agents in charge of distribution have an obligation, and thus a reason, to distribute the good equally.

To illustrate, suppose that Tommy and Annika spend the weekend picking apples and bring them to a juice-maker, who makes 100 bottles of apple juice out of the apples. Suppose that, for some reason, Tommy and Annika collect their shares of juice separately and the juice-maker is in the position to choose between two distributions. She could either give each of them 50 bottles, or she could give one 60 bottles and the other 40 bottles. The juice-maker knows that neither of them has invested more time or effort in collecting the apples and there is no other fact of the matter that grounds a claim to a greater share. In such a case, it seems compelling to think that the juice-maker has a moral reason to choose the equal distribution.

If there is a reason to distribute equally, then VBT entails that distributing equally is good. But what value is promoted or instantiated by equal distribution? Appealing to the law of diminishing marginal utility, one might argue that the ten

[10] For the longer version, see Kiesewetter (2022, 32–44).

bottles in question have a greater benefit for a person who has 40 bottles than for a person with 50 bottles, and that for this reason equal distribution promotes welfare (in the sense of maximizing the sum of welfare that Tommy and Annika receive). But we can stipulate that this is not the case. It seems conceivable that Tommy and Annika get the same benefit from each bottle of juice, so that their overall welfare is not promoted by an equal distribution. That does not change the fact that they have a claim to an equal share.

It is also plausible to think that by and large, equal distribution of goods will promote valuable social relationships and work against power imbalances that can create a danger for valuable forms of societies. But this is not to say that such a value will be promoted in each particular case. And it seems that if we assume that an unequal distribution of apple juice in this particular case will have no impact on social relationships and power balances, this does not change the fact that Tommy and Annika have a claim to an equal share.

According to what Derek Parfit calls *teleological egalitarianism*, equality is intrinsically valuable.[11] Drawing on this assumption, proponents of VBT might say that the value that is promoted by equal distribution is equality itself. But the view that equality is intrinsically valuable has forcefully been criticized. As Harry Frankfurt points out, equality is a purely formal property and it is difficult to see how such a property could be intrinsically valuable.[12] Moreover, as Parfit and others have argued, the assumption that we have reason to promote equality entails that we have reason to *destroy* substantive goods if this is what it takes to establish equality (the so-called *Levelling Down Objection*).[13] For example, if the juice-maker has the possibility to choose only between an unequal distribution of bottles (60:40), on the one hand, and destroying all bottles and leave both Tommy and Annika with nothing (0:0) on the other, then teleological egalitarianism entails that there is a solid value-based reason in favour of destroying all bottles. Of course, teleological egalitarians can also say that there are stronger, welfare-based reasons against destruction. Intuitively, however, we do not weigh a welfare-based reason against destruction against an equality-based reason for destruction in such situations. Unless further values are promoted by the destruction of some good, we assume that there is no reason to do that *at all*.[14]

The situation here contrasts with cases of conflicting values. Assume, for example, that people's liberties are restricted in order to protect the vulnerable from

[11] See Parfit (1997, 204). This view is held, among others, by Temkin (1993, 282).

[12] See Frankfurt (1997).

[13] See Parfit (1997, 210–11). See also Raz (1986, Ch. 9) and Temkin (1993, 247).

[14] Schroeder (2007, 92–97) argues that intuitions about the non-existence of reasons are not trustworthy in cases where these reasons would be massively outweighed. For some reasons to doubt this, see Kiesewetter and Gertken (2021, 275–76). Even if Schroeder is right, however, his view does not help with the Levelling Down Objection, for those who think that equality is an important value cannot plausibly assume that reasons based on this value are *massively* outweighed by welfare-based reasons.

severe risks, as it happened in most countries during the pandemic. Even if we grant, as many did, that this was justified or even required, we do not feel the temptation to say that there was *no reason* against restricting the liberties. There is a clear residual sense in which there would have been something good about not restricting the liberties, even if doing so was, all things considered, for the best. This sense of conflict seems to be absent in the case in which we have to decide between destroying a good (on the assumption that doing so promotes nothing but equality) and distributing it unequally.[15]

If equality is not intrinsically valuable, and only contingently related to other goods such as welfare, why do we have reason to distribute equally even in cases where this does not promote welfare or other values? According to *deontological egalitarianism*, unequal distribution is (under certain circumstances) unjust, or violates moral claim rights.[16] If an equal distribution of goods is possible in cases like the apple-picking example, then the persons involved have a right to an equal share. If they have a right to an equal share, then others have an obligation not to deny them their equal share by choosing an unequal rather than an equal distribution, and this obligation involves a moral reason for equal distribution. But saying this does not entail that equality has intrinsic value, or that persons also have a right to the destruction of goods if equal distribution is not possible. It thus avoids the above-mentioned objections to teleological egalitarianism.

In summary, there are strong reasons for thinking that the reason to distribute goods equally that figures in the equal distribution principle is not based on a presumed final value of equality, nor on any other value that equal distribution typically promotes. Rather, this reason seems to be a constitutive part of a claim to an equal share (or its corresponding obligation). If this is right, the equal distribution principle suggests an extensional argument against VBT: since there can be claims to an equal share even if there is no value in equal distribution, and such claims entail reasons to distribute equally, there can be reasons for valueless actions.

Proponents of VBT might reply that in cases in which the equal distribution principle entails a reason to distribute equally, equal distribution is good in virtue of being *just*. And indeed, this strikes me as a successful response to the extensional argument. But for an act to be just in the relevant sense is, in part, for it to satisfy a

[15] Temkin concedes that the Levelling Down objection has "tremendous force", but rejects it by way of arguing against another principle that he takes to be "at the heart" of the objection, namely: "*The Slogan:* One situation cannot be worse (or better) than another if there is no one for whom it is worse (or better)" (Temkin 1993, 248). The problem is that this claim is significantly stronger than the Levelling Down Objection, and the Levelling Down Objection has independent plausibility (see also Parfit 1997, 220). To say that for inequality to be bad, it must be bad for someone, is not to say that *nothing* can be bad if it isn't bad for someone. For this reason, one cannot reject the former claim by arguing against the latter.

[16] See Parfit (1997, §3) for an illuminating discussion of the differences between deontological and teleological egalitarianism.

claim.[17] And for someone to have a claim to a good is, at least in part, for others to have obligations and thus reasons to not deny her that good.[18] The reason to distribute equally is thus part of what *makes* equal distribution just and, in this respect, good. Consequently, it cannot be explained by this goodness. So even if reasons for distributing goods equally are always accompanied by a value, this value cannot explain these reasons.

A related reply on behalf of VBT is that in cases in which the equal distribution principle requires it, equal distribution is good in virtue of *showing due respect* for the involved parties. Here, again, proponents of VBT must be cautious not to assume that equal distribution is required by respect *because* the relevant people have a claim to an equal share; it must be disrespectful to distribute unequally independently of any presumed claim to an equal share. This seems to be Frankfurt's view. Frankfurt characterizes respect as follows:

> Treating a person with respect means, in the sense that is pertinent here, dealing with him exclusively on the basis of those aspects of his character or circumstances that are actually relevant to the issue at hand. Treating people with respect precludes assigning them special advantages or disadvantages except on the basis of considerations that differentiate relevantly among them. Thus, it entails impartiality and the avoidance of arbitrariness.[19]

On the basis of this characterization, Frankfurt argues that "it is the moral importance of respect and hence of impartiality … that constrains us to treat people the same when we know nothing that provides us with a special reason for treating them differently".[20] If we add to this the assumption that it is good to treat people with respect in Frankfurt's sense, we seem to be able to provide a value-based explanation of the obligation (and hence the reason) to distribute equally.

However, unless we presuppose that people have a claim to an equal share under the conditions specified by the equal distribution principle, it is not clear why impartiality should require equal distribution. A distributor of goods might also treat

[17] There may be a sense of 'just' that does not involve the notion of a satisfied claim, but merely the idea that underserved inequality is absent (as in "it's unjust that place A has so much better weather than place B" or "it's unjust that player A was so lucky with cards, while player B was so unlucky"). But the view that justice in this latter sense is non-derivatively valuable is vulnerable to the Levelling Down Objection, and can thus be put aside.

[18] One might hold that the claim or obligation *provides* a reason for compliance rather than being constituted by one. But this view faces the challenge to explain the sense in which claims or obligation are intrinsically normative – after all, many things that provide reasons are not intrinsically normative. Moreover, even if claims did not constitutively entail reasons, the present argument would still show that *claims* to an equal share cannot be explained by the value of compliance, and this is a conclusion that proponents of VBT are unlikely to accept (cf. Kiesewetter 2022, 36–37, for analogous points).

[19] Frankfurt (1997, 8–9).

[20] Frankfurt (1997, 10).

two potential beneficiaries impartially by letting a fair coin decide who will get a greater and who will get a smaller share. Intuitively, however, this is not a permissible distribution procedure in the circumstances specified by the equal distribution principle unless the affected parties have consented to it. And yet it cannot be disrespectful in the Frankfurtian sense of assigning advantages or disadvantages partially or arbitrarily. If it is disrespectful to distribute unequally on the basis of a fair coin toss, then this must be because the two parties have a claim to an equal share and the distributor lacks the authority to let the coin decide who will get a greater share unless these claims are waived. I conclude that respect in Frankfurt's sense cannot ground the reason mentioned in the equal distribution principle, as this is a reason that persists even if the requirements of impartiality and non-arbitrariness are consistent with unequal distribution.

As in the case of the arguments against VBT that were based on the promising and the authority principle, the argument based on the equal distribution principle is *explanatory* rather than *extensional*. It is not that there is no value in complying with the reasons that these principles postulate, but that the only value that is necessarily promoted or instantiated by complying with these reasons does itself presuppose these reasons and thus cannot explain them.

4. Conclusion

Reasons based in considerations of egalitarian justice plausibly belong to a group of reasons that cannot be explained by the value that complying with them instantiates or promotes. Is there anything they have in common with other reasons that defy an explanation in terms of value, such as choice-based reasons? It is plausible (though not uncontroversial) to think that promissory obligations and obligations to obey correlate with moral claim rights. Likewise, reasons of egalitarian justice seem to correlate with rights to an equal share. Moreover, all these rights are waivable: Just as a promisee can release a promisor from his promissory obligations and an authority can relase a subordinate from his obligation to obey, people who have a claim to an equal share can also release those in charge of distribution from their obligation to give them their equal share. My hypothesis is that reasons correlating with waivable rights all defy an explanation in terms of the value of compliance.[21]

Why would this be so? It is natural to think that the point of a waivable right quite generally is not primarily to protect a value that is instantiated or promoted by compliance, but rather to protect the value that consists in *having normative control* over whether compliance is required.[22] It is good to be able to create promissory

[21] See Kiesewetter (2022, 47). A further plausible case are reasons to refrain from using other people's property without permission (*ibid.*, 44–46).

[22] On normative interests, see Owens (2012, esp. 6–12).

bonds even if a specific promised action has no value (i.e., no value independently of the value of discharging promissory obligations). It can be good that an authority is in charge even if some commanded action is (independently of the command) pointless. And there can be value in having a claim to an equal share even if one has, in the particular case, no use for the share in question. Perhaps in order for the normative control involved in a waivable right to have value, it must be that compliance with the right is often or typically valuable as well. But as we have seen, it need not be. And yet does a moral right entail a moral obligation, and thus a moral reason for compliance.[23]

As indicated in the introduction, those who find the conclusion of this paper convincing are left with the task of explaining the Guise of the Good Thesis in a manner that is compatible with the rejection of VBT, or rejecting it in a way that explains why it has seemed attractive to so many. Though I can only very briefly sketch a response to this challenge here, one option that strikes me as promising is to reject the Guise of The Good Thesis and claim that its attractions can be accounted for in terms of a Guise of Reasons Thesis instead.[24] It is worth noting, however, that there also is a way to preserve the Guise of the Good Thesis for those who reject VBT. For recall that the argument against VBT was an explanatory rather than extensional one. Rejecting VBT on the basis of this argument is thus consistent with maintaining that compliance with practical reasons is necessarily valuable. This would mean that agents can take their actions to be good even if they act for reasons that are not value-based, without thereby committing any kind of mistake. And this suggests that a vindicating explanation of the Guise of the Good Thesis is available even to the opponents of VBT.[25]

[23] Owens also argues that there are obligations to avoid what he calls "bare wrongings", i.e. wrongings that do not breach anyone's interest (cf. Owens 2012, 12–17). Being in the grip of VBT, however, he holds that these are obligations we need not have any reason to perform. His assumptions commit him to rejecting a weak form of moral rationalism, according to which moral obligations entail at least *pro tanto* reasons for action. I agree with Portmore (2011, 38–51) that such a view cannot accommodate plausible connections between moral obligation and blameworthiness.

[24] See Gregory (2013).

[25] I would like to thank Felix Koch and Thomas Schmidt for very helpful feedback on an earlier draft. Work on this essay was funded by the European Union (ERC Grant 101040439, REASONS F1RST). Views and opinions expressed are however those of the author only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

References

Anscombe, G.E.M. 1957. *Intention*. Cambridge, MA: Harvard University Press (2nd ed., repr. 2000).

Crisp, Roger. 2006. *Reasons and the Good*. Oxford: Oxford University Press.

Foley, Richard. 1987. *The Theory of Epistemic Rationality*. Cambridge, MA: Harvard University Press.

Frankfurt, Harry. 1997. 'Equality and Respect'. *Social Research* 64 (1): 3–15.

Goldman, Alvin I. 2001. 'The Unity of the Epistemic Virtues'. In *Pathways to Knowledge*, 51–70. Oxford: Oxford University Press (2002).

Gregory, Alex. 2013. 'The Guise of Reasons'. *American Philosophical Quarterly* 50 (1): 63–72.

Kiesewetter, Benjamin. 2022. 'Are All Practical Reasons Based on Value?' *Oxford Studies in Metaethics* 17: 27–53.

———. forthcoming. '*Pro Tanto* Rights and the Duty to Save the Greater Number'. *Oxford Studies in Normative Ethics* 13.

Kiesewetter, Benjamin, and Jan Gertken. 2021. 'How Do Reasons Transmit to Non-Necessary Means?' *Australasian Journal of Philosophy* 99 (2): 271–85. https://doi.org/10.1080/00048402.2020.1745252.

Maguire, Barry. 2016. 'The Value-Based Theory of Reasons'. *Ergo* 3 (9): 233–62.

Owens, David. 2012. *Shaping the Normative Landscape*. Oxford: Oxford University Press.

Parfit, Derek. 1997. 'Equality and Priority'. *Ratio* 10 (3): 202–21.

Portmore, Douglas W. 2011. *Commonsense Consequentialism: Wherein Morality Meets Rationality*. New York: Oxford University Press.

Rabinowicz, Wlodek, and Toni Rønnow-Rasmussen. 2004. 'The Strike of the Demon: On Fitting Pro-Attitudes and Value'. *Ethics* 114 (3): 391–423.

Raz, Joseph. 1986. *The Morality of Freedom*. Oxford: Oxford University Press (repr. 1988).

———. 2001. *Value, Respect, and Attachment*. Cambridge: Cambridge University Press.

Rønnow-Rasmussen, Toni. 2002. 'Instrumental Values – Strong and Weak'. *Ethical Theory and Moral Practice* 5 (1): 23–43. https://doi.org/10.1023/A:1014422001048.

———. 2011. *Personal Value*. Oxford: Oxford University Press.

———. 2022. *The Value Gap*. Oxford: Oxford University Press.

Schroeder, Mark. 2007. *Slaves of the Passions*. Oxford: Oxford University Press.

Temkin, Larry S. 1993. *Inequality*. Oxford: Oxford University Press.

Wedgwood, Ralph. 2009. 'Intrinsic Values and Reasons for Action'. *Philosophical Issues* 19: 321–42.

———. 2017. *The Value of Rationality*. Oxford: Oxford University Press.

Collective Obligations and the Moral Hi-Lo Game

Kirk Ludwig

Introduction

Olle Blomberg and Björn Petersson (2023) argue that collective moral obligations, at least in some cases, are irreducibly collective. By this they mean the subject of the obligation is a group and their having a moral obligation collectively cannot be analyzed into individual obligations of its members to do their parts in what the group has an obligation to do. The main argument focuses on a choice situation that looks like a moral Hi-Lo game, in which we have the intuition that the group is responsible for pursuing the best moral outcome. Blomberg and Petersson argue that we cannot account for this intuition by deriving it from individual obligations of the parties to do their parts in bringing about the best moral outcome. In contrast, I will argue that the case has not been made and that we can plausibly account for the intuition that the group has a moral obligation while seeing it as grounded in the independently derived obligations of the members to do their parts.

We typically attribute obligations to informal groups with plural referring terms as in [1].

> [1] We ought to save the children

[1] is ambiguous between a distributive and a collective reading. On the distributive reading, [1] is understood as equivalent to [1d].

> [1d] Each *x* of us: *x* ought to save the children.

On the collective reading, in contrast, the group is, in some sense, the locus of the obligation, as in [1c].

> [1c] We are such that we ought to save the children (working together).

Just as an individual obligation requires of its subject, *ceteris paribus*, action to fulfill the individual obligation, so group obligation requires of its "subject", *ceteris paribus*, collective action to fulfill the group obligation. I put "subject" in scare quotes in the second clause because at this point we do not want to assume that [1c] involves commitment to the group per se being the bearer of the obligation rather than it distributing obligations to its members *to contribute to their jointly saving the children*. We make attributions of collective obligation when

> (a) group action is necessary in order to bring about a moral good or avoid a moral harm or
>
> (b) group action, even if not necessary, nonetheless will be more effective or carry less risk of failure in pursuing a moral good or in avoiding of a moral harm.

One can take two different stances on collective obligations. First, one can regard the collective obligations of groups as grounded in their individual obligations. This is the bottom up approach and entails that collective obligations are reducible to individual obligations to contribute to collective action. Second, one can regard collective obligations of groups as primary and any individual obligations, when present, as derived from them. This leaves open that collective obligations may not always entail or be accompanied by individual obligations. This is a top down approach. I endorse the bottom up approach.

The Case Against Reduction

Blomberg and Petersson make a case for the top down approach by appeal to cases. Here is the central example (2023, 1; page number citations are to the online first version of the paper).

> *Burning Building*: Three children are trapped in a burning building. One of them is in one room, and the other two are in a second room some distance away. The neighbours Agnetha and Benny see each other approaching the building from opposite sides. Agnetha breaks in and has enough time to do her part of rescuing either the child in the first room or the two children in the second room. The rescue can succeed only if Benny heads straight for the same room with his fire extinguisher. If both go to the first room, they will only rescue the first child. If both go to the second room, they will only rescue the two other children. If each goes to a different room, no child will be rescued. Suppose that Agnetha and Benny can make these choices without any significant risk to their own or each other's life or health. All this is common knowledge between them, but they do not have any opportunity to communicate with each other—each must choose which room to head for independently of the other. (Adapted from (Colman, Pulford, and Lawrence 2014, 36))

Blomberg and Petersson share the intuition that Agnetha and Benny have an obligation jointly to save the two children. They call the basic intuition, following

Schwenkenbecher (2019, 28). The reason Agnetha and Benny have a collective moral obligation is that: "Only together do they have the ability to rescue the two children, and rescuing the two children is the best they can do, morally speaking" (2023, 2). Blomberg and Petersson aim to vindicate the basic intuition in the face of a puzzle about how it is possible. The puzzle is that it seems that for the group of Agnetha and Benny to have an obligation, they must be together an agent. But the only agents present are Agnetha and Benny, who can't communicate and must act independently. Blomberg and Petersson argue that it suffices for them to have a collective obligation that they are each able to ask not just "What ought I to do?" but "What ought we to do?" More specifically, what is necessary for this is that Agnetha and Benny are able to (i) identify with the group, (ii) to "we-frame" the situation, and (iii) deliberate about what the group ought to do. They develop this account drawing on Michael Bacharach's development of the team reasoning framework in decision theory (Bacharach 2006). The argument that collective obligations are not analyzable into individual obligations of its members rests on the claim that the latter are answers to the question "What ought I to do?" and that that starting point is insufficient to recover the basic intuition in cases like Burning Building.

Blomberg and Petersson note that Stephanie Collins (2019, 140) has argued that the individual agents can reason from individual obligations to participating in group action. Blomberg and Petersson claim that this cannot vindicate the basic intuition.

> We will now explain why Agnetha and Benny would not be able to have a collective deliberative obligation if they could only ask and answer the question: "What ought I do?" We claim that if Agnetha starts by asking what she ought to do in Burning Building, then she cannot rationally settle on any determinate answer. The same is true of Benny. They could therefore not have any collective deliberative obligation to save the two children. But why is this so? (2023, 5)

The core of the argument is that Agnetha and Benny face a moral Hi-Lo game. In a Hi-Lo game, there are two Nash equilibria (if the other(s) maintain their choices, no one has an incentive to change theirs), but one of the two has a higher payoff for both. In Burning Building, the structure is represented in Figure 1 (the first number in each box represents Benny's payout, the second Agnetha's).

Moral Hi-Lo Game		Agnetha	
		Room 1	Room 2
Benny	Room 1	1, 1	0, 0
	Room 2	0, 0	2, 2

Figure 1: Hi-Lo moral dilemma. Agnetha and Benny can save 1 or 2 children if both go to room 1 or room 2, but none otherwise.

The Hi-Lo game is a problem for classical game theory, which treats agents as individual strategic reasoners who make choices in the light of the choices that others make or are likely to make. In a Hi-Lo game, if you do not have any evidence bearing on what choice the other will make, it seems you can only engage in conditional best response reasoning. Room 1 is best for Benny if Agnetha goes to room 1, but room 2 is best for Benny if Agnetha goes to room 2. *Mutatis mutandis* for Agnetha. Thus, in the moral Hi-Lo game, if Benny asks "What ought I to do?", it seems that his answer will depend on how Agnetha answers the question "What ought I to do?" For each should aim for the morally best outcome, it seems, *given what the other does*. But then they seem to be at an impasse, and neither can reach a rational decision about what he or she ought to do.[1]

The argument that an individualist account of collective obligation cannot accommodate the basic intuition can be put as follows.

1. Agnetha and Benny have a collective deliberative obligation to save the two children in room 2 only if Agnetha and Benny first have individual obligations to contribute to their saving the two children in room 2.
2. Agnetha and Benny have individual obligations to contribute to their saving the two children in room 2 only if each can reason correctly that the answer to the question 'What should I do?' in Burning Building is to go to room 2.
3. Neither Agnetha nor Benny is in a position to reason correctly that the answer to the question 'What should I do?" in Burning Building is to go to room 2.
4. Therefore (from 1-3), Agnetha and Benny do not have a collective deliberative obligation to save the two children in room 2.

Premises 1 and 2 express the individualistic approach to collective moral obligation. Premise 3 is supported by the claim that they cannot move past conditional best response reasoning in Burning Building if they are focused on the question "What should I do?" The conclusion 4 contradicts the basic intuition. To hold onto the basic intuition, then, we must give up the conjunction of 1 and 2, that is, the individualistic approach to collective moral obligation. I will call this the moral Hi-Lo argument. If this is correct, then the basic intuition cannot be accounted for by the bottom up approach, which takes individual obligations to be explanatorily basic.

In cases of decision making under uncertainty, the indifference principle suggests giving each option equal weight. Assume this is a rational strategy. (Surely it is better than doing nothing when doing nothing is guaranteed to have no payoff and there is no cost to action.) In the Hi-Lo game, this gives choosing Hi (room 2 in our case) a higher expected utility ($.5 \times 2 > .5 \times 1$). This would provide a way of moving past conditional best response reasoning which leaves Agnetha and Benny at an impasse and thereby show that premise 3 is false.

[1] For development of this idea see (Sugden 2000, 179-182, Bacharach 2006, 35-68).

However, Blomberg and Petersson argue that we should doubt that the indifference principle is a sound general principle for reasoning in the absence of information about probabilities, on the grounds that its application to some Stag Hunt games leads to the wrong result. In a Stag Hunt game, to take a simple case, two hunters have the option of hunting rabbits or stags. There's one stag in the hunting range and a number of rabbits. One stag yields as much nutrition as six rabbits. But a successful stag hunt requires the hunters to cooperate. We suppose they can't communicate and must make a choice of proper hunting tools before going on the hunt. If one chooses stag and the other rabbit, the one who chooses rabbit will be able to trap two rabbits in the time available but the other can hunt neither stags nor rabbits. If both choose rabbit, there are enough rabbits in the range for each to trap two. This game structure is shown in Figure 2.

Stag Hunt		Player 1	
		Stag	Rabbit
Player 1	Stag	3, 3	0, 2
	Rabbit	2, 0	2, 2

Figure 2: Stag Hunt. There are two Nash equilibria, Stag-Stag and Rabbit-Rabbit.

In this game, the indifference principle doesn't help to choose the payoff dominant Nash equilibrium because the expected payout is higher for each in choosing Rabbit (.5 x 2 + .5 x 2 > .5 x 3 + .5 x 0, that is, 2 > 1.5). Here Blomberg and Petersson assume that it is, contrary to the result of applying the indifference principle, not irrational[2] for the hunters to play (stag, stag) rather than (rabbit, rabbit).[3]

[2] While I will not appeal to the indifference principle below in defending a reductive account of group obligation, it is worthwhile asking whether it really delivers the wrong result in this case. Why is it supposed to be rational for them to hunt stag rather than rabbit? The intuitive idea is that they'd be better off, so surely if they both choose stag, they cannot be charged with being irrational. But doesn't this depend on what they think about the other? If player 1 thinks player 2 is risk averse, then it would not be rational for player 1 to choose stag. If player 2 thinks player 1 thinks player 2 is risk averse, it would not be rational for player 2 to choose stag. So intuitions about whether it is rational for them to choose stag depend on how we fill in the picture about what they think about the other. If each thinks the other is likely to choose stag, then it is certainly rational for each to do so. But what about the case in which they have no idea what the other thinks or is disposed to do? Is it rational to assume that the other is more likely to choose stag than rabbit or vice versa? Neither, evidently. It is in this circumstance that the indifference principle looks like a reasonable basis for a decision. Blomberg and Petersson think that this gives the wrong verdict, and that it is in fact rational for the players to choose (stag, stag). But why? They think group identification and team reasoning can make sense of it, for if both team reason, they are better off. But it is reasonable for each to team reason only if they each have good reason to think the other is doing so as well. But in this case they do not.

[3] These are not the only options. Perhaps it is a reasonable principle, for example, to choose the Pareto optimal equilibrium in cases in which there is more than one Nash equilibrium and they are strictly ordered by Pareto dominance. An outcome *x* Pareto dominates outcome *y* iff it *x* is strictly better for at

Blomberg and Petersson also argue that the indifference principle and the common knowledge of rationality requirement would lead Agnetha and Benny to have contradictory beliefs. If each knows the other applies the principle, each can then reason that the probability of the other going to room 2 is higher than 50%. But this, they say, leads Agnetha and Benny to have contradictory beliefs: each thinks the probability that each will go to room 2 is 50% and also higher than 50%.

A second objection to a reductive account of collective obligation focuses on whether starting from the question "What should I do?" Agnetha and Benny could have normative reasons to save the two children.

> Plausibly, the subject of an obligation to Φ must not only have the ability to Φ, but also the ability to Φ for the normative reasons that make Φ-ing morally obligatory (see (Lord 2015); cf. Collins 2019: ch. 3). Agnetha and Benny would lack this ability if each of them were restricted to asking and answering the question: "What ought I do?" In its first-person singular form, the deliberative question concerns what to do solely on the basis of the person's own agentive abilities. Hence, if Agnetha were limited to "I-reasoning," then she would not have a normative reason to do her part in saving the two children. Nor would Benny if he were limited in the same way. Only together can they have a normative reason to save the two children (cf. (Dietz 2016, 960-963)); and only if they together have this normative reason will each of them have a normative reason to do their part. … they can therefore at most each have an obligation to do their part in a collective endeavour such as a joint rescue. (2023, 8-9)

I take the central assumption here to be

> *I-Reasoning*[4]: Reasoning in the first person always concerns only what the agent can do by herself with no contributions from others.

A secondary assumption is a cousin of the Ought Implies Can principle:

> *Having a Normative Reason to* $\boldsymbol{\varphi}$ *Implies that You Can* $\boldsymbol{\varphi}$ (*NORIC*):
> If one cannot do something, then one cannot have a normative reason to do it.

Then, since neither Benny nor Agnetha can alone save any children, I-Reasoning would preclude them from seeing anything they can do that would save any children, and, hence, by NORIC, they cannot have a normative reason to save any children. Hence, they cannot have an obligation to do so, since they can have an obligation to

---

least one player than *y* and no worse for any. An equilibrium is Pareto optimal iff no change can make anyone better off without making someone worse off. This principle would select (Hi, Hi) in the Hi-Lo game. See (Harsanyi and Selton 1988).

[4] Blomberg (personal communication) states that he would not endorse this principle in general but maintains it is true about I-reasoning in Hi-Lo. Substituting 'in Hi-Lo' for 'always' will not affect the critical remarks below.

do so only if they can act on a normative reason to do it. I'll call this the argument from I-reasoning.

Blomberg and Petersson argue further that the key to Benny and Agnetha having a collective obligation is their having the capacity to we-reason, to adopt the team reasoning perspective in which each thinks in terms of maximizing the (moral payoff) for group action, not for individual action, and then derives from that what they ought to do as part of the team. This makes obligation relative to agential perspective. They argue that in some situations the "I"-relative oughts and "we"-relative oughts require or permit different responses, and this yields a kind of moral incommensurability. The capacity to we-reason is itself context relative because the capacity for group identification, which is the key to we-framing their decisions, will be affected by the context.

These further conclusions rest on the idea that the basic intuition that Benny and Agnetha have a collective obligation in Burning Building cannot be accommodated by starting with a standard picture of obligations that individuals have. In the following, I will argue that the arguments against reduction are not successful.

Reply to the Hi-Lo Argument

The basic intuition is that in such cases as Burning Building agents can be morally obligated to do something as a group, something we express by saying that *they* have an obligation to do something together. In particular, Benny and Agnetha have an obligation to save the two children by their going to room 2 because that is the morally best outcome. This cannot be taken as a judgment that they have an irreducibly collective obligation without begging the question. Our question is whether the basic intuition can be vindicated without accepting that the attribution of collective responsibility is irreducible.

It will be helpful to begin with a reductive account of collective obligation. We may put it this way (for pro tanto obligation add 'pro tanto' before 'collective obligation' on the left hand side of the biconditional and before 'moral obligation' in (b)):

> *Collective Obligation as Distributed Obligation to Engage (CODE):*
> A group *G* has a collective obligation to *J* in context C if and only if in C:
>
> (a) *G J*-ing is necessary in order to bring about a moral good or avoid a moral harm
> or
> *G J*-ing will be more effective or carries less risk of failure in bringing about a moral good or in avoiding of a moral harm, and
>
> (b) each member *x* of *G* individually has a moral obligation to contribute to *G J*-ing which is not derived from *G* having independently a collective obligation to *J*.

The Hi-Lo Argument aims to show that our intuition in Burning Building cannot be vindicated by appeal to the CODE conception of collective responsibility because, the argument goes, condition (b) is not met in that case.

The moral Hi-Lo game is supposed to present a problem for Benny and Agnetha if they are thinking in terms of their individual moral responsibilities. The problem is supposed to be that what each should morally do, if asking 'What should I do?', depends on what the other in fact does. Each is able, the claim is, only to reason, "If the other goes to room 2, I ought to go to room 2; if the other goes to room 1, I ought to go to room 1."

The scenario stipulates that they can't communicate and that the basic setup is common knowledge between them. But in many other ways the case is under described. Therefore, the intuition that they have a collective obligation to save the two children may rely on how we fill in the case. This will be relevant also to what resources they have for reasoning in the first person about what they should do. In the case of Burning Building, how do we fill in our understanding of what Benny and Agnetha are likely to know or believe about each other? We likely think that if they could communicate, they would confirm with each other their understanding of the situation and what it is best to do—which is obvious in this case—and then do it. That is why we stipulate that they can't communicate.[5] But that already shows that we are thinking that both of them have pro moral attitudes and expect that the other does as well. If this is our default view of their characters and attitudes, then we will also suppose that each of them assigns a relatively high likelihood to the other *being motivated to achieve the best moral outcome in the situation*. Let us say that each assigns a probability of .8 to the other aiming for the best moral outcome (the exact number turns out not to be material). We can assume that each knows or has a subjective probability close to 1 (let's say 1) that he or she wants the best moral outcome. Then each will assign a probability of .8 that they both want the best moral outcome. They know that the best outcome is saving the two children in room 2 and that working together is necessary to bring it about. So each will know, or have good reason to believe: each of us wants to cooperate on saving the two children in room 2 because that is the best moral outcome. Neither will think: we want to cooperate on saving the one child in room 1. The result of their communicating would be to confirm this. But communication is not necessary for them to have good reason to believe it. They will thus be able to identify in the circumstances a goal that each knows that each wants to promote over other things that they can do and that it is what they would agree to do if they could communicate. Thus, each has a *reasonable expectation* that his or her going to room 2 will promote what is morally the best outcome in the circumstances. (The reasoning here does not start by adopting the perspective of the group and so is not team reasoning; rather it draws on background

[5] Of course, one of them, Agnetha, say, might simply announce that she is going to room 2, counting on Benny to follow suit since it settles what he should do. But that shows also that she already thinks he wants what is morally best as she does.

information about the circumstances to increase confidence in both that the other will go to room 2.) Each has an individual obligation to contribute to the best moral outcome they can achieve in the circumstances. Therefore, each has an obligation to do his or her part in their rescuing 2 children in room 2. Thus, they have a collective obligation to rescue 2 children in room 2. This bit of reasoning vindicates the basic intuition, but it does so by deriving individual obligations from their expectations about the other sharing the goal of achieving the best moral outcome, which is a natural way to fill in the background in Burning Building, and so via satisfying (b) in CODE. This shows that premise 3 in the argument above from a reductive account of collective obligation to Agnetha and Benny not having a collective obligation in Burning Building is false.

We standardly expect, when we have common interests with others which we can only achieve by working together and there are no costs in cooperating, that the other will cooperate with us, that is, that we will cooperate. Imagine a Burning Building scenario where children are in only one room. There is still a question about what one should do when it requires two to rescue them, but you cannot communicate. Your expectation is that the other will contribute to bringing about the best moral outcome and you act on that assumption. Burning Building differs from this case in having not just two possible outcomes (save the children or don't) ranked in terms of their moral desirability but three (save two, save one, save none). Yet on the assumption that you both will act from moral motives, it is obvious that the morally right thing to do is to rescue as many children as you can.

The methodological point is that our judgements about a group being collectively responsible in a scenario is very likely to be conditioned by our making default assumptions, which reflect a natural way of filling in the background, when nothing contrary is explicitly stated in the description of the scenario (and sometimes even when it is). The natural background for Burning Building (and one implied by how it is framed in terms of a matrix that represents only moral goods and the exclusion of communication) is that each of Agnetha and Benny want to do the right thing and believe this about each other. If the intuition rests on this way of filling in the background, then the basic intuition can be recovered from the individual perspective without difficulty.

Blomberg and Petersson might level the objection that whatever assumptions Agnetha and Benny make about each other's tendency to go to room 2 will lead to their having contradictory beliefs about probabilities when combined with the assumption of common knowledge of rationality and common knowledge of the circumstances. The argument has the same form as in the case of the appeal to the principle of indifference. If each knows the other has reason to assign a probability p to the other's going to room 2 and that that yields a higher expected payoff than going to room 1, then the probability that the other will go to room 2 should be higher than p. In fact, this style of objection, if successful, would work no matter what the probability (other than 1 or 0) they each assign to the other's doing something, as long as it selected one of the options as superior to the other. The

problem, if there is one, is also independent of the Hi-Lo game. It arises even if there is just one room with children in a burning building. Yet, surely in this case they need not have contradictory beliefs.

There are a number of threads to disentangle here.

(1) The reasoning I sketched above did not in the first instance assign a probability to the other going to room 2 but instead to the other aiming for the morally best outcome. If they each have reason to assign .8 to the other aiming at the best outcome, and this is common knowledge, then they can each reason that there is a .8 probability that they both aim at the best moral outcome. As they can each see that the other will reason in this way, it will be common knowledge among them that there is a .8 probability that they both aim at the best moral outcome. On the assumption they both aim at the best moral outcome, they should each go to room 2. It is more reasonable for each to adopt that assumption than any other. Given that they are rational, they will adopt the more reasonable assumption. Assume this is all common knowledge. Then each knows the other assumes they both aim at the best moral outcome and so knows that each believes that he/she should go to room 2. If they each believe that he or she should go to room 2, then each of them will. Assume each believes this. Then each of them believes that each of them will go to room 2 and each of them will. I have not here represented the reasoning in the last stages as involving a calculation of an expected value for going to room 2, but as a matter of adopting an assumption for the purposes of determining what to do. Subsequent reasoning takes it as a fixed point.

(2) Assume, however, that at some point each assigns a probability to the other going to room 2 that yields going to room 2 as having the higher moral expected value. If the reasoning proceeds in this way, each has sufficient normative reason to go to room 2. Further reasoning is superfluous. What matters for action is not the size of the difference in expected value between going to room 2 and other options, but whether it is higher. If it is, then each can just act without further reflection, and they would not be involved in any incoherence, if there is any threat of it, from further reflection.

(3) However, even if they do engage in further reflection, it should not result in their holding inconsistent beliefs. Either the reasoning that is supposed to generate a greater confidence level takes some time or it takes no time. If it takes some time, then inference to a higher probability than initially assigned will require a simple update of the probability and the agent does not need to hold contradictory beliefs.[6] If it takes no time, it is not an inference at all. All we could have in mind is that the agent's beliefs are from the beginning where they should be on working out the consequences of all their basic beliefs and assumptions. But this would not involve any contradictory beliefs.

[6] Blomberg and Petersson suggest that the reasoning would amount to assuming the probability that the other will choose one option is *p* and then inferring from that that the probability is different from *p*, which might be taken to be a reductio of the original assumption. But this need not be how it

One might grant that the intuition *in Burning Building* can be generated by thinking of each of Benny and Agnetha having individual obligations to do their parts in saving two children based on calculation of the expected value of going to room 2 versus room 1 given the background assumptions sketched above, but deny that this generalizes on the grounds that we can always find a *Moral Stag Hunt* which can't be solved no matter how confident (short of certainty) the hunters are of the other(s) hunting stag. (I am not attributing this response to Blomberg and Petersson, but it will be instructive to consider it.)

First, let's assume that the reasoning of Agnetha and Benny must proceed by assigning a probability of .8 that the other will choose the morally best option. Then it seems that there is a set of payoffs in the Stag Hunt game that will make it less rational morally to hunt stag than rabbit even if one knows that the others are .8 likely to hunt stag. For example, if you know that the payoff for hunting stag if the others do is 10 and 0 otherwise, but 9 if you hunt rabbits, then the expected value for hunting stag is .8 x 10 = 8 but for hunting rabbit, which is a sure thing, it is 9. See Figure 3.

Moral Stag Hunt		Player 1	
		Stag	Rabbit
Player 1	Stag	10, 10	0, 9
	Rabbit	9, 0	9, 9

Figure 3: Moral Stag Hunt. Even with a .8 probability that the other will choose Stag, hunting Rabbit appears better for each.

In general, for any probability assigned to the others' hunting stag, we can find a ratio of payoffs for hunting stag versus rabbit in the stag hunt scenario that will have the result that hunting rabbit is better than hunting stag. Let s be the utility of hunting stag when all parties do, r be the utility of hunting rabbit alone, and p be the probability that the others will hunt stag. If $r > s \times p$, then the expected value for hunting rabbit will be greater than hunting stag. For any $p < 1$, for any s, there is an r such that $r < s$ and $r > s \times p$. Thus, the argument goes, no matter the probability assigned to the other hunting stag, we can find a payoff structure in which the

proceeds. One can reason as follows. The antecedent probability, without taking into account how the other will reason, that the other will go to room 2 is .5. If it is .5, then going to room 2 is best. Therefore, I should go to room 2. But the other will reason in the same way that I just did. If the other reasons in the same way I just did, then the other will reach the conclusion that going to room 2 is what he should do. Given that the other is rational, that is what he will do. Therefore, taking into account additionally how the other will reason, the probability that the other will go to room 2 is 1. This is a revision of the assignment on the basis of taking into account additional evidence relevant to the assignment of the probability, namely, how the other will reason. There is nothing incoherent in assigning a probability of .5 to **α** given background knowledge K and assigning a probability of 1 to **α** given K + **δ**.

morally best outcome is not chosen from the standpoint of individual moral reasoning.

But *not so fast!* There are two points we need to attend to. The first is what we mean by saying the utilities are interpreted as moral outcomes—we need a story that makes sense of this. The second is how we are to think of the utilities for the two players given that they represent moral outcomes.

On the first point, it is hardly clear that the group is morally required to maximize the amount of meat summed across all hunters or to maximize the amount each takes home from the hunt. To construct a case where it seems that there is a clear forward looking moral obligation to work together, let us suppose that the purpose of hunting is to procure food to save children from dying of starvation. Let's suppose that each unit of value in the matrix represents enough meat to save one child from starvation. Then if Agnetha and Benny both hunt stag, 20 children are saved. If they each hunt rabbit, 18 children are saved. If either hunts stag and the other rabbit, then nine children are saved.

On the second point, the good for each contributor is the total number saved, just as in Burning Building, for they aim for the moral good, and so all should share equally in the good that results from the intersection of their choices. Given this, for each the value of the intersection of any pair of actions is the total number of children saved. This alters the payouts as shown in Figure 4.

Moral Stag Hunt Revised		Player 1	
		Stag	Rabbit
Player 1	Stag	20, 20	9, 9
	Rabbit	9, 9	18, 18

Figure 4: Moral Stag Hunt revised. 20 children are saved if both hunt stag, 18 if both hunt rabbit, and 9 if one hunts rabbit and the other stag.

But now the altered payoff structure gives us a different result. In effect, this transforms it into a Hi-Lo game.[7] Give the payoffs, as long as the probability that

[7] Blomberg and Petersson agree with this conclusion (see p. 30), which is reached on the basis of the assumption that moral payoffs are agent-neutral in the sense that it is irrelevant who brings them about. They suggest, however, that we can generate a genuine moral Stag Hunt scenario if morality is, at least in some cases, agent-relative (see p. 31) in the sense that one can assign a higher moral value to saving someone oneself rather than another doing it. The common-sense argument for agent-relativity is that we can assign a higher moral value to helping someone to whom we have a special relation, such as a child or spouse, than to someone else doing so. Blomberg and Petersson do not take a stand on whether morality can be agent-relative. They aim rather to illustrate how the I-perspective and team-perspective may ground different moral judgments about what one ought to do, if morality is agent-relative. However, even if we grant morality can be, in some cases, agent-relative, in the sense that an agent may place greater moral value on saving someone herself than another doing so (saving your child for example), it is not clear that this avoids the collapse into a Hi-Lo game. For if it is better for a mother to save her child than a stranger because of her special relation to her child, then plausibly that is better

the other hunts stag is greater than 9/20 (.45) it will be best to hunt stag. Thus, for example, simply applying the principle of indifference would weigh the options equally, and then the payout for choosing to hunt stag is 20 x .5 + 9 x .5 = 14.5 which is greater than the payout, 9 x .5 + 18 x .5 = 13.5, for choosing to hunt rabbit. If we can assume, as we did above, the likelihood is higher than .5, the gap in the expected values will be greater. So the result is not that *individually* they morally ought each to hunt rabbit when they are asking 'What ought I to do?' but instead that they should each hunt stag with the other. The central point is that if we think of the payoffs as representing the moral good, then positive outcomes are the same for everyone (see note 8 on whether agent-relative morality makes a difference). When *N* children are saved, that is not less morally good when I don't save them. Thus, we don't get the standard structure of the stag hunt when we think about the outcomes as in terms of what is morally best. I will say more about the significance of this below.

I'll return to some other ways of filling in the scenario in Burning Building where Agnetha and Benny may have more reason to think the other *will not* help at all or not go to room 2. We will see that this does not help the argument. But before that it will be useful to turn to the argument from I-reasoning. For what I have just described may sound as if it is already we-reasoning, and that if Benny and Agnetha were really restricted to thinking about their individual obligations and reasoning about what they could do, they could never arrive at the conclusion that they were each obligated to contribute to their saving two children.

Reply to the Argument from I-reasoning

The two assumptions of this argument were I-reasoning and NORIC, repeated here.

> *I-Reasoning:* Reasoning in the first person always concerns only what the agent can do by herself with no contributions from others.
>
> *Having a Normative Reason to* φ *Implies that You Can* φ *(NORIC*):
> If one cannot do something, then one cannot have a normative reason to do it.

---

from everyone's perspective, not just the mother's. But then that extra value will simply be added to the sum total good brought about by the combination of actions, leading to a transformation of the matrix into a Hi-Lo game (preserving symmetry for the players). That is, everyone sees the value added when someone helps those they have a special relation to. Agent relativity would have to be interpreted as meaning not just that an agent may assign a greater value to helping someone she has a special relation to, but also that that fact is morally irrelevant to everyone else. But *that* does not seem to be of a piece with common sense morality. We surely do all agree that, other things being equal, it is better for parents to feed their children than strangers, for example, because of their special relation to them.

I think that both of these assumptions are mistaken.

I-Reasoning is false because in thinking about what I can bring about, I do not need to exclude calculations about what others are doing or would do given what I do. Given any goal I have, I can ask myself what I can do to achieve it. This is reasoning in the first person, but it precludes no means to the end. Sometimes I can get what I want by getting others to do things whether they know that is my intention or not. I can clear a building by setting off a fire alarm. Clearing the building requires contributions from others. But what I do gets them to make their contributions. I can also get things done by working with others. If I want to move a large table from one room to another, one answer to the question 'What should I do?' is to get some help. So reasoning in the first person does not always concern only what the agent can do by herself with no contributions from others. So when someone asks, "What ought I to do?" in circumstances like Burning Building, the answer can be that I should work with someone else to achieve the best moral outcome. That is a perfectly fine answer to the question about individual obligation. That gives one a normative reason to contribute to saving the two children.

NORIC is controversial.[8] I believe that it is false. There is not space here to do justice to the issues, but here is a quick brief. We have both positive and negative moral duties. A positive moral duty is a duty to do something, either to bring about a good or prevent a harm, construed broadly. A negative moral duty is a duty not to do something because of the harm it will bring about. These general duties are standing requirements. If I am not in a position to fulfill a duty, it does not disappear. What the lack of ability to fulfill a duty pertains to is whether it is proper to blame me for not fulfilling it. The duty I have to prevent harm gives me a normative reason to save children in a burning building even if I am the only one around. I have that reason even if I cannot act on it because I do not have the ability alone to do it. This is true of all our reasons. A reason to do something is a consideration that speaks in favor of it (it need not speak decisively in favor of it because we can have conflicting reasons). That I cannot do something does not entail that I have no reason to do it, that is, that nothing speaks in favor of it. If I develop a craving for water in the desert, I have a reason to drink some water. If there is none around, I can't act on it. But I am motivated to look for some because I have a reason to drink some water.

[8] Bart Streumer (2007) and (Lord 2015) argue it is true; Ulrike Heuer (2010) and Kimberley Brownlee (2010) argue it is false; Streumer replies to Heuer in his (2010). One could *insist* that a consideration that speaks in favor of something isn't sufficient for there to be a reason in favor of it (here we do not mean all-in or everything considered) but that there must also be some prospect of acting on it successfully. Suppose we call these c-reasons. By definition you don't have c-reasons if you can't do what they are putative reasons for. But now we should reject the claim that you can't have an obligation to do something if you don't have a normative c-reason for it. For sometimes we fail to fulfil duties we indisputably have, such as caring for our children, because we can't do it, for example, because we suffer from a debilitating fatal disease. In this case, we would have the duty but no normative c-reasons to do it. We are not blamable in this case, but it is not because we fail to have any duties to our children.

If I find an oasis, then my reason becomes practically relevant, but it was there all along, as evidenced by its motivating me to look.

Whatever the verdict on NORIC, the argument from I-reasoning fails because in Burning Building Agnetha and Benny are not restricted in first person reasoning to thinking about what they can do alone.

Burning Building Redux

Return to the case of Burning Building. Surely all that is needed is a set of circumstances, even if it is not the default way of filling in the case, in which we judge that Agnetha and Benny are collectively obligated to save the two children, but it is not the case that each can arrive at an obligation to do their parts by answering the question "What ought I do?" Suppose that Agnetha and Benny live in a dystopian future in which there is extreme scarcity and almost everyone has developed a reflexive tendency to mind their own business. Agnetha and Benny know that most people would not make any attempt to save the children even if they could do it alone. They know nothing about each other. They can't get any information about what the other is doing before they actually undertake to arrive in room 1 or room 2. Thus, each (Agnetha/Benny) has more reason than not to think the other (Benny/Agnetha) will not contribute even knowing the other (Agnetha/Benny) wanted to save the children. We'll call this the Dystopian Burning Building case.

Let's grant that they collectively have an obligation to save the children, that is, the answer to the question 'What should they do?' is 'They should save the two children'. They also then collectively have a normative reason to save the children. However, given what I said above about the NORIC principle, we should also grant that they have a normative reason to do their parts in saving the children. Perhaps this then supports also the judgment that they, in some sense, still have an obligation to do their parts in saving the children.

Still, each of them has *most* reason to think that *they can't fulfill that obligation*. So we might focus on the question whether there is a contrast between what they have most normative reason to do assuming they act together and what each has most normative reason to do individually. Perhaps each of them has most reason not to do anything, given the high probability that the other will not do anything. The group perspective might be said to be expressed by the question, 'What do we have most reason to do together?' If the answer to this question is to save the two children, then we apparently get a contrast in judgments about what the group has most normative reason to do and what its members, reasoning from the first person, have the most normative reason to do.

In response, first, although it is clear that each has most reason to think *they cannot fulfill* an obligation to contribute to their saving the children in Dystopian Burning Building, it does not follow that each does not have *most reason to (undertake to) do their part in their saving two children*. Benny believes it is unlikely that Agnetha will make any effort to save any of the children. Still, it is not out of the question for Benny that she will. Since Benny's going into the Burning Building does not involve any risk to himself, it still makes sense for him to go to one of the rooms on the chance that Agnetha will go to one of them. But which one? While from Benny's point of view, it is epistemically unlikely that Agnetha will act to save any of the children, if she does, it is because she is motivated by moral considerations. In that case, she is more likely to go to room 2 because that is morally the best outcome, as we argued above. So Benny should act to go to room 2, even if the chances of success are low. Mutatis mutandis for Agnetha. The decision matrix in this case includes three options: go to room 1, go to room 2, and don't do anything. This matrix is represented in Figure 5.

Moral Hi-Lo Game 2		Agnetha		
		Room 1	Room 2	Nothing
Benny	Room 1	1, 1	0, 0	0, 0
	Room 2	0, 0	2, 2	0, 0
	Nothing	0, 0	0, 0	0, 0

Figure 5: Hi-Lo with the option of doing nothing.

Given that, from the point of view of Benny/Agnetha there is some chance that Agnetha/Benny will go to room 1 or room 2, doing nothing will have a lower expected value that going to one of the rooms. Once the choice is made to set aside doing nothing, this reduces the decision matrix to the original. So in fact Dystopian Burning Building does not introduce any essentially new elements.

When Agnetha Knows that Benny Will Go to Room 1

It might still be argued that we can get a divergence in what obligations Agnetha and Benny have together and individually by considering a case in which it is common knowledge among them that one will go to room 1, no matter what.[9] The

[9] These arguments are not the ones that Blomberg and Petersson offer for individual and collective obligations coming apart, but theirs depend on the idea that team reasoning may generate obligations independently of individual obligations to do their part. This rests in part on the idea that we can't recover the basic intuition about Burning Building, which I have disputed. I'll also argue in the last section that team reasoning has to be justified before we engage in it, and so can't function as an

thought is that since they clearly can together save two children and not just one, and that is strictly morally better, they ought to save the two children. But if, for example, Agnetha knows with certainty that Benny will go to room 1, then her moral obligation is to go to room 1. Surely if Agnetha has an obligation to do her part in their saving the child in room 1, then she does *not* have an obligation to do her part in their saving two children. But as *they* have an obligation to do that, group obligation comes apart from individual obligation, and group obligation cannot be reduced to obligations of the members to do their parts.

The argument here goes as follows:

(1) Agnetha and Benny have an obligation collectively (together) to save two children by their going to room 2 and cooperating in removing the two children from the burning building.

(2) It is common certain knowledge among Agnetha and Benny that Benny will go to room 1 to try to save the child there.

(3) If Benny will go to room 1, then the morally best outcome that Agnetha can contribute to in that circumstance is saving the child in room 1 by going to room 1 and helping Benny to rescue the child there.

(4) If the morally best outcome that Agnetha can contribute to in the circumstance that Benny will go to room 1 is saving the child in room 1 by going to room 1 and helping Benny to rescue the child there and this is known to Agnetha, then Agnetha has an obligation to go to room 1 to help Benny save the child there.

(5) Therefore (2-4), Agnetha has an obligation to go to room 1 to help Benny save the child there.

(6) If Agnetha has an obligation to go to room 1 to do her part in saving one child with Benny, and that is incompatible with her going to room 2 to do her part in saving two children with Benny, then Agnetha does not have an obligation to go to room 2 to do her part in saving two children.

(7) Therefore (5, 6), Agnetha does not have an obligation to go to room 2 to do her part in saving two children.

(8) Therefore (1, 7), while Agnetha and Benny have an obligation collectively (together) to save two children by their going to room 2 and cooperating in removing the two children from the burning building, they do not both have an obligation to go to room 2 to do their parts in saving two children.

We have an obligation to prevent harm from happening through action or inaction. In the latter case, we are called on to act. In practice, we are limited by what harm

independent standpoint from which to generate collective moral obligations. The arguments I respond to in this section help to bring out another source of the intuition that collective and individual obligations come apart but I will show that in fact they do not come apart with respect to the same background circumstances.

we can foresee. Even then, we can't in general prevent all the harm that we can foresee. If what I have been saying is correct, then we still have the obligation, but we can't be blamed for what we cannot prevent (excepting non-cooperators who can contribute). But then what should we do when we cannot prevent all the harm we can foresee, that is, fulfill the strict requirements of duty? We should prevent as much harm as we can, taking into account what options are open to us. If I alone can save 1 or 2 children by going to room 1 or 2, but not all, then I should go to room 2 in those circumstances. If I cannot save the children in room 2 because it is inaccessible, I should go to room 1 in those circumstances. These are obligations of the form: I should **φ** given C. 'C' is replaced by a specification of the circumstances which delimit what you can expect reasonably to be able to bring about.

I want to allow that there is a sense in which (1) is true. And I want to allow that there is a sense in which (5) is true. But I will argue that the circumstances in which we understand (1) to be true are not the same as the circumstances that make (5) true.

Why is (1) true? Here we are thinking that each of them can do their part in their saving the children in room 2 and each of them knows (or ought to know) this and that it is the morally best thing to do *in the circumstances.* Here we do not include in the circumstances *what they intend or are willing to do and what they will or are likely to do*, that is, we bracket their intentions and likely actions. Call these circumstances C_1. So in C_1 they ought to save the two children and each of them ought to do their parts in that. (Here I assume that they can, relative to C_1, rationally arrive at the conclusion that they morally ought each to do their parts in their saving the children.) A crucial fact about these circumstances is that we have not included in them anything about the actual inclinations or intentions of either Agnetha or Benny. In thinking of what each *should do*, and so of what they should do, we treat their wills as *compliant to* the requirements of morality. For otherwise simply not intending or being willing to do something would alter the circumstances so that there was nothing they could do in the circumstances. We ask, bracketing what they intend or will actually do, given the circumstances, what are the requirements of morality on each of them: they each should go to room 2 and cooperate with the other in saving two children. So they ought to both go to room 2 and save the two children there.

Now consider Agnetha's position when she knows that Benny will not go to room 2 but will instead go to room 1. This adds an additional fact to the circumstances against which she evaluates what it would be best for her to do. Call these circumstances C_2. Fixing that Benny will go to room 1, the morally best outcome that Agnetha can contribute to is saving one child by going to room 1 and cooperating with Benny in doing that. Thus, in (5) the obligation Agnetha has is relative to C_2. In (1) in contrast the obligation Agnetha has is relative to C_1. When we then relativize the conclusion to the different circumstances, we get

> While Agnetha and Benny have an obligation collectively (together) in C_1 to save two children by their going to room 2 and cooperating in removing the two children from the burning building, they do not both have an obligation to go to room 2 to do their parts in saving two children in C_2.

But this is compatible with Agnetha having an obligation relative to C_1 to go to room 2 to cooperate with Benny in saving two children. So the argument does not show after all that they can have an obligation to save two children when Agnetha does not, relative to the same circumstances, which is what is needed to show that their obligation together can come apart from their obligations to do their parts.

What should we say about Benny in these circumstances? Benny is violating his duty in going to room 1. He is blamable for not choosing the best option, while Agnetha is not. Agnetha is not because she is responding to what he does, and she would go to room 2 conditional on Benny doing so. Benny's will is out of line with the requirements of morality; Agnetha's is not.

Can we answer the question what Benny and Agnetha should do together relative to C_2? There is a difficulty with this. As noted above, when we ask what a group of people should do when they can intervene to prevent some harm, we bracket what they intend or will actually do. We imagine that each will expect others to focus on what they can do together to achieve the best end. Each expects that the others, where the circumstances are public, will be willing to do her part in the ensemble of acts that will have the best outcome. Then we conclude each has an obligation to do her part in that. But if we ask about Agnetha and Benny in C_2, then the question is what each should do, given as a fixed point that one of them will go to room 1, full stop. The oddness of the question of what they should do lies in the fact that Benny has already decided what he will do. There is no question left for Benny to consider about what he should do in the circumstances in which he goes to room 1. It is a presupposition of asking what he should do regarding some matter that we don't include what he does. This then is also a presupposition of the question of what they should do in the circumstances. Thus, the question of what they should do in C_2 presupposes it is open what Benny does while the circumstances determine that it is not open.

In sum, the difficulty is that to show that a group can have an obligation when its members do not have obligations to do their parts, we need to relativize the obligations to the same circumstances, for here we are thinking about what they have most moral reason to do, and that depends on the circumstances. When we ask what an individual should do, we bracket what they intend to or will do on the matter. We do this because if we were to treat as part of the circumstance against which we evaluate what she should do *what she will in fact do*, then it is not open to her to do otherwise than she will *in those circumstances*, and the question is moot. The question of compliance to what duty requires in an agent's circumstances must focus only on the circumstances exclusive of the will of the agent and what she will do in response to the circumstances. We assume a compliant will and ask what the

agent would do to satisfy the requirements of duty. This same constraint applies when we ask what a group of individuals should do in some circumstances. Here we have in view the wills of each of the members. We then bracket, in asking what the group should do, what the individual members intend, and what they will or are likely to do, and ask essentially what they would do on the assumption that their wills are compliant to duty. But for each of them then the circumstances include the assumption of the wills of the others being compliant to their duties. When we focus not on the group, but on an individual member of the group, then the requirement that we bracket the wills of the others is lifted. We can then take into account what others intend, or will or are likely to do. This then can alter what it is morally best to do from the point of view of the individual, but this is also an evaluation relative to different circumstances. Thus, we do not get the verdict that relative to fixed circumstances the group obligation to do something and its members obligations to do their parts come apart.

Takeaways

First, we can see that one reason the basic intuition seems compelling is because a presupposition of asking what a group should do is that their wills are compliant to the demands of morality. So if we then ask what each should do, we think about this in the light of each of the others being compliant to doing what is best in those circumstances. Then the best that each can do is to make their contribution to the best outcome the group can achieve by all doing their parts in that. But this, I suggest, is a matter of the presupposition of the question, namely, that the parties' wills are not determined and are compliant to the demands of morality. When we focus on the question 'What should I do?' asked from the perspective of one member of the group, we presuppose only that the agent's will is compliant. We can then in principle get different answers to what *they* should do (given the presuppositions of the question) and what *individual agents* should do (given what additional facts can enter into the calculation given the different presuppositions). But this does not show that group obligations are not reducible to or derivable from individual obligations.

Second, when we ask what individual agents ought morally to do, we are restricting our attention to moral considerations. We are not asking all-in what the agent should do from the standpoint of individual rationality. If moral considerations are not overriding, then it can be rational for an agent to do something other than what is morally best or required. When we are evaluating outcomes with respect to moral considerations alone, however, differences in individual perspective on the value of outcomes disappears. The best moral outcome is by its nature best for everyone when we are restricting attention to moral considerations (see note 8 on agent-relative morality). A consequence of this is that what the

individual engaged in moral reasoning aims at in the context of group action is the best outcome for the group. There is no question about whether the best outcome from the group could come apart from the best outcome for the individual, as long as we are focusing on moral considerations alone. If participants can reasonably assume that it is public that all the potential participants are acting from moral considerations and informed about the outcomes, then except in cases in which there is a tie for the best moral outcome, there is a unique answer to the question 'What should I do?' which is determined by what the best outcome the group can achieve is, namely, my part (perhaps to be determined) in our achieving the morally best outcome.

This shows that moral reasoning by its nature requires something akin to group benefactor reasoning, in Bacharach's terms. In group benefactor reasoning, agents facing a decision about how to act in concert with others prioritize, in their individual preferences, group utility (however defined). If we think of maximizing group utility in terms of the best moral outcome, then reasoning from moral considerations alone is group benefactor reasoning.

Bacharach argued that team reasoning, which involves agency transformation, is not equivalent to group benefactor reasoning, which involves preference transformation (Bacharach 2006). The cash value of this distinction is that in team reasoning the individual starts with the question what is best for the team (if the team were an agent and I were the team, what would I do?), and then derives his part in that. If there is a strict ordering of the action ensembles in terms of group utility, team reasoning chooses the ensemble that is ranked highest. In group benefactor reasoning, the individual aims for the best team outcome, but still must think about it in terms of what the other players do. This allows that team reasoning, the argument goes, may resolve (by transforming) certain Prisoner's Dilemma games when group benefactor reasoning does not. For example, in a standard Prisoner's Dilemma such as that shown in Figure 6 (Petersson 2017),

Prisoner's Dilemma		Player 1	
		Cooperate	Defect
Player 1	Cooperate	4, 4	0, 5
	Defect	5, 0	3, 3

Figure 6: Prisoner's Dilemma

both team reasoning and group benefactor reasoning will look at a transformed matrix, as shown in Figure 7, which treats each square's value as the sum of the values of the players.[10]

[10] Donald Regan (1980, chapter 2, esp. p. 62) observed that Prisoner's Dilemmas are not possible for act utilitarians (AU) because there can be no agent-relative differences in the payoffs of combinations

Prisoner's Dilemma Transformed		Player 1	
		Cooperate	Defect
Player 1	Cooperate	8	5
	Defect	5	6

Figure 7: How the Prisoner's Dilemma in Figure 6 looks from the perspective of team reasoning and group benefactor reasoning.

Team reasoning recommends the combination of (Cooperate, Cooperate) because that is the highest payout for the group. For the group benefactor, that is clearly best, but the claim is that since each agent is engaged in individual strategic reasoning, what she should do depends on what the other does, for this is a Hi-Lo game.[11]

If we are focusing on moral outcomes, and the values in Figure 6 represent summable goods, then we will arrive at the payouts in figure 7. But I argued above that in this situation moral agents reasoning from the first-person perspective can rationally arrive at the decision to cooperate if it is common knowledge (or just reasonably assumed) that the players aim (or are likely to aim) for the best moral outcome or, equivalently, are paying attention only to moral considerations. In fact, in presenting the values in the decision matrix as moral values we presuppose each is reasoning from moral considerations, and if each knows the matrix that is relevant each knows that is true of the other. So team reasoning is not necessary to arrive at the correct decision in these cases.

Moreover, in many cases, even if you have very good reason to think the other will not pursue the morally best outcome, the highest expected value may favor

of actions. The result is the same for any theory (AU or not) on which the moral values realized by any combination of actions are the same for all those acting. As pointed out in note 8, this is compatible with agent-relative moral value based on special relations as long as it adds to the total for everyone. Wlodzimierz Rabinowicz (1989), in a very interesting paper, has argued that one can construct Prisoner's Dilemmas for AU. The arguments are too involved to go into here in detail. One depends on construing AU as requiring agents to maximize the utility of their own acts (which is a form of agent-relativity that introduces a bias toward bringing about a good oneself) and to ignore the past (even if relevant to the total value contributed to the universe by one's present actions) and then involves considering temporal sequences of acts by different agents and an organic value principle (the whole may be better than the sum of its parts). Another depends on adopting preference utilitarianism and future orientation and considering preference changes from the perspective of a single agent at different times. There are assumptions here which I would want to raise some questions about, but, in any case, the scenarios don't impinge on the discussion of the decision matrices considered in the text, which involve simultaneous choices by distinct agents.

[11] Paul Weirich (2018) responds to this argument by noting that what act one chooses can carry information about what act the other will choose when we assume "each player can predict his counterpart's response to his choices" (218). Player 1 choosing Hi "supplies a reason for its performance because it carries information that the other player, predicting his act, will choose High in response" (218). Choosing Hi is self-supporting. The same goes for choosing Lo, but Hi has greater self-conditional utility, so Hi is the rational choice.

cooperation. In Figure 7, given the values, the probability that the other will defect must be greater than .75 for it to be best to defect. So cooperating will be reasonable unless the prospects are quite dim.

Finally, say that in the game in Figure 7, player 1 knows player 2 will defect with probability .8. In these circumstances, player 1 should defect as well because that promotes the better outcome. As noted above, these are not the circumstances in which we would say that they ought to cooperate, so we do not get a mismatch between individual obligations and collective obligations. But what does team reasoning recommend? If player 1 engages in team reasoning, he will choose to cooperate, despite knowing this does not lead to the morally best outcome. Thus, team reasoning is inappropriate in the sense that it leads to what are known to be suboptimal outcomes. It would be morally wrong for player 1, given what he knows, to engage in team reasoning. It would be morally permissible to engage in team reasoning only if one independently had reason to think that would lead to the best outcome given how the other is likely to reason. This generalizes. When one ought not to act on the assumption that the other aims at the morally best outcome, one ought not to engage in team reasoning. It is permissible to engage in team reasoning only when one has independent reason to think the expected value of doing one's part in what would be the best outcome objectively is higher than other choices. It is therefore not appropriate to decide what one ought to do by first adopting the strategy of team reasoning and then deriving what one ought to do from that. The fundamental standpoint from which to derive one's obligations remains that of the individual agent thinking about what is best given the actual facts about what it is reasonable to think others will or are likely to do.[12]

References

Bacharach, Michael. 2006. *Beyond Individual Choice: Teams and Frames in Game Theory*. Princeton: Princeton University Press.

Blomberg, Olle, and Björn Petersson. 2023. "Team Reasoning and Collective Moral Obligation." *Social Theory and Practice*. doi: https://doi.org/10.5840/soctheorpract2023120177.

Brownlee, Kimberley. 2010. "Reasons and ideals." *Philosophical Studies* 151 (3):433-444.

Collins, Stephanie. 2019. *Group Duties: Their Existence and Their Implications for Individuals*. Oxford: Oxford University Press.

Colman, Andrew M., Briony D. Pulford, and Catherin L. Lawrence. 2014. "Explaining Strategic Coordination: Cognitive Hierarchy Theory, Strong Stakelberg Reasoning, and Team Reasoning." *Decision* 1 (1):35-58.

[12] My thanks for very helpful comments from Olle Blomberg and Mattias Gunnemyr.

Dietz, Alexander. 2016. "What We Together Ought to Do." *Ethics* 126 (4):57-69.

Harsanyi, John C. , and Reinhard Selton. 1988. *A General Theory of Equilibrium Selection in Games*. Cambridge, Mass.: MIT Press.

Heuer, Ulrike. 2010. "Reasons and impossibility." *Philosophical Studies* 147 (2):235 - 246.

Lord, Errol. 2015. "Acting for the Right Reasons, Abilities, and Obligation." *Oxford Studies in Metaethics* 10:26-52.

Petersson, Björn. 2017. "Team Reasoning and Collective Intentionality." *Review of Philosophy and Psychology* 8 (2):199-218.

Rabinowicz, Wlodzimierz. 1989. "Act-utilitarian prisoner's dilemmas." *Theoria* 55 (1):1-44.

Regan, Donald. 1980. *Utilitarianism and Co-operation*. Oxford: Oxford University Press.

Schwenkenbecher, Anne. 2019. "Collective Moral Obligations: 'We-Reasoning' and the Perspective of the Deliberating Agent." *The Monist* 102 (2):151-171.

Streumer, Bart. 2007. "Reasons and Impossibility." *Philosophical Studies* 136 (3):351-384.

Streumer, Bart. 2010. "Reasons, impossibility and efficient steps: reply to Heuer." *Philosophical Studies* 151 (1):79 - 86.

Sugden, Robert. 2000. "Team Preferences." *Economics and Philosophy* 16:175-204.

Weirich, Paul. 2018. "Rationality and Cooperation." In *The Handbook of Collective Intentionality*, edited by Marija Jankovic and Kirk Ludwig, 209-220. New York: Routledge.

Pragmatic Challenges in Practical Ethics

Christian Munthe

Abstract. This brief essay traces a development of orthodox applied ethics into a present-day variant of practical ethics, where pragmatic reasons may upset ideal theoretically and empirically informed epistemically supported ethical prescriptions when these are to be implemented in a real context. This shift comes with a development where the applied ethicists of older days are nowadays aiming for much more specific and practically useful action-guidance, and for activist involvement to support feasible implementation of ethical prescriptions. This results in a radical and a moderate activist variant of practical ethics, both of which face specific challenges due to the necessity of considering pragmatic reasons. I argue that the radical variant has trouble managing these challenges. The moderate variant may manage them, but this may require substantial methodological development.

The aim of this brief essay is to conduct a sort of "pilot" study of a set of challenges emerging out of recent trends to allow pragmatic considerations to play an increasing role in applied, or practical, ethical analysis.

While pragmatism continues to be a debated view in general epistemology, philosophy of science and metaethics (Legg & Hookway 2021; Sayre-McCord 2014), the role of pragmatic arguments in substantial normative ethics has been much less scrutinized. I will briefly outline several distinct ways in which pragmatic considerations have started to be viewed as good reason to modify otherwise theoretically well-founded normative positions in practical ethics, due to an expansion of aims compared to more orthodox applied ethical approaches. I will not

defend this expansion as such, but rather trace a particular set of implications of it that may be viewed as challenges to the field.

There are partial precursors of addressing the role of pragmatic arguments in practical ethics. Since John Rawls introduction of conditions of "well ordered societies" and "stability" in his theory of the "original position" supposed to justify his substantive theory of justice (Rawls 1971) – constraining what principles may be chosen for what contexts by the idealized parties in it – a still rather tentative discourse on so-called *non-ideal theory* goes on in the contemporary social contract tradition of political philosophy (Jubb 2012; Schmidtz 2011; Thompson 2020; Valentini 2012). However, my aim here is not bound by either this philosophical tradition, the narrow focus on justifying political institutions, or the conditions of "well-ordered societies" and the import of political legitimacy highlighted by Rawls' stability condition. Rather, I will describe a broader conception of pragmatic reasons in practical ethics, of which the non-ideal theory debate in political philosophy is one instance. I will provide some concrete examples of what the impact of including pragmatic reasons may have on the conclusions of specific practical ethical arguments.

I will start by explaining in more detail what I include in the notion of a *pragmatic reason* and distinguish two distinct variants of such reasons. I will then present an analysis of how to determine the impact of these respective types of pragmatic reasons on the soundness and/or validity of practical ethical arguments. This includes reviewing the development from the orthodox applied ethics that emerged in the 1970's and into present-day practical ethics. Lastly, I will present and preliminary assess some challenges for practical ethics created by the recognition of the type of pragmatic reasons I have outlined. My tentative conclusion is that, while these challenges should be considered and may motivate further adaption of argumentative models in practical ethics, they do not undermine the basic rationale for acknowledging pragmatic reasons in practical ethics.

Pragmatic Reasons – Weak and Strong

The notion of a pragmatic reason employed in this essay is meant to capture a notion of a type of normative reasons that is broader but related to the idea of non-ideal political theory in Rawlsian and related social contract theoretical discourse. More specifically, a pragmatic reason is a consideration that provides a reason for what to do in a practical context, albeit not grounded in appeals to the epistemic rationality of accepting such normative claims. Typically, such reasons are activated when the ethical analysis moves from asking what should (ideally) be done, to asking whether this is practically feasible to implement effectively (without changing the conditions of the ideal theoretical justification). These reasons are here called pragmatic, since they arise out of no epistemic argument for or against the truth of ethical

judgements, but rather from people's responses to such judgements and attempts to implement them.

More specifically, there are two main types of sources of such pragmatic reasons:

> **Disagreement**: Key actors for successful implementation do not embrace the ideal ethical theory in question, are not persuaded by arguments based on it, and thus will likely not adhere to its prescription regarding the practice in question.
>
> **Illegitimacy**: Attempting to implement the (ideally) recommended practice will not be tolerated by key actors and/or those to which the recommendation applies, and it will therefore not be effectively implemented, enforced, or respected.

Now, in many cases, there are options to aid implementation of ideal theoretical prescriptions with measures directed to influence factors such as these, mostly meant to either persuade people to change their minds (rhetoric and manipulation), or to force or lure them into compliance (politics). Sometimes, the outcome of adding pragmatic considerations to an ideal theoretical ethical argument when deciding what to do comes down to no more than this: the pragmatic reasons recommend further actions that complement and are compatible with the ideal theoretical recommendation. For instance, this is what Rawls does when he advocates having societies tell the narrative of how the original position (allegedly) justifies his substantive theory of justice as a way of increasing the stability of a society implementing that theory, even making this part of a political enforced educational system justified by the same theory (Rawls 1971). Such pragmatic reasons I consider weak, as they do not upset the initial ideal theoretical conclusion. However, as Rawls recognized, the impact of such weak pragmatics has limits. Any attempt to persuade people of the plausibility of a recommendation, or to lure or force them to comply with it may be wasted, or even backfire so that the unwillingness to accept and comply increases even more. In addition, the additional measures may incur costs in monetary or other terms that undermine the initial ideal ethical justification of the recommendation.

This leads over to the strong notion of pragmatic reasons, where these are thought to provide considerations which are *at odds with* the ideal theoretical conclusion. In the following, this is the pragmatics I will be focusing on. These strong pragmatic reasons tell us to adapt an ideal theoretical ethical conclusion to the pragmatics, so that the threat of the pragmatic factors against feasible effective implementation is mitigated. In the Rawlsian case, this is the non-ideal theoretical option of weakening those normative aspects of the substantive theory of justice which are resisted by key actors and the population, which has been extensively questioned and debated in the social contract theory context. My contention, however, is that the notion of a (strong) pragmatic reason is viable across (applied or practical) normative ethics. There exists a wide variety of applied or practical normative ethical domains of inquiry (different areas of politics, various domains of private life and personal

relationships, professional conduct and practice, non-political institutions such as business or religious communities and so on), and a manifold of normative theories to use as assumptions regarding what may justify specific normative conclusions in all such discourses. Regardless of discourse and theoretical framework, we may distinguish ideal theoretical reasons to justify conclusion related to some discourse, based on some theory, from reasons not based on this theoretical framework (or any of its ideal-theoretical competitors) and/or not directly a part of the discourse in question (albeit allegedly relevant for it) that refer to pragmatic factors of disagreement and/or illegitimacy, and that are claimed to provide strong reasons at odds with the ideal ethical conclusion.

Let me illustrate with the help of a well-known contested practical ethical issue, that of (voluntary) euthanasia (that is the practice of a health care professional to intentionally kill a patient on this patient's request). Suppose that your favorite ideal ethical theory supports the notion of legalizing this practice. The types of pragmatic considerations described above may provide additional reasons against attempting to act on this prescription. This since disagreement and illegitimacy undermine the feasibility of the prescription; although legalization may be a real option, it is unlikely that attempting to have this option performed will be successful. One may then, of course, try to make people change their minds and behaviors – e.g., to have medical professional organizations adopt voluntary euthanasia as accepted practice (as in Belgium and the Netherlands), or to force them to comply (as in Canada). But suppose not much comes out of that, or that the actions needed to be taken have substantial downsides (such as excessive force or liberty restriction). Then the question arises whether there is some way to modify the suggestion in a way that may increase support for it. For instance, rather than hopelessly petitioning parliaments to make exceptions for certain types of murder (in the legal sense), one rests content with developing a clinical routine for physician's assisted suicide (which we here assume to be legal) that could be incorporated into accepted medical practice (in the ordinary way, via clinical trials, consensus conferences, et cetera). This in spite of the fact that the ideal theoretical ethical reasons supporting physician's assisted suicide lend equal support to voluntary euthanasia.

Of course, all these types of pragmatic reasons against an ideal theoretically supported legalization of euthanasia may also appear to undermine ideal theoretical cases for banning euthanasia. My point here is not to plant doubts against any specific position regarding this ethical issue. Moreover, the pragmatic reasons may occur also in relation to other ethical issues than those regarding the legal banning or permission of some practice, such as issues about what should be an accepted practice or ethos of some profession, or how groups of people, such as informal communities, or individuals should act in different situations.

From Orthodox Applied to Present-day Practical Ethics

The orthodox notion of (philosophical) applied ethics, as it emerged throughout the 1970's and -80's, has mainly been to produce arguments of the following form:

> (Allegedly justified) ideal normative ethical/political theory
>
> Relevant factual assumptions about the nature of available options and consequences of these with regard to some practice
>
> ---
>
> Conclusion (*ceteris paribus*) of what of the options should be chosen

This form has served moral (and political) philosophy quite well in helping to elucidate the specific concrete implications of abstractly formulated theories. Such elucidation has further enrichened both the critical assessment and justification of philosophical theories, as well as the development of new, more sophisticated theories. It has also been quite useful to help practical deliberation trace different arguments back to specific sets of factual assumptions and/or ethical/political theory-elements. But that was considered to be the end of the ethical analyst's job. Figuring out the factual details to unpack the *ceteris paribus* clause in specific, concrete cases, or fixing successful effective implementation was not considered to be on the applied ethicist's agenda. The typical reaction of an orthodox applied ethicist when facing the disagreement or illegitimacy factors would be to note that the ideal theoretical ethical conclusion still stands, and that key actors and people are irrational and/or immoral if they do not accept it or do not adhere to it (this follows from the ideal theoretical ethical conclusion).

Over the years, as applied ethicists started to become engaged more closely with decision-makers and specific practical problems, applied ethicists started to interact more with practical details and other research areas to figure out how useful specific guidance could be teased out of the orthodox applied ethical conclusions. Among the complexities uncovered by such interactions, as well as the general experience of how allegedly firmly justified proposals were often ignored or distorted in the practice of policy making and implementation, were the type of pragmatic factors highlighted in the former section. The aims of the field were expanded and made more ambitious.

The first part of this development produced a strong shift towards *empirically informed* applied ethics, and today it is considered standard practice to have applied ethics research projects involve not only philosophical ethicists, but also practitioners and researchers from relevant fields, such as the bio- and

technosciences, law, politics, psychology, and so on. The aim is no longer to be content with theoretically assumed research questions and generic *ceteris paribus* conclusions resting on contested normative assumptions. Rather, this step is now seen as a first step towards further advanced analysis of factual and normative uncertainties, institutional complexities and public opinion and values (see Beauchamp 2003 for an illustrative example). This development has helped orthodox applied ethics to become more *practical*, more capable of assisting decision-makers with real solutions to the ethical problems they are in fact facing. While aiming for more useful, specific and action-guiding prescriptions, this development is still compatible with the orthodox shrug response to pragmatic feasibility factors. However, it has given rise to critical debates on both the role of empirical research in practical ethical analysis (Davies et al. 2015) and what has sometimes been termed "the ethics of ethics", i.e. questions about the ethical limits of practical ethicists to aid decision-makers with regard to questions selected by the latter and the obligations of ethicists to promote the good and the right (Eckenwiler & Cohn 2007).[1]

The second part of this development consists of ethicists expanding their ambitions to go beyond the production of normative advice and prescriptions, and to participate in the practice of implementing these prescriptions in the form of specific policy and activities on the ground to pave the way for rolling out such policies. My own view of this development is that it is expected on the basis of the increased awareness of practical complexities and pragmatic factors coming out of the empirically informed applied ethics, and the increased focus of the "ethics of ethics" discourse on the importance of not only identifying the good and the right, but to actually see it done. This step has eventually led to debates on the soundness of such scholarly practical ethical "activism" (Brody 2009; Draper et al. 2019; Eckenwiler & Cohn 2007), but I will sidestep these here. Instead, my point is to highlight that this step undermines the aforementioned shrug response to pragmatic complications. These must now be considered by the practical ethicist as part of the "activist" work to aid the effective implementation of a prescription. Pragmatic reasons may thus undermine the ideal theoretical reasons to act on a prescription, however well justified it may be from an (empirically informed) theoretical standpoint. Of course, such reasons may serve to motivate policy measures to persuade, lure and force people in view of widespread disagreement and illegitimacy. But as observed earlier, the justification for such additional measures will always have limits, and in many cases these limits can be expected to be transgressed. Then, the "activist" aim would seem to provide reasons to revise the

[1] A case in point may be the role of self-labelled "pragmatic" or "practical" bioethicists in facilitating bogus stem cell treatment hoaxes, or what we now know as the opioid epidemic scandal in the US. See this string of blog posts for more information about these matters:
https://philosophicalcomment.blogspot.com/search?q=McGee&max-results=20&by-date=true

prescription one seeks to implement, at odds with ideal theoretical justification, in order to adapt it to the pragmatic circumstances to make it more feasible.

At the same time, this step into "activism" can be radical or moderate. The radical step is when the activism leaves all further normative ethical considerations aside, thus abstaining from further critical inquiry into one's own assumed basic ethical standpoint. Depending on which standpoint this is, such an approach may express itself as fanatism (as with some examples of some identity political activist scholars[2]) or nihilism (where the ethicist becomes a brain for hire, no matter the purpose, as in the scandals mentioned in footnote 1). A more common move, however, is that the aim to aid and participate in implementation is *added to* (rather than replacing) the more ideal theoretical and empirically informed parts – and this is the type of activism I label moderate.

Challenges Due to Increased Room for Pragmatics

Allowing pragmatic considerations to enter normative ethical (including political) arguments with an independent force of their own implies a number of philosophical challenges. Some of these have been noted and debated in the context of non-ideal social contract theorizing to some extent, but not all. I will here work through these challenges in relation to a generally conceived practical ethics (as sketched in the foregoing section) rather briefly. My modest aim is an attempt at an overview and a sketch of how to manage them in a systematic and justified manner.

> *Challenge 1: Undermined normativity?*
>
> If the soundness of practical ethical conclusions is allowed to be constrained by what is feasible in view of the factors of disagreement or illegitimacy, the substance of normative ethics is undermined, since these conclusions may be entirely determined by what people, communities and institutions actually do or prescribe, not what they should do or prescribe.

This challenge seems false both for radical variants of practical ethics, and for moderate ones. In both cases, normativity is there (fixed by the assumed ideal theoretical basis, or in continuous critical question). What both variants do is to trace new sources of normativity, rather than abandoning normativity. At the same time, the room for pragmatically based critique of ideal theoretical recommendations in the moderate variant makes the way it traces these normativity sources very different from that of the radical variant. This leads over to the next challenge.

[2] See, e.g., Munthe (2020).

Challenge 2: No room for objective ethical truth?

> If the soundness of practical ethical conclusions is allowed to depend on what people, communities and institutions actually prescribe and how they will in fact respond to their (attempted) implementation, the possibility of objective ethical truth regarding specific practical matters is ruled out.

The force of this challenge, of course, depends on to what extent "objective ethical truth" is a viable prospect in the first place. But suppose it is, and suppose that some (perhaps not yet formulated) ideal ethical theory unveils it. Then it seems to follow that truths about more specific practical ethical matters are necessarily more dependent on subjective factors, such as what people believe is right and good in practice, or the extent to which they are prepared to act in accordance with such beliefs. The force of the challenge also depends on what "objective" is supposed to mean, and to what extent the subjective elements introduced by pragmatic considerations really undermine objectivity in any *serious* way.[3]

I, for one, am not at all certain that they do for the moderate variant. First, the pragmatic aspects are not in any way opposed to the notion of an ethical judgement as true or false, they merely introduce new truth-conditions for such judgements. Moreover, these new truth-conditions, while having subjective aspects detached from normal epistemic reasons, are mostly compatible with the possibility of a person, group or institution being mistaken. The subjectivity introduced does not entirely determine ethical truth. So, while the unwillingness of a population to embrace or act in accordance with an ideal theoretically supported ethical prescription may undermine the ethical soundness of applying the prescription to this population, there is still room to claim that this population *should* embrace and act in accordance with it. This holds even if we consider only one single individual, albeit in that case we may face intricate normative ethical problems of how to think ethically about people we know to be irrational or suffering from weak character.

For the radical variant, however, this challenge seems more difficult to get around. This since the radical variant of practical ethical activism has severed any link to ideal theoretical critical reflection. And it is this type of reflection that allows the moderate to both say that we have to adapt to what people actually believe to gain legitimacy for practical proposals and that they should believe other things that would make such adaptions needless.

[3] At the same time, as pointed out by one anonymous reviewer of this chapter, acknowledging the need to consider pragmatic reasons may turn out to make a difference to *what* ideal normative ethical theory about objective moral truths best holds up to scrutiny. Theories such as traditional Kantian ethics, with very strict and abstract criteria of rightness may be undermined by requirements to consider pragmatic reasons. This may be taken by those independently convinced of the truth of these theories to question the move into "activism" taken by contemporary practical ethics.

Challenge 3: Vulnerability to strategic manipulation?

> If the soundness of practical ethical conclusions is allowed to depend on what people, communities and institutions think about these arguments and how they will in fact respond to their implementation, practical ethics becomes vulnerable to strategic manipulation by partisan, vested or similarly partial or biased interests.

This is a true challenge that comes with the Marx-inspired move of practical ethics from merely trying to explain what is right and good in our world to also trying to change this world for the better. That change places the ethicist in a social context where the responses of others to the fruits of one's labour are not merely to be reckoned with intellectually, considering standards of epistemic rationality. The practical ethicist also must consider *practical rational* aspects about the consequences of and grounds for devising and advancing particular arguments and conclusions. From such a practical rational standpoint, it is desirable to avoid actions or patterns of behaviour that tend to make the practical ethicist vulnerable to exploitative manipulative strategies, such as blackmail or lures of a *dutch book*[4] nature. Certain ways of adapting to pragmatic factors may thus be considered practically irrational. The basic reason for this is that adapting to these types of strategy inevitably leads to an outcome where the agent gives up everything, while the responding player gains everything. I will not here try to solve this challenge, merely point out three aspects of it that, to my mind, make it solvable, at least for moderates.

First, the pragmatic considerations of practical ethical turn do not erase the ideal theoretical considerations of orthodox applied ethics, merely complement them. This undermines the necessity of sliding all the way to the complete loser of a completed *dutch book* or a blackmail scenario. Just as political negotiation parties have a limit for how much to concede in compromises, ideal theoretical considerations temper the impact of pragmatics on the conclusion of a practical ethical argument. This response is only available to moderates. Second, the risk of strategic manipulation (and reasons to avoid it) can itself be considered as a pragmatic aspect. Third, in many cases there exist practical options that may serve to change the pragmatics of a situation, e.g. make people less disposed to strategic manipulative behaviour, and such options may be worked into the practical ethical analysis. Still, there remains to explore how the logic and rationale motivating a particular balancing of ideal-theoretical and pragmatic reasons should look like.

Challenge 4: A heuristics paradox?

> Since pragmatic considerations constrain (sometimes considerably) what conclusions may be supported in practical ethics, while these considerations themselves may change (or be changed) over time due to ideal theoretical ethical considerations,

[4] A *dutch book* is a hazard game odds strategy where each step of the game (rationally) lures a player to continue to bet in a series that will necessarily make this player lose the game all things considered.

> practical ethics faces a paradox regarding whether to start ethical analysis from the pragmatic or the ideal theoretical end.

Also this challenge for practical ethics seems to me to be a genuine one. At the same time, it arises only in those cases where pragmatic considerations considerably constrain what ideal theoretical conclusions are feasible, and where the prospect are good for having ideal theoretical disputation change prevailing opinions, attitudes and behaviours that affect feasibility. In those cases, however, I believe there is a practical way forward that can be grounded in the aims of practical ethics, at least for moderates.

Just as with the handling of the strategic aspects of the preceding challenge, the threat of genuine paradox can be avoided by building the prospect of having ideal theoretical reasons change pragmatic considerations over time into these pragmatic considerations themselves. As I pointed out regarding the first two challenges, the need to consider pragmatics in practical ethics does not undermine the normativity of ideal theoretical ethical reasons, especially not if these reasons are true. There may thus be excellent ideal theoretical reasons to advocate these very reasons to key actors, communities and institutions to have them change in a way that relaxes the tension between what is pragmatically feasible and what should ideally be done or occur in a certain context. At the same time, such reasons have to be scrutinized for pragmatic feasibility as well: in some contexts, trying to change the pragmatics may be a waste of time and resources and therefore unethical.

As this response rests on the availability of ideal theoretical ethical inquiry and discourse, it would not seem to be available to radicals. However, I conjecture that such radicals, faced with the threat of the heuristics paradox, would most likely find reasons to abandon the radical approach.

In all, I thus see potential for a moderately activist practical ethics to handle the challenges coming out of the necessity of considering pragmatic reasons to revise and adapt ideal theoretically supported ethics conclusions. At the same time, it is obvious that this potential management needs a developed methodology that is not yet in existence.

Conclusion

I have traced a development of orthodox applied ethics into the present day of a practical ethics aiming not only for philosophical analytical rigour, but also for advanced action-guidance and practical usefulness, stretching into the realm of "activism" to implement practical ethical prescriptions. I have argued that the last element brings with it a necessity to consider pragmatic reasons, outside of the epistemic reasons to accept or deny different ethical conclusions, and to adapt practical ethical conclusions to promote feasible effective implementation. The resulting practical ethics comes in a radically and a moderately activist variant. I

have identified four distinct challenges for this approach to practical ethics and argued that the radical variant may handle some of them, but not all, and that this provides arguments to abandon the radical stance for a more moderate one. This moderately activist practical ethics may escape or manage all of the challenges, but to do so, it needs to further develop its own methodology.[5]

References

Beauchamp, Tom L, 2003, "Methods and Principles in Biomedical Ethics", *Journal of Medical Ethics,* 29, 269-274, doi: http://dx.doi.org/10.1136/jme.29.5.269

Brody, H, 2009. *The Future of Bioethics*. Oxford: Oxford University Press.

Davies, R. Ives, J, Dunn, M, 2015, "A systematic review of empirical bioethics methodologies", *BMC Med Ethics* 16, doi: https://doi.org/10.1186/s12910-015-0010-3

Draper, Heather, Moorlock, Greg, Rogers, Wendy, Scully, Jackie (eds.), 2019. "Special Issue: Bioethics and Activism", *Bioethics*, 33(8), 51-978

Eckenwiler, Lisa A, Cohn, Felicia G (eds.), 2007. *The Ethics of Bioethics: Mapping the Moral Landscape.* Baltimore, MA: Johns Hopkins University Press.

Jubb, Robert, 2012. "Tragedies of Non-ideal Theory, *European Journal of Political Theory*, 11(3), 229-246, doi:10.1177/1474885111430611

Legg, Catherine and Hookway, Christopher, 2021. "Pragmatism", In: *The Stanford Encyclopedia of Philosophy* (Summer 2021 Edition), Edward N. Zalta (ed.), URL = https://plato.stanford.edu/archives/sum2021/entries/pragmatism/

Munthe, Christian, 2020. "Bioethics, Disability, and Selective Reproductive Technology: Taking Intersectionality Seriously", In: *The Oxford Handbook of Philosophy and Disability*, Cureton, A & Wassermann, DT (eds.), Oxford: Oxford University Press, doi: https://doi.org/10.1093/oxfordhb/9780190622879.013.41

Rawls, John, 1971. *A Theory of Justice*. Oxford: Oxford University Press

Sayre-McCord, Geoff, 2014. "Metaethics", In: *The Stanford Encyclopedia of Philosophy* (Summer 2014 Edition), Edward N. Zalta (ed.), URL = https://plato.stanford.edu/archives/sum2014/entries/metaethics/

Schmidtz, David, 2011, "Nonideal Theory: What It Is and What It Needs to Be", *Ethics,* 121(4), 772-796, doi: https://doi.org/10.1086/660816

Thompson, Christopher, 2020. "Ideal and Non-ideal Theory in Political Philosophy", *Politics*, doi: https://doi.org/10.1093/acrefore/9780190228637.013.1383

Valentini, Laura, 2012. "Ideal vs. Non-ideal Theory: A Conceptual Map", *Philosophy Compass*, 7, 654-664

[5] I am very grateful to all who gave comments when I presented a first, rough draft of this essay at the research seminar in Practical Philosophy and Political Theory at the University of Gothenburg.

In Defence of Mooreanism

Jonas Olson

Abstract. In his recent book *The Value Gap* (2021), Toni Rønnow-Rasmussen defends a pluralist view of final goodness and goodness-for, according to which neither concept is analysable in terms of the other. In this paper I defend a specific version of monism, namely so-called 'Mooreanism', according to which goodness-for is analysable partly in terms of final goodness. Rønnow-Rasmussen offers three purported counterexamples to Mooreanism. I argue that Mooreanism can accommodate two of them. The third is more problematic, but this is in the end not a decisive objection.

1. Introduction

In his recent book *The Value Gap* (2021), Toni Rønnow-Rasmussen argues that there is a fundamental gap between final goodness and goodness-for (a person or some other kind of entity). A plausible theory of value should therefore be in a crucial sense pluralist rather than monist. As Rønnow-Rasmussen explains, '[w]e are value pluralists (or at least dualists) if we believe that 'good' and 'good for' denote two kinds of value neither of which can be understood in terms of the other.'[1] Monist views do not recognize a fundamental gap and attempt either to analyse final goodness in terms of goodness-for, or goodness-for in terms of final goodness, or, more radically to eliminate the one in favour of the other. Rønnow-Rasmussen calls the view that takes final goodness to be fundamental 'Mooreanism', because of its close affinities with some claims G. E. Moore made in his seminal work *Principia Ethica* (1993 [1903]).[2]

[1] Rønnow-Rasmussen 2021: 24.

[2] 'What then is meant by "my own good"? In what sense can a thing be good *for me*? It is obvious, if we reflect, that the only thing which can belong to me, which can be *mine*, is something which is good,

No version of monism is tenable, according to Rønnow-Rasmussen. In this paper, I shall defend Mooreanism against his criticism. The gist of the criticism of monist views is that 'they must on purely formal grounds renounce certain value claims as being nonsensical.'[3] This criticism rests on a methodological principle of substantive neutrality, according to which no formal view of value—whether monist or pluralist—should rule out coherent substantive views about value.[4] It is not entirely clear whether Rønnow-Rasmussen intends this principle to state a necessary condition of adequacy or merely a desideratum. I shall return to this point in section 4.

Rønnow-Rasmussen's challenge to Mooreanism takes the shape of three examples of substantive evaluative claims about goodness-for, which Mooreanism allegedly fails to make sense of. I shall argue that in at least two of the three cases, Mooreanism can make good sense of the claims. While the third case is more debatable, it is in the end not a decisive challenge to Mooreanism. All of this will be dealt with in section 4. Before I get there, I shall make some further points about Mooreanism in section 2, and in section 3 I shall formulate a version of Mooreanism to use for the illustrative purpose of showing how Mooreans can handle Rønnow-Rasmussen's counterexamples.

2. Kinds of Mooreanism

Rønnow-Rasmussen distinguishes between Radical Mooreanism and Mooreanism, accordingly:

> *Radical Mooreanism:* 'Final goodness-for' expresses either an incoherent value notion or one that can (and should) be replaced entirely by the non-relational notion of what is finally good.
>
> *Mooreanism:* 'Final goodness-for' expresses either an incoherent value notion or a derivative notion that is ultimately dependent on the non-derivative notion of what is finally good.[5]

One might think that Radical Mooreanism implies Mooreanism, but this is not obviously so for at least two reasons. First, the view that one 'value notion' can and

---

and not the fact that it is good. When, therefore, I talk of anything I get as "my own good", I must mean either that the thing I get is good, or that my possessing it is good. In both cases it is only the thing or the possession of it which is *mine*, and not *the goodness* of that thing or the possession.' (Moore 1993 [1903]: 150, italics in original).

[3] Rønnow-Rasmussen 2021: 26.

[4] Rønnow-Rasmussen 2021: 26, 37.

[5] Rønnow-Rasmussen 2021: 34, *Final* goodness-for is to be contrasted with *instrumental* goodness-for. What is instrumentally good-for is that which is conducive to goodness-for.

should be replaced by another does not have to be based on the view that the one is ultimately dependent on, or derived from, the other. Second, while Mooreanism implies that final goodness is non-derivative, Radical Mooreanism does not (provided that final goodness being non-relational does not imply its being non-derivative).[6]

It will be useful to anchor the two versions of Mooreanism in the recent debate on value theory, and to identify representatives of each. I take Tom Hurka (2021) to be a Radical Moorean and Guy Fletcher (2012) to be a Moorean. Hurka argues that 'there's no philosophically useful understanding of "good for" and related terms that's *both* evaluative rather than merely descriptive or naturalistic *and* irreducible to other evaluative concepts, in particular "simply good"'.[7] Thus far, Hurka's view seems compatible with both versions of Mooreanism. But since he adds that there is 'no understanding [of "good for"] on which it makes a substantive contribution to ethics'[8] and that 'it would serve clarity if philosophers used only [the] phrase ["simply good"]'[9], I take him to be a defender of Radical Mooreanism. (Hurka uses the phrase 'simply good' to express roughly the same concept as Moore and others following him called 'intrinsic goodness' and that many philosophers nowadays call 'final goodness'.[10])

In contrast to Hurka, Fletcher thinks that the property of good for 'does normative work' in generating normative reasons.[11] But Fletcher also defends a so-called 'locative' analysis of what it is for something to be good for a person. The details of the analysis need not concern us here. What is important to notice is that one of its necessary conditions for something, G, being good for a person, is that G is 'non-instrumentally', or finally, good.[12] Hence, G's being good for a person would seem to be 'ultimately dependent on' G's being finally good. This makes it apt to interpret Fletcher's locative analysis of goodness-for as a version of (non-radical) Mooreanism.[13]

---

[6] I am indebted to Jens Johansson for the second point.

[7] Hurka 2021: 804, emphases preserved.

[8] Hurka 2021: 804.

[9] Hurka 2021: 821.

[10] Mooreans (in the sense of 'Mooreanism' relevant to this paper) take different views on whether final goodness can only depend on properties intrinsic to what is finally good, and may therefore differ over whether 'intrinsic goodness', 'final goodness', 'goodness *simpliciter*', or some other phrase is to be preferred. For discussion, see, e.g., Rønnow-Rasmussen 2015.

[11] Fletcher 2012: 17.

[12] The other two necessary conditions—that together with the first are jointly sufficient—for G being non-instrumentally good for X are that 'G has properties that generate, or would generate, agent-relative reasons for X to hold pro-attitudes towards G for its own sake, [and that] G is essentially related to X.' (Fletcher 2018: 3)

[13] Fletcher claims, however, that the locative analysis retains all of its merits even if the property of being good for does no work in generating normative reasons (2018: 17). The locative analysis might thus be consistent with Radical Mooreanism.

Another notable feature of the locative analysis is that Fletcher offers it not as a conceptual analysis, but as a metaphysical analysis of the property of being good for.[14] This point is relevant to Rønnow-Rasmussen's general charge that Mooreanism, *qua* monist view, must renounce apparently coherent substantive claims about goodness-for as 'nonsensical' (see section 1). If Mooreans analyse the property of being good for—but not the concept of goodness-for—partly in terms of final goodness and if Rønnow-Rasmussen is right that there are some substantive claims about what is good for that they cannot affirm, it is not obvious that they have to renounce these claims as nonsensical rather than simply false. However, it is clear that for an analysis of goodness to renounce a substantive view about what is good on purely 'formal' grounds is also to violate Rønnow-Rasmussen's methodological principle of substantive neutrality.

3. A(n) (Over)simplified Moorean Analysis of Goodness-for

To investigate whether Mooreanism can handle Rønnow-Rasmussen's purported counterexamples, let us for the sake of illustration adopt a simplified Moorean analysis of goodness-for. Taking a cue from one of Hurka's suggestions, let us say that for something, G, to be good for a person, *a*, is for G to be finally good and appropriately related to *a*.[15] The question of what it is to be 'appropriately related' to *a* is a difficult one that will receive no definite answer here. Suffice it to say that it is an independently plausible assumption that something is good for a person only if it is appropriately related to the person. And I infer from this that any account of goodness-for that aspires to be fully explanatory should tell us something about what it is to be appropriately related to a person.

For illustrative purposes, we can once again follow Hurka, who adopts Peter Railton's suggestion that, necessarily, whatever is good for a person is such that the person finds (or would find, if she were to reflect) it compelling or attractive; it finds (or would find) a 'resonance' in the attitudes of the person.[16] We can then say that for G to be good for a person *a*, is for G to be finally good and to resonate with *a*'s attitudes. Let us add to the proposal that whatever is *bad* for a person is such that the person finds it repelling or unattractive; it finds (or would find) a 'dissonance' in the attitudes of the person. We can then say that for G to be *bad* for *a*, is for G to be finally bad and to dissonate with a's attitudes. We can also say that whatever does not resonate or dissonate with *a*'s attitudes is neutral for *a*, regardless of

[14] Fletcher 2012: 5.

[15] Hurka 2021: 806.

[16] Hurka 2021: 811; Railton 2003: 47.

whether it is finally good, bad, or neutral. The analysis is related to Fletcher's locative analysis but clearly simpler than it, and quite possibly oversimplified, but it will serve our illustrative purpose well.

No doubt, the analysis would have to be qualified in several ways to be ultimately defensible. For example, the attitudes in question should presumably be suitably idealized. I assume that this can be done in a way that is consistent with taking internal resonance to be a naturalistic relation, which is Railton's idea. If this assumption holds, the simplified analysis is a version of Mooreanism that analyses goodness-for in terms of final value and a naturalistic condition. Something would also have to be said about degrees of goodness-for and how the Moorean analysis accounts for them. For simplicity's sake, I shall assume that while meeting the resonance/dissonance condition is necessary for something's being good/bad for a person, *how* good or bad for a person something is, is determined entirely by its final goodness or badness. Degree of resonance or dissonance is thus not relevant to degree of goodness-for. This is very likely an oversimplification, but to repeat, I do not mean for the analysis to be ultimately defensible. I intend to use it merely as a kind of proxy, to investigate whether Mooreanism can respond to Rønnow-Rasmussen's purported counterexamples.

4. Rønnow-Rasmussen's Counterexamples to Mooreanism

The order in which I consider the purported examples of Mooreanism does not follow Rønnow-Rasmussen's. I begin with the two I believe can be accommodated by Mooreanism, and take the more troublesome example last.

4.1 Counterexample I: Final Goodness and Goodness-for: Overall and *Pro Tanto*

Just as something can be finally good (or bad) *overall*, or *all things considered*, it can be finally good (bad) *pro tanto*, or *in some respect*. When we evaluate a state of affairs, such as the current state of the world, we can focus either on its final goodness (badness) overall or on its final goodness (badness) *pro tanto*. The same distinction applies to evaluations in terms of goodness-for (and badness-for); when we evaluate whether something is good (bad) for a person, we can focus either on its overall goodness (badness) for the person, or on its *pro tanto* goodness (badness) for the person. Now consider the following claim:

(I) The world is overall good, but it is overall bad for *a*.[17]

[17] Rønnow-Rasmussen 2021: 40.

This looks problematic to Mooreanism, since it appears that Mooreans must take the world's being overall bad for *a* to consist partly in its being overall finally bad, whereas claim (I) implies that the world is overall finally good! Given our (over)simplified version of Mooreanism about goodness-for, stated in the previous subsection, I suggest that Mooreans can offer the following interpretation of claim (I): The world in its entirety is all things considered finally good (rather than bad or neutral) but the proper parts of it that meet with resonance or dissonance in *a*'s attitudes are all things considered finally bad (rather than good or neutral), making the worlds in its entirety overall bad for *a*.[18]

This is a Moorean rendering of what claim (I) amounts to that seems to me fully intelligible.[19]

It presupposes that not everything about the world that is finally good (bad) meets the relevant resonance (dissonance) condition. I take this to be an innocuous presupposition, since it is independently plausible that there are many parts of the world that are finally good without being good for a person, *a*, because *a* lacks the relevant attitudes (or would lack the relevant idealized attitudes). Indeed, it is plausible that not everything about the world, but only some of its proper parts, *can* resonate or dissonate with *a*'s attitudes, regardless of their final value. Among things or events that cannot resonate or dissonate with *a*'s attitudes are things and events that *a* cannot be aware of, for example, because they are too remote from *a*, spatially, temporally, or psychologically.

In counter-response, critics of Mooreanism might want to modify claim (I) accordingly

> (I*) The proper parts of the world that meet with resonance or dissonance in a's attitudes are overall finally good, but they are overall bad for a.[20]

While claim (I*) might look more troubling than claim (I), the scenario it expresses is also compatible with our proxy version of Mooreanism. To see this, suppose that the relevant proper part of the world, *L* (*a*'s life, let us say), dissonates with *a*'s attitudes. (In other words, *a* is a gloomy character, discontent with her life.) Suppose also that in *L*, the final goodness outweighs the final badness, rendering *L* all things considered finally good. Since the things or episodes in *L* that are finally good do not resonate with *a*'s attitudes, however, they are not good for *a.* Suppose also that

[18] I shall not try to give ontological precision to my talk of the 'world' and its 'proper parts'. Since it is independently plausible that we can talk of the final value of the world in its entirety and about the final values of some of its proper parts, the onus to give such precision is no more on Mooreans than on pluralists like Rønnow-Rasmussen.

[19] I take my proposed Moorean rendering of claim (I) to be an instance of what Rønnow-Rasmussen calls the 'localization manoeuvre'. Rønnow-Rasmussen discusses and rejects a different instance of the localization manoeuvre that seems to me less plausible (2021: 41-43).

[20] I am indebted to Jens Johansson for this suggestion.

L contains things or episodes that are finally bad. Those things or episodes dissonate with *a*'s attitudes (since they are included in *L*, which, according to our previous assumption, dissonates with *a*'s attitudes). They are consequently bad for *a*. *L* thus contains some things and episodes that are bad for *a* and no things or episodes that are good for *a*. It seems a fair conclusion that *L* is bad for *a*, all things considered, although *L* is finally good, all things considered.

4.2 Counterexample II: The Intuition of Neutrality

The next challenge for Mooreans comes from the field of population axiology. According to the 'intuition of neutrality', there is a positive range of wellbeing, such that adding to a population a person whose level of wellbeing is within that range is morally neutral.[21] Put in terms of value, the intuition of neutrality can be formulated accordingly:

> (II) There is a positive range of wellbeing, such that adding to a population a person whose level of wellbeing is within that range does not increase the final value overall of that population.

In other words, although the person's life has positive wellbeing, and is to that extent good for her, adding her life to a population does not add to the final value overall of the population. Accommodating this intuition seems like a challenge to Mooreans, for recall that according to our (over)simplified version of Mooreanism, something's being good for a person implies that it is also finally good. If we assume that final value is strictly additive, to the effect that the final value of a whole equals the sum of the added final values of its proper parts, it is easy to agree with Rønnow-Rasmussen that '[m]onist[s] will need to be quite inventive … [t]o make proper sense of this intuition'.[22]

However, while some monists will have to struggle to accommodate the intuition of neutrality, Mooreans will not have to stretch their inventiveness far beyond their chief source of inspiration. It is a familiar implication of Moore's doctrine of organic unities that the final value of a whole (on the whole) need *not* equal the sum of the final values of its proper parts.[23] For example, the final value overall of a population need not equal the final values of the individual lives comprising that population. The doctrine of organic unities thus allows that adding to a population a life, or several lives, whose levels of wellbeing are within the neutral range does not increase the overall final value of the population. None of this is in conflict with the claims that such a life is finally good, and that it is good for the person whose life it

[21] Broome 2004: 143-5.

[22] Rønnow-Rasmussen 2021: 44.

[23] Moore 1993 [1903]: 81-5.

is. Moore's doctrine of organic unities thus provides Mooreanism with the principled resources to make proper sense of the intuition of neutrality.

4.3 Counterexample III: The Totality of Good

Now, consider the following claim:

(III) Causing the totality of good for a person is itself the totality of good.[24]

By 'the totality of good', Rønnow-Rasmussen means 'all the goodness there is'.[25] I take it that the kind of goodness in question can be either goodness-for or final goodness, so we can talk about the totality of goodness for a person, just as we can talk about the totality of final goodness.

It is not entirely clear to me what claim (III) amounts to, but here is an interpretation of it that seems inconsistent with Mooreanism: The act of causing all the goodness there is for a person is the sole bearer of final value. This strikes me as a peculiar substantive evaluation. I shall say more about why presently. Setting such questions and concerns aside for a moment, I agree with Rønnow-Rasmussen that '[a] Moorean cannot make sense' of the evaluation expressed in claim (III).'[26] The reason why should be clear. According to our Moorean analysis of goodness-for, something's being good for a person implies its being finally good. So, the act of causing the totality of (or indeed some) good for a person cannot be the *sole* bearer of final value. It must also be that that which is caused—i.e., that which is good for the person in question—is a bearer of final value.

What can Mooreans say in response, given that they accept that claim (III) is a coherent substantive evaluation with which it makes sense to agree or disagree? Recall Rønnow-Rasmussen's methodological principle of substantive neutrality (section 1), according to which analyses of value should not rule out coherent substantive evaluations on purely 'formal' grounds. This sounds fair enough, but it is not clear whether Rønnow-Rasmussen views the principle as a necessary condition of adequacy or merely as a desideratum. Some of what he says (e.g., on p. 26) suggests the former. If that is plausible, Mooreans will have to concede defeat: any plausible analysis of goodness-for must make sense of claims such as (III), but Mooreanism fails to do so.

On the other hand, Rønnow-Rasmussen at one point at least calls the methodological principle of substantive neutrality a 'core desideratum'.[27] If it is merely a desideratum, the failure of Mooreanism to accommodate claim (III) is not

[24] Rønnow-Rasmussen 2021: 43.

[25] Rønnow-Rasmussen 2021: 43.

[26] Rønnow-Rasmussen 2021: 43.

[27] Rønnow-Rasmussen 2021: 37.

sufficient ground for rejection. It seems to me that a moderately holistic approach to value analyses will view the methodological principle as merely a desideratum. Rønnow-Rasmussen claims that a theory of value must meet it in order to be adequate[28], but he also holds that 'whatever position we arrive at regarding the dualist/monist issue, our conclusion is bound to have repercussions … for many substantive views about what is in fact valuable'.[29] If he is right about the latter, as I believe he is, a moderately holistic approach to value analysis seems called for.

As I have indicated, the interpretation of claim (III) that Mooreans cannot make sense of is a rather exotic and contrived evaluation. Why would the act of causing all (why all?) the goodness for a person be the only thing of final value? What about causing only some goodness for a person? What about bringing about, but not *causing*, states that are good for a person or finally good? Moreover, if 'a person' in claim (III) does not pick out a particular person, the question arises why causing the totality of goodness for a random person is the only bearer of final value; why would it not be finally good to cause (some) goodness for other people too? If 'a person' in claim (III) means a particular person—Toni, say—the claim seems absurd. Although Toni is a very nice person and a highly supportive supervisor, it is far from plausible that causing the totality of goodness for him is the only thing of final value.[30]

Therefore, Mooreanism's failure to accommodate claim III does not seem like a great cost. Matters had been different, had Mooreanism been forced to renounce as incoherent or false a substantive evaluative view that is intuitively compelling, or at least generally recognised as such. Claim (III) does not fall into that category.

Conclusion

For all I have said, Mooreanism may in the end be less plausible than pluralism. But establishing that it is requires us to look beyond Rønnow-Rasmussen's three purported counterexamples. Mooreanism has the resources to respond adequately to two of them. It might not be able to accommodate the third one, but I conclude that this is not a decisive objection.[31]

---

[28] Rønnow-Rasmussen 2021: 37.

[29] Rønnow-Rasmussen 2021: 31.

[30] Francesco Orsi suggested that (III) might be a less exotic evaluation if the person in question is God. But the view that causing the totality of good for God is itself the totality of God would seem to rule out God's own goodness. Moreover, since it seems exotic to propose that one can cause any changes in God's condition, the proposal that one can cause *the totality of good* for God seems highly exotic.

[31] I am grateful to Krister Bykvist, Jens Johansson, Francesco Orsi, and Caj Strandberg for their helpful comments on earlier versions. A grant from the Research Council (grant no 2019-02-828) is gratefully acknowledged.

References

Broome, John (2004) *Weighing Lives*. Oxford: Oxford University Press.

Fletcher, Guy (2012) 'The Locative Analysis of *Good For* Formulated and Defended', *Journal of Ethics and Social Philosophy* 6: 1-27.

Hurka, Thomas (2021) 'Against "Good For"/"Well-Being", For "Simply Good"', *Philosophical Quarterly* 71: 803-22.

Moore, G. E. (1993 [1903]) *Principia Ethica*, rev. edn., ed. T. Baldwin. Cambridge: Cambridge University Press.

Railton, Peter (2003) *Facts, Values, and Norms: Essays Toward a Morality of Consequence*. Cambridge: Cambridge University Press.

Rønnow-Rasmussen, Toni (2015) 'Intrinsic and Extrinsic Value', in I. Hirose & J. Olson (eds.) *The Oxford Handbook of Value Theory*. New York: Oxford University Press:

Rønnow-Rasmussen, Toni (2021) *The Value Gap*. Oxford: Oxford University Press.

Happy Egrets Strike Back?

Francesco Orsi[1]

1. Introduction

The fitting attitude account of value (FA) claims that to be good or bad is to be a fitting target of a pro-attitude or a fitting target of a con-attitude. Toni Rønnow-Rasmussen has recently defended FA from a new version of what is variously called the solitary goods objection (Bykvist 2009), the wrong kind of value problem (Reisner 2015), or the too little value problem (Rowland 2019, chapter 7): [2] there seem to be objects or states of affairs which are good (or bad), but it is not the case that it is fitting for anyone to favour (or disfavour) them. If the objection is correct, then, contrary to what FA holds, for x to be good or bad cannot be for x to be a fitting target of a pro-attitude or a fitting target of a con-attitude. In this contribution I argue that advocates of FA have a better reply to give to the new version of the solitary goods objection than Rønnow-Rasmussen's somewhat defeatist defence. (For the record, I say this as someone who has often been on the side of those who are sceptical about FA, see Orsi & Garcia 2021, and explored alternatives to it, see Orsi 2013b. But I do find the solitary goods objection to FA unconvincing, as I did in Orsi 2013a.)

A typical example of the solitary goods objection asks us to consider a state of affairs that seems good, but which by its very nature implies that no one is in a position to have a fitting attitude towards it:

> *Happy Egrets*: there being happy egrets but no past, present or future agents (Bykvist 2009: 5).

[1] This is a nod to "The Strike of the Demon" (Rabinowicz and Rønnow-Rasmussen 2004), which contains a section titled "The Demon Strikes Back" (pp. 419 ff.).

[2] Dancy (2000) should be credited for first stating the problem. Bykvist (2015) replies to Orsi (2013a).

No one *in the same world* where *Happy Egrets* obtains was, is, or will be in a position to favour *Happy Egrets*. One might conclude that FA is false, because *Happy Egrets* is (or at least can be) clearly good even if it is not fitting for anyone to favour it.

A natural reaction is to say that at least it is fitting for us, contemplating *Happy Egrets* from the actual world, to favour it, for example by taking contemplative pleasure in it (Orsi 2013a). However, Kent Hurtig (2019) has recently argued that this kind of reply may not always work. In particular, when a state of affairs akin to *Happy Egrets* is indexed to the actual world, it cannot be the case that it is fitting for subjects in a *non-actual world* to favour or disfavour that actual state of affairs, with the result that such states of affairs may be good (or bad), without anyone's attitudes being fitting towards them.

In what follows I first articulate Hurtig's argument—making it more precise, if possible, than Hurtig himself does. Then I discuss Rønnow-Rasmussen's response to it and show why it is somewhat defeatist. Finally, I provide a response to Hurtig that illustrates a broader point about why arguments from solitary goods against FA are doomed to fail: if—due to their location in modal space—the relevant states of affairs cannot even be evaluated as good or bad (and a fortiori favoured or disfavoured) by readers, then such cases are dialectically powerless against FA; if, on the other hand, they can be evaluated as good or bad—despite their location in modal space—then they can also be favoured or disfavoured, and it will be fitting for us (if for no one else) to favour or disfavour them, thus defusing the challenge.

2. World-specific Values

Hurtig argues that FA fails to account for the value of a state of affairs such as this:

> *S*: [*p* is a significant true proposition, and no one in the actual world at any time has any attitude toward *p*] (3245)[3]

Hurtig suggests that *S*'s actually obtaining is bad for its own sake, presumably because if *p* is a significant true proposition, it would be good to know that *p*, and a fortiori it would be good that someone in the actual world had some attitude toward *p*.[4] In order to account for the value of *S*, FA must find a suitable truth or fact about fitting attitudes towards *S*, for example, the fact that it is fitting to disfavour *S* for its own sake. What are the available candidates?

[3] All page-only references are to Hurtig (2019). I have explicitly included 'is a significant true proposition' to Hurtig's own formulation, because he himself describes *p* as a significant true proposition (3245).

[4] I say "presumably" because Hurtig himself appears to just stipulate the badness of *S*.

Since it is a feature of *S* that no one in the actual world has any attitude towards *p*, and *S* includes *p*, it follows that no one in the actual world can have an attitude of disfavour towards *S*, because if they did, then they would have an attitude towards *p* as well, be that as non-committal an attitude as merely entertaining *p*. In other words: there is no coherent scenario where a subject is both part of the actual world *and* disfavours *S*. And if there is no such coherent scenario, then it is not possible to actually disfavour *S*. On the assumption that fittingness implies can, it follows that *actually* disfavouring *S* cannot be the fitting attitude towards *S*.[5]

The natural alternative is to say that it is fitting for a *non-actual* subject, i.e. for a subject existing elsewhere than in the actual world, to disfavour the actual state of affairs *S*. Given the content of *S,* only a non-actual subject could have some attitude towards *p*, and thus towards *S*. This is analogous to the move I suggested (Orsi 2013a) in response to *Happy Egrets*. Since in the world of *Happy Egrets* it is not fitting for anyone to favour *Happy Egrets*, FA can only locate fitting responses to *Happy Egrets* in a world where *Happy Egrets* does not obtain, for example, in our own world. Hurtig's case would seem to be the reverse of that: as he writes, "the evaluating—if there is to be any at all—*has* to take place from a non-actual world" (3247, his italics). In both cases, it seems that FA will resort to what has been called trans-world fittingness (Reisner 2015).

Hurtig, however, rejects the idea that it can be fitting for a *non-actual* (i.e. counterfactual) subject to disfavour the *actual* state of affairs *S*. Here is a reconstruction of his argument (3247):

> P1. If it is fitting for a non-actual subject N to disfavour the actual state of affairs *S*, then N must be able to disfavour the actual state of affairs *S*.
>
> P2. In order for N to be able to disfavour the actual state of affairs *S*, N's evaluation must be able to uniquely be about *S*'s obtaining in the actual world.
>
> P3. There is no causal link between the actual world and N's world.
>
> P4. If there is no causal link between the actual world and N's world, then N's evaluation cannot uniquely be about *S*'s obtaining in the actual world.
>
> C1. Therefore, N's evaluation cannot uniquely be about *S*'s obtaining in the actual world.
>
> C2. Therefore, N is not able to disfavour the actual state of affairs *S*.
>
> C3. Therefore, it is not fitting for a non-actual subject N to disfavour the actual state of affairs *S*.

[5] A reviewer pointed out that, e.g., A. C. Ewing did not accept "fittingness implies can". But the "can" in this case is one of logical possibility.

And of course, if *S* is bad for its own sake, but it is not fitting for N (or any other actual or non-actual subject) to disfavour it, then FA is false.

P1 is an application of the idea that "normativity implies can" (3243, 3248). Premises P2 to P4 are Hurtig's paraphrases of passages in Brogaard and Salerno (2019), where the latter cast doubt on the possibility of counterfactual knowledge of actual truths.[6] It seems that their doubts, if sound, would carry over to the case of counterfactual evaluation and favouring of actual states of affairs that are impossible actually to evaluate, such as *S*. Due to lack of a causal link, counterfactual evaluators have no way of "latching onto" the actual world as opposed to any other world in which *S* obtains, and thus have no way of latching onto the value of *S as it obtains in the actual world* (3247). I will now discuss Rønnow-Rasmussen's response.

3. Rønnow-Rasmussen's Response

Rønnow-Rasmussen responds to Hurtig by essentially conceding to the challenge: "Like Hurtig, I believe it is impossible to have an attitude in the evaluating world that latches onto the actual world" (2022: 116). However, he goes on to explain why "the problem is not quite as serious as it appears to be" (ibid.). He distinguishes two scenarios: (a) non-universalizable features of a state of affairs are not value-makers; (b) non-universalizable features are or can be value-makers.

In the first scenario, then, the value of a state of affairs like *S* depends only on its universalizable value-making features, and therefore not on non-universalizable features like the identity of the individuals involved (say, "Charlie") or—as I understand Rønnow-Rasmussen—even the particular modal location of the state of affairs. If so, then "a proponent of FA analysis ought to be quite satisfied with the counterfactual evaluator evaluating a class of Charlies (i.e. those individuals in possible worlds that share Charlie's universal value-making features) even if his attitude does not 'latch on' to the Charlie in the actual world" (ibid.: 116-117). In the case of state *S*, then, the counterfactual evaluator can still evaluate a class of states of affairs that share *S*'s universal value-making features, even if her evaluating does not latch onto *S* as belonging to the actual world. The assumption made by this reply to Hurtig is that evaluating the class of such states of affairs does not itself require latching onto any particular world or individuals, or (practically equivalently) that the 'latching onto' required in this case is possible for any

[6] "If there is such non-actual knowledge, there is non-actual thought about an actual situation. So the non-actual thinker somehow has a concept of an actual situation. But how is it possible for a non-actual thinker to have a concept that is specifically about situations in this the actual world. It will not do for the thinker to express the thought 'actually *p*', since 'actually' will designate rigidly only situations in her own world. Moreover, since there is no causal link between the actual world w1 and the relevant non-actual world w2, it is unclear how non-actual thought in w2 can be uniquely about w1" (Brogaard & Salerno 2019, in turn referring to Williamson (1987)).

evaluator whatever their location in modal space. (I'll grant this assumption in what follows.)

I believe that this response concedes a much larger defeat for FA than Rønnow-Rasmussen supposes. If Hurtig is right, then *whenever* the allegedly required latching onto a certain world or individual does not or cannot take place, it will not be fitting to favour exactly *that* valuable state of affairs (or the individuals therein), but only the class of states of affairs sharing the universalizable features. But the required latching onto can fail in a myriad of cases, even when the evaluating world and the evaluated world coincide. It is a highly contingent matter whether anyone's attitudes do or do not latch onto a given states of affairs. In turn, it will be a highly contingent matter whether anyone can favour that particular state of affairs, and thus whether it is fitting to favour that particular states of affairs.

This predicament puts FA before two unpleasant alternatives. The first alternative is to hold a disjunctive account of the objects of fitting attitudes: a state of affairs P is good if and only if either it is fitting to favour P or, failing that, it is fitting to favour something like P's better relative P*: that is, P minus any non-universalizable feature which would require the evaluator's latching onto exactly P's world or P's individuals. But this account is not a great solution: ideally, we would like the objects of fitting attitudes (*what* it is fitting to favour) to be, always, exactly the same—i.e. the same tokens under the same description—as the objects bearing the value property (*what* is good). If Charlie's being happy is good, then it is Charlie's being happy that should be favoured, and not, even as a second best, simply the state [someone's being happy]. Even if one agrees that the value of a state of affairs depends only on a state of affairs' universalizable features, one may still require that the object of the fitting attitude be the state of affairs including its non-universalizable features, because that is after all how the relevant value bearer is presented in this case (e.g. as Charlie's being happy). One thing is the question about legitimate value-makers, another thing is the question about what are the legitimate targets of fitting attitudes.

The second unpleasant alternative is to stipulate that FA only account for the value of value bearers stripped of any non-universalizable feature. On this view, FA should account neither for the value of Charlie's being happy, nor for the value of someone's being happy in the actual world, but only for the value of someone's being happy. Of course FA advocates are free to select the subject matter of FA as they please. But it seems to me FA would lose some of its appeal. After all, we ordinarily ascribe value to states of affairs that include non-universalizable features such as the identity of individuals, times, places. Moreover, we may be tempted to hold the view that, for example, "agent *a*'s pleasure is valuable, but […] no other agent's pleasure is valuable, however similar it is in terms of its universalizable features" (Rønnow-Rasmussen 2022: 117). FA, then, had better find a way to also cover the value of value bearers with non-universalizable aspects rather than ignore them.

This point is particularly pressing when we consider Rønnow-Rasmussen's second scenario: non-universalizable features are or can be value-makers of states like Hurtig's *S* or Charlie's being happy in the actual world. As in the example above, one could hold that only a certain agent's pleasure is valuable, regardless of similarities with other agents, thus making the value dependent, in part, on who the agent is. In fact, one could hold that God's pleasure is the highest good, or that beatific vision of God is the highest good, as distinct from, say, "the pleasure of perfect beings is the highest good" and "beatific vision of perfect beings is the highest good". In these cases it is God's identity that matters, over and above his perfection or other universalizable features he (and maybe only he) possesses. Whether these substantive axiologies are plausible or not, it would be a significant cost for FA to decide to leave them outside of its sphere of analysis.

In this connection, Rønnow-Rasmussen notes that "*any* view suggesting that non-universalizable features can be value-making features will owe us an explanation" (ibid.). This is true. But the special problem for FA is that FA faces the extra burden of making sense of the object of fitting attitudes in these cases, *if*, as Rønnow-Rasmussen appears to concede to Hurtig, having fitting attitudes towards these particular states of affairs requires the possibility of referring or latching onto particular worlds or individuals, and this possibility may not always be given to the relevant evaluator. In this sense, *if Hurtig is right*, then it is probably better for FA to altogether give up on accounting for the value of value bearers with non-universalizable features (despite the cost of this move), and focus on finessing the first alternative above to make it more digestible.[7] However, it will be clearly even better for FA if one can reject Hurtig's argument in the first place. This is what I do in the next section.

4. A Different Response: Modal Relocation

The first thing to note is that Hurtig's argument, as it stands, would generalize to cases that, for all Hurtig says, FA *can* account for. As reminded above, FA already needs trans-world fittingness in order to explain the value of states of affairs that are non-actual, like *Happy Egrets*. In this case, all we know is that such a state obtains in some possible, non-actual, world. We do not seem to need any causal link with

[7] I find Rønnow-Rasmussen's first response (a counterfactual evaluator can still have fitting attitudes towards a class of states of affairs, if not towards *S* itself) similar to the response I gave to the 'distance problem' in Orsi 2013a. I argued that 'x is good to degree n' can be defined as 'it is fitting to favour x to degree n from behind a veil of ignorance regarding the evaluator's distance (personal, temporal, modal, even epistemic) from x'. However, I am not sure that all factors to be 'veiled' necessarily match non-universalizable features of the state of affairs—they are rather facts about me than facts about the state of affairs. The question also remains whether, after veiling all these factors, we still need to be able to latch onto x in order to evaluate *it* and not something else.

any of the possible worlds where *Happy Egrets* obtains in order to contemplate and favour it. But if lack of a causal link is a challenge for FA in the case of *S*, then it should be so also in the case of *Happy Egrets*. Since it doesn't seem to be a challenge in the latter case—and Hurtig appears to agree (3245)—then it is down to Hurtig to explain why it is a challenge in the case of *S*.

Hurtig is likely to answer that *S* is different from other solitary goods or evils in that *S* is world-specific (3246). In fact, in the article there is a crucial shift in the content of the relevant state of affairs from simply

> *S*: [*p* is a significant true proposition, and no one in the actual world at any time has any attitude toward *p*]

to what I will call

> *Actual S*: [*S* obtains in @]—where '@' designates the actual world (3246-7).

Hurtig does not seem to appreciate that *Actual S* is different from *S*. *Actual S* is a state of affairs *indexed* to a particular location, namely the actual world, while *S* is not. *S* only says something *about* the actual world, namely that in this world no agents have attitudes at any time towards *p*. But *S* itself could in principle be indexed to a different world. So, to make Hurtig's argument work, we should now insert '*Actual S*' in place of 'the actual state of affairs *S*'.[8]

So *Actual S* (rather than just *S*) is supposed to be importantly different from *Happy Egrets*. Since the latter is not indexed to any specific modal location, an evaluator's latching onto the possible worlds where it obtains comes on the cheap, i.e. without the need for any causal link. (Or perhaps there is no need for latching onto them in the first place in order to have the relevant fitting attitude.) But when a state of affairs is indexed to the actual world, like *Actual S*, the thought must be that a causal link between evaluating world and evaluated world is required, so that the evaluator is able to pick exactly the right location of the valuable state of affairs. Were she to pick a state exactly like *Actual S*, however located in a world that is not the actual world, she would make a mistake and end up evaluating not *Actual S* but a different state of affairs. In the absence of a causal link, then, a counterfactual evaluator is not able to evaluate *Actual S*, hence it cannot be fitting for her to disfavour *Actual S*.

I will now provide a response to this modified version of Hurtig's argument. The starting point is that Hurtig does not sufficiently explain whether there is supposed to be a special problem for non-actual evaluators to uniquely pick the actual world among other similar worlds, or whether there is also a parallel problem for actual evaluators to uniquely pick a specific non-actual world W1, when a valuable state

[8] In fairness to Hurtig, by the time he presents the argument reconstructed above, the shift to *Actual S* has already occurred. But he doesn't register the shift.

of affairs is indexed to W1 (think *Happy Egrets in W1*). But the latter claim seems more plausible.[9] In other words, it seems that if a causal link is required for a counterfactual evaluator to have an attitude towards a world-indexed actual state of affairs, then by parity of reasoning a causal link must also be required for an actual evaluator to have an attitude towards a world-indexed non-actual state of affairs. And if this is true, then it follows that in the absence of a causal link with the relevant non-actual world we, as actual evaluators, would not be able to favour or disfavour a world-indexed non-actual state of affairs, because we would not be able to uniquely pick this one among other, similar, states of affairs occurring in other worlds.

However, the latter implication should give us pause. As I noted in (Orsi 2013a), whenever a seemingly good or bad state of affairs is put up for consideration as a counterexample to FA, it must at least be fitting *for the readers* to regard it as good or bad. If it is not even fitting for readers to evaluate it, then it can hardly be eligible as a counterexample to FA. But if it is fitting for readers to evaluate it, then it must be possible to evaluate it. And if we, as readers, can evaluate it, then this means that our ability to evaluate such states of affairs holds regardless of our location in modal space vis-à-vis the location of the state of affairs. Where we stand with relation to the state of affairs doesn't seem to matter. A fortiori, the absence of a relevant causal link between the reader's world and the world where the state of affairs obtains is neither here nor there. Since there does not seem to be any additional challenge in going from evaluating a state as good or bad to favouring or disfavouring it, it also follows that we, as readers, can have the relevant fitting attitude. In other words, it is tempting to suggest that the whole literature on solitary goods must be premised on the assumption that it is at least possible, and fitting for someone, namely the reader, to favour or disfavour the putative solitary good or evil in some way.[10]

World-specific or indexed solitary goods and evils are no exception. In fact, Hurtig concurs, as he writes that "it is coherent to think that *S*'s obtaining *in the actual world* [i.e. *Actual S*] is bad for its own sake" (3245, his italics). Now, this needs to be refined, since it is not coherent for us to *both* regard ourselves as actual evaluators of *Actual S* and think that *Actual S* is bad. By hypothesis, *Actual S* cannot have *actual* evaluators. What is coherent, instead, is for the reader to think that *Actual S* is bad while regarding herself as non-actual—placing herself in a non-actual world. This sort of modal relocation must be possible for us, or else it is not clear *for whom* it is coherent to think that *Actual S* is bad for its own sake. And here is the catch: if it is possible for us, readers, to place ourselves in a non-actual world and evaluate *Actual S from there*, without there being any apparent causal link between the non-actual world we would inhabit and the actual world where *Actual*

[9] It seems that Hurtig would agree, as he writes that the challenge for FA is, in general, to show "*how* it is possible to favour *specific* worlds, situations, or states of affairs" (3248, his italics).

[10] This point applies also to Reisner's "causal entanglement" case (2015). See Rowland (2019, ch. 7) for a detailed response to Reisner's arguments.

S obtains, then it must be possible for *any* counterfactual evaluator to evaluate and disfavour *Actual S*, without the need for any causal link between *their* non-actual world and the actual world. *Qua* placed in a non-actual world, we are in a no more privileged position with respect to *Actual S* than any other non-actual evaluator of *Actual S*. (It's not as if by virtue of, in fact, inhabiting the actual world, we can somehow smuggle our way to *Actual S*. When we evaluate *Actual S*, we stand firmly in a non-actual world.) Hence either one should reject premise P4 in Hurtig's argument (if there is no causal link between the actual world and N's world, then N's evaluation cannot uniquely be about *S*'s obtaining in the actual world), or Hurtig must accept that *Actual S* is a state of affairs not even his readers can coherently evaluate.

Hurtig may want to buy into the second horn of the dilemma. He might suggest that putting up *Actual S* for consideration as a legitimate counterexample to FA does not require the readers' ability to evaluate *Actual S*. It only requires the ability to contemplate its general features, namely *Actual S* minus the world-specific index. In other words, Hurtig may claim to be entitled to present *Actual S* as a counterexample, even if *Actual S* is a state of affairs that (by Hurtig's own lights) *as such* we cannot properly grasp and evaluate, since we—forced by the nature of *Actual S* to take the position of non-actual evaluators—cannot uniquely pick *Actual S* from other, similar states of affairs. It is not fitting for us to think that *Actual S* is bad, yet we are to somehow take it that *Actual S* is bad.

It is not clear whether Hurtig would endorse this reply line, as this involves the same mismatch between value bearer (here, *Actual S*) and object of the fitting attitude (*Actual S* minus the world-specific index) pointed out in my reply to Rønnow-Rasmussen above. Of course, since Hurtig is arguing *against* FA, such a mismatch need not be a problem for *him*. But this reply still involves two hefty commitments that plausibly *everyone* should steer clear of.

First, if, despite our inability to uniquely pick it from similar states of affairs, *Actual S* works as a counterexample to FA, then it follows that there are, or could be, good or bad states of affairs that, by their very nature, are not *as such* graspable by any subject in any world, and a fortiori it cannot be fitting for anyone to evaluate them, let alone favour or disfavour them. The best we can do is relate to similar states of affairs that do not include a problematic index to a world we have no causal link with. I will not discuss whether such a view is coherent. Arguably it is a view that those who think that normativity is optional to value might be happy to endorse: there are valuable states of affairs that it is fitting for no one to even evaluate—because no one could evaluate them. But at this point we might wonder whether, overall, FA offers a better package than any such view. It is worth remarking that extant theories on the relation between value and normativity do not go so far as to *deny* the rather trivial claim that if something is good, then it is fitting to regard it as good. If a theory can account for the value of *Actual S* only by denying this trivial claim, then such a theory earns a benefit at a very significant cost.

The only way out of this problematic commitment is to deny premise P4 in Hurtig's argument: despite the absence of an appropriate causal link between *Actual S* and the non-actual world into which we "relocate" when contemplating *Actual S*, we are able to evaluate *Actual S* as such, and so are other counterfactual evaluators. That is why it is coherent and fitting for us (more precisely, for the counterfactual "us") to think that *Actual S* is bad. And if it is coherent and fitting for "us" to think that *Actual S* is bad, then it is a short step to it being fitting for "us" to disfavour *Actual S*, as FA has it. Note: on this view, the object of fitting disfavour is indeed *Actual S*, not just the class of states of affairs sharing universalizable features with *Actual S*. Indexing states of affairs to particular modal locations cannot make them completely inaccessible to evaluation and other fitting attitudes. It thus seems that, whatever may be true regarding belief or knowledge, evaluation is a kind of attitude that can tolerate lack of appropriate causal links between evaluating world and evaluated world. The broader implications of this point will need to be explored elsewhere.[11]

The second hefty commitment is not so much one in value theory, but rather in the ethics of argumentation. Suppose Hurtig does endorse the idea that *Actual S* is a valid counterexample to FA, even though we cannot really judge *it* bad. Then he would need to defend the fairness of objecting to a view on the basis of a counterexample that, by its very nature, readers (and author) are unable to properly grasp, but only able to "get somewhere near". Whether such argumentative moves are ever legitimate is a complicated question I cannot address here, but it seems Hurtig has taken upon himself the burden to address it.[12]

References

Brogaard, B. & Salerno, J. (2019). "Fitch's Paradox of Knowability", The Stanford Encyclopedia of Philosophy (Fall 2019 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2019/entries/fitch-paradox/>.

Bykvist, Krister (2009). 'No Good Fit: Why the Fitting Attitude Analysis of Value Fails'. *Mind* 118, 469: pp. 1–30.

[11] A reviewer helpfully pointed out that states like *S* and *Actual S* may be 'axiological blindspots', analogous to epistemic blindspots like 'It is now raining but no one believes that it is now raining'. Epistemic blindspots are true propositions that cannot be truthfully or rationally believed. Axiological blindspots would be states of affairs that are good (bad) but cannot be fittingly (dis)favoured or even evaluated. One question, here, is whether epistemic blindspots can at least be coherently *entertained* by someone—not much more than this would seem to be needed for evaluation.

[12] This research has been supported by the European Union through the European Regional Development Fund (the Centre of Excellence in Estonian Studies, TK 145), and the University of Tartu, grant PHVFI21914. I thank the editors for their invitation to contribute and a reviewer for their insightful comments.

Bykvist, Krister (2015). 'Reply to Orsi'. *Mind* 124, 496: pp. 1201–5.

Dancy, Jonathan (2000). 'Should We Pass the Buck?' In A. O'Hear (ed.), *Philosophy, The Good, the True, and the Beautiful.* Cambridge: The Press Syndicate of the University of Cambridge: pp. 159–73.

Hurtig, Kent (2019). 'The Fitting Attitude Analysis of Value: An Explanatory Challenge'. *Philosophical Studies*, 176, 12: pp. 3241-49.

Orsi, Francesco (2013a). 'Fitting Attitudes and Solitary Goods'. *Mind* 122 (487): 687-698.

Orsi, Francesco (2013b). 'What's Wrong with Moorean Buck-Passing?'. *Philosophical Studies* 164, 3:727-746.

Orsi, F. & Garcia, A. G. (2021). 'The Explanatory Objection to the Fitting Attitude Analysis of Value'. *Philosophical Studies* 178 (4):1207-1221.

Rabinowicz, W. and Rønnow-Rasmussen, T. (2004). 'The Strike of the Demon: On Fitting Attitudes and Value'. *Ethics* 114, 3: pp. 391–423.

Reisner, Andrew (2015). 'Fittingness, Value, and Trans-World Attitudes'. *Philosophical Quarterly* 65, 260: pp. 464–85.

Rønnow-Rasmussen, Toni (2022). *The Value Gap*. Oxford: Oxford University Press.

Rowland, Richard (2019). *The Normative and the Evaluative: The Buck-Passing Account of Value*. Oxford: Oxford University Press.

Williamson, Timothy (1987). 'On the Paradox of Knowability'. *Mind* 96, 382: pp. 256-61.

A Kantian Reading of 'Good' and 'Good For'

Some Reflections on Toni Rønnow-Rasmussen's Fitting Attitude Analysis of Value

Herlinde Pauer-Studer

Abstract. The paper argues that Toni Rønnow-Rasmussen's fitting-attitude analysis of 'good' and 'good for' allows us to interpret and justify Kant's Formula of Humanity (FH) in a constructive way. His classification of 'good' as a non-relational intrinsic final value and 'good for' as a relational extrinsic final value sheds light on two main features of FH, namely that it requires us to display a specific attitude to human beings, while also obligating us to recognize this value in the relational dimension. Based on a reflection of what attitudes toward persons are fitting, we might well come to endorse that persons are "ends in themselves" and merit respect and recognition. I then argue (by way of an ethical reading of Kant's demand to leave the state of nature and move to a rightful civil condition) that we have, in addition to a fitting attitude, deontic normative reasons (not mere pro tanto reasons) for making this very attitude toward persons the principal standard for our relations to others and to ourselves.

1. Introduction

A widely accepted assumption is that not everything we consider valuable may depend on our desires and preferences. This is especially true of the basic values that govern our social life, foremost the idea that persons have special value. Kant

articulated this thought in his famous Formula of Humanity (FH), which requires us to treat humanity "whether in your own person or in the person of any other, always at the same time as an end, never merely as a means". In other words, the appropriate attitude to persons is one of respect.

Exactly how this Kantian formula is to be understood remains controversial among moral philosophers to this day. In this paper, I will show that Toni Rønnow-Rasmussen's taxonomy of values, in combination with his fitting attitude analysis of value (FA), offers an attractive way to interpret Kant's requirement. In particular, Rønnow-Rasmussen's distinction between 'good' as a non-relational intrinsic value and 'good for' as a relational extrinsic value enables us to flesh out the full potential of Kant's principle and, moreover, to read Kant's conception of morality in a relational way, thus overcoming the limitations of a first-person only understanding.

The paper is structured as follows. Section 2 outlines the main features of Toni Rønnow-Rasmussen's analysis of value. In section 3, I present Kant's derivation of the Formula of Humanity (FH). Section 4 argues that Christine Korsgaard's attempt to ground this formula in the constitutive conditions of agency is not successful. I will then argue (section 5) that Rønnow-Rasmussen's distinction between a relational and a non-relational reading of value does justice to both aspects of FH, namely that it requires us to display a specific attitude to human beings, while also obligating us to recognize this value in the relational dimension. That persons have value in themselves thus guides our relations to ourselves and to others.

2. Two Kinds of Final Value: 'Good' and 'Good For'

In his impressive book *The Value Gap* (Rønnow-Rasmussen, 2022), Toni Rønnow-Rasmussen proposes a fine-grained taxonomy of values that exceeds the classic opposition between what is valuable for its own sake and what is instrumentally valuable. It is beyond the scope of this paper to do justice to the details of Rønnow-Rasmussen's intricate analysis. In what follows, I will only sketch those aspects that are relevant to our topic, namely how to understand the full scope of the Kantian principle that persons deserve to be valued as "ends in themselves".

The hallmark of Rønnow-Rasmussen's analysis is the distinction between two basic values, 'good' and 'good for'. The meaning and scope of these notions become clearer when we examine how Rønnow-Rasmussen specifies them in terms of other concepts commonly used in value theory, such as final and non-final, intrinsic and extrinsic, relational and non-relational.

Fundamental for Rønnow-Rasmussen are the categories final and non-final value (Rønnow-Rasmussen, 2022, 17). 'Final value' refers to objects that are valuable for their own sake. 'Good' and 'good for' might both refer to final values, according to Rønnow-Rasmussen. He endorses value dualism, according to which *final goodness* and *final goodness for* are both coherent value notions, though semantically distinct

("they cannot be fully understood in terms of one another", Rønnow-Rasmussen, 2022, 35).

'Good for' might, as Rønnow-Rasmussen points out, refer to an instrumental or to a non-instrumental value. That something, x, is good *for* someone therefore either means that x is instrumentally valuable or it means that x is valuable for its own sake (yet valuable for someone). In the latter case, 'good for' is tied to a final value. An instrumental value is, as Rønnow-Rasmussen argues, a non-final value and not valuable for its own sake, but as a mere means.

Rønnow-Rasmussen also explains 'good' and 'good for' in terms of the distinction between intrinsic and extrinsic values; intrinsic and extrinsic are "subcategories within the class of final values" (Rønnow-Rasmussen, 2022, 17). Something has intrinsic value when it is valuable in terms of the bearer's internal features alone; things do have extrinsic value when they are valuable because of the bearer's external relational properties. 'Good-for' presents, according to Rønnow-Rasmussen, a relational value, while 'good' is a non-relational value. Moreover, 'good' amounts to an impersonal value (i.e. good period), while 'good for' is a personal value (good for someone), not least because of its extrinsic relational aspects.

To summarize the main points of this taxonomy: The set of final values, which includes what is valuable for its own sake, contains two kinds of value, namely 'good' and 'good for'. 'Good' is a final intrinsic value insofar as it is valuable for its own sake in virtue of the bearer's internal features alone, and it is a non-relational value. 'Good for', on the other hand, is a final extrinsic and relational value, given that it refers to things that are valuable for their own sake in virtue of "some of [the bearer's] external relational properties" (Rønnow-Rasmussen, 2022, 17).

Rønnow-Rasmussen's overall aim is to provide a fitting-attitude analysis (FA) of 'good' and 'good for'. 'Fittingness' refers to the relation between an object and a response; relevant is whether an 'object' merits a certain response or is worthy of that response. A fitting-attitude analysis must take into account that the attitudes fit the different 'objects' at stake. Persons, for instance, merit respect or admiration (Rønnow-Rasmussen, 2022, 15).

An attractive feature of a fitting-attitude analysis is, according to Rønnow-Rasmusssen, that it "connects our attitudes with valuable objects in a straightforward way" (Rønnow-Rasmussen, 2022, 5). We do not need to appeal to yet another value in explaining why something is valuable.

However, as Rønnow-Rasmussen emphasizes, the fittingness relation by itself does not provide a full account of values. Attitudes as such are not value-constitutive; they just indicate that something is worthy of being valued. According to Rønnow-Rasmussen, the missing element for a full account of FA is the notion of a reason. As he writes:

> A person is valuable, according to the FA analysis, because there is something about this person that provides us with a reason to respect him or love him. The connection between value and attitudes is intermediate; you need to add reasons to the picture—only then do we have value (Rønnow-Rasmussen, 2022, 5).

For Rønnow-Rasmussen, it is the object that provides us with a reason to favour it (Rønnow-Rasmussen, 2022, 113). He thus understands 'reason' in a contributory sense. The reasons needed for a fitting-attitude analysis of value are, as Rønnow-Rasmussen states, pro tanto reasons, because it is important that the reasons that speak in favour of an object being valuable may be outweighed by other pro tanto reasons (Rønnow-Rasmussen, 114). This suggests that thinking about what is valuable is for him a matter of weighing considerations.

Rønnow-Rasmussen suggests the following principle as central to an FA analysis:

> FA1: For something to be valuable is for x to be (or provide) a reason for an agent who is rightly placed to favour x (Rønnow-Rasmussen, 2022, 114).

And on the basis of this principle, he then proposes the following analysis of the final value 'good for':

> For *x to be* finally good is for x to be, or provide, a reason for any agent who is rightly placed to favour *x* for its own sake *where this favouring is not for the sake of someone or something other than x* (Rønnow-Rasmussen, 2022, 137, italics in the original).

> For *x* to be finally good for *a* is for *b* to be, or provide, a rightly placed *b* with a reason to finally favour *x* for *a*'s sake (where *a* and *b* might, but need not be, identical) (Rønnow-Rasmussen, 2022, 137, italics in the original).

All three formulations for defining the final values 'good' and 'good for' involve agents. Note, however, that Rønnow-Rasmussen restricts the role of agents to having a (pro tanto) reason to favour x.

This will not suffice when it comes to explaining the value of persons. The fitting attitude toward persons cannot be understood in terms of a mere favouring relation. Rather, the appropriate attitude toward persons is that of respect and recognition. Moreover, it is one we ought to have toward others. Since pro tanto reasons may be outweighed by other pro tanto reasons, they do not conceptually correspond to an 'ought' in the deontic sense.[1] Instead of pro tanto reasons that provide us merely with considerations in favour of valuing x, we need to rely on a special kind of normative reasons that reflect the deontic character of the fitting attitude when it comes to explaining the value of persons as persons. I will use the term 'deontic reasons' for this special kind of normative reasons.[2]

[1] Following Korsgaard, we might say that pro tanto reasons are tied to a weighing model of practical reason.

For a critique of such an understanding of practical reason see (Korsgaard, 2009b, 49-51).

[2] In his book *Personal Value* (2011), Rønnow-Rasmussen speaks of normative reasons in connection with favouring. He explains, for instance, the final value of an object x in terms of normative reasons favouring x for its own sake and personal value in terms of normative reasons for favouring x for a person's sake (Rønnow-Rasmussen 2011, 27, 48, 78, 90). This means that for Rønnow-Rasmussen pro tanto reasons are normative reasons in so far as they are considerations in favour of something, for

However, I think that Rønnow-Rasmussen's account allows us to make room for normative reasons (deontic ones) that provide a warrant for obligatory fitting attitudes.

In the next two sections, I will first outline Kant's derivation of the Formula of Humanity (FH) (section 3) and then Korsgaard's reformulation of Kant's argument (section 4). Both attempts run into serious problems. Subsequently (section 5), I try to show how Rønnow-Rasmussen's fitting-attitude analysis of value (in a slightly modified form) offers an illuminating interpretation of Kant's Formula of Humanity.

3. Kant's Derivation of the Formula of Humanity (FH)

Kant seeks to justify the categorical imperative in the following way: first, he outlines the conditions the highest principle of morality has to meet, then arguing that the categorical imperative (particularly in the Formula of Law formulation and the Formula of Humanity formulation) fulfills those conditions. Second, he provides an additional argument (Kant calls it a deduction) for why the categorical imperative is indeed the principle of morality.

We might describe the structure of this complex argument as a two-step procedure: first, to lay bare the conditions for the possibility of x, and, secondly, to provide a justification that provides (deontic) normative reasons for the reflective endorsement of the principle that embraces those conditions.[3]

---

instance, that object x has value. Note, however, that in *Personal Value* Rønnow-Rasmussen does not tie the FA analysis to pro tanto reasons; he relies on a general notion of normative reasons. He considers it as a specific strength of fitting attitude analyses that they do not appeal to a "specific 'reason notion'", but "understand value in terms of a general notion of reason or normativity" (Rønnow-Rasmussen, 2011, 91). In *The Value Gap*, Rønnow-Rasmussen is more specific in tying his FA analysis of value to pro tanto reasons as a specific kind of normative reasons. In order to distinguish my understanding of 'normative reasons' from Rønnow-Rasmussen's use of the term 'normative reasons', I will speak of 'special normative reasons' or 'deontic normative reasons'. The difference between a pro tanto reason and a deontic reason is that the former amounts to a consideration in favour of something and considers deliberation as a weighing of those considerations, whereas the latter provides a more robust warrant since a deontic normative reason for x (e.g., for the acceptance of a principle which attributes special value to something) is backed by an argument that aims to provide a justification for why x (e.g., a principle attributing special value to something) is normatively binding. Deliberation is here more complex, involving a reflective assessment of reasons in light of principles and testing procedures. Let me add that in *Personal Value*, Rønnow-Rasmussen argues that the FA analysis suggested by him seeks to reduce evaluative claims to deontic claims about the attitudes that are fitting and that, therefore, one ought to have (Rønnow-Rasmussen, 2011, chapter 2). In my view, his account misses the deontic level. My approach in this paper, which relies on deontic normative reasons, can be seen as an attempt to do justice to this deontic element in the FA analysis (my discussion is restricted to the attitude we owe human beings). I thank two anonymous referees for pressing me to clarify my understanding of 'normative reasons'.

[3] My understanding of a transcendental argument is that such an argument consists not merely in the exposition of necessary presuppositions of x (i.e. the conditions of the possibility of x), but also in an additional justification why the principle that meets those conditions is justified. In the process of laying bare the presuppositions we gain some insight why those presuppositions are indispensable and

Kant's reasoning in the *Groundwork* leading to FH follows that line. The first step, namely exposing the categorical imperative as a condition for the possibility of moral reasoning and, thus, of acting morally, is performed by what Kant calls a 'regressive argument'.[4]

With respect to FH, the regressive argument consists of depicting the conditions that underpin our self-valuing as human beings and rational agents. The argument begins by claiming that characteristic of rational beings is the capacity of self-determination by a will.

Kant then states that the ground of the will's self-determination has to be an objective end, not a relative end. The latter would be merely conditionally valuable since its worth would depend on certain subjective desires and incentives. Relative ends are "only the ground of hypothetical imperatives" (GMS, AA 04: 428.02).[5] But the ground of a categorical imperative as a practical law must be something that, as an end in itself, has absolute worth—for "if all worth were conditional and therefore contingent, then no supreme practical principle for reason could be found anywhere" (GMS, AA 04: 428.31-33).

Kant concludes by maintaining that all rational and human beings exist as ends in themselves and, thus, possess unconditional—i.e., absolute—worth. This is so since human beings can never have merely conditional value and can never serve merely as means. Kant adds that all human beings must see themselves in this way, given that we all share the rational basis for this kind of self-understanding (GMS, AA 04:429.05-07).

Given "the representation of what is necessarily an end for everyone because it is an end in itself", the sought principle has to meet the following conditions: it must hold categorically, it must be "an objective principle of the will", and it must be able to "serve as a universal practical law" (GMS, AA 04: 428.36-37, 04: 429.01-02). Kant concludes that the categorical imperative maintains: "*So act that you use humanity, whether in your own person or in the person of any other, always at the same time as an end, never merely as a means*" GMS, AA 04: 429.10-12, italics in the original).

---

hence why their assumption seems justified; however we need an additional argument for why the principle embracing those presuppositions seems justified. A detailed discussion of transcendental arguments is beyond the scope of this paper.

[4] The most transparent form of a regressive argument is, in my view, Kant's reasoning in the Groundwork leading to the Formula of Universal Law (FUL). FUL comes up at the end of an analysis intended to identify the principle underlying the good will. Kant proceeds by exposing the conditions such a principle has to meet (to have the form of a categorical, not a hypothetical, imperative; to be a formal law holding universally). The argument ends by stating that those conditions – formality, universality and categorical bindingness – are exactly met by FUL. Kant's formulates (FUL) as follows: *"(A)ct only in accordance with that maxim through which you can at the same time can will that it become a universal law."* (GMS AA 04:421; italics in the original).

[5] Note: All references to Kant's Groundwork follow the notations of the Academy edition of the Groundwork, reprinted in Kant (1785/1996).

Kant's argument relies on two assumptions. The first is that our capacity for setting ends defines us as human beings. Pursuing ends involves that we, as rational agents making choices, confer value on our ends. The second assumption holds that the rational capacity of ascribing objective value to ends presupposes that one must ascribe objective and unconditional value to oneself and to one's rational will, which amounts to respecting the humanity within us. The source of all value lies in our human and rational nature.[6]

The derivation of FH rests on the self-legislation of the rational will and thus on the principle of autonomy. Kant confirms as much when he states:

> For, nothing can have a worth other than that which the law determines for it. But the law-giving itself, which determines all worth, must for that very reason have a dignity, that is, an unconditional, incomparable worth; and the word respect alone provides a becoming expression for the estimate of it that a rational being must give. Autonomy is therefore the ground of the dignity of human nature and of every rational nature (GMS, AA 04: 436.01-07).

The importance of autonomy for grounding FH is also made apparent by Kant's caveat that his regressive exposition of the conditions leading to FH merely amounts to a conditional justification. Kant's derivation relies, moreover, on the assumption that "rational nature exists as an end in itself", which, as "the ground of this principle", i.e. FH, holds equally for all other rational agents (GMS, AA 04: 429.05-07).

In an annotation to the text, Kant tells us that he introduces this assumption (rational nature exists as an end in itself) as a mere postulate, and that he is going to provide its full justification, a "deduction", in section 3 of the Groundwork. The argument offered there proceeds from the claim that a rational will is an autonomous will, and that an autonomous will and a moral will are one and the same, to the conclusion that a rational will is a moral will.

Kant himself was uneasy with his argument; he suspected it to be circular, simply presupposing the autonomy of the will without further argument. That he indeed

---

[6] Allen Wood reconstructs Kant's argument for the derivation of FH in the following way:

(1) This (that rational nature exists as an end in itself) is how the human being necessarily represents his own existence; to this extent, therefore, it is a subjective principle of human actions.

(2) But every other rational being also represents its existence consequent to precisely the same rational ground which is valid for me;

(3) Therefore, it is at the same time the rational ground of an objective principle, from which, as a supreme practical ground, all laws of the will must be able to be derived.

(4) The practical imperative will therefore be the following: *Act so that you use humanity in your own person, as well as in the person of every other, always at the same time as end, never merely as a means* (Wood, 1999, 124f.).

For a discussion of Kant's argument for FH see also (Korsgaard, 1996d, 122 f.).

had every reason to be concerned is evident given his underlying assumption: "If, therefore, freedom of the will is presupposed, morality together with its principle follows from it by mere analysis of its concept" (GMS, AA 04:447.08-09). Kant's argument just relies on an analytic connection between autonomy and morality.

The upshot is this: Kant's deduction, which is meant to complete the transcendental argument, is not successful, because it simply postulates that morality follows analytically from autonomy. Kant fails to provide additional reasons for why the reflective endorsement of FH seems inevitable.

However, there is an alternative for justifying FH. Based on a reflection of what attitudes toward persons are appropriate and fitting, we might well come to endorse that persons are "ends in themselves" and merit respect and recognition. This much is the specific contribution of a fitting-attitude analysis of value. The additional task, however, is to identify the normative reasons that make this attitude toward others also normatively compelling and binding. As we will see, Rønnow-Rasmussen's FA analysis of value provides some essential tools for accomplishing this task.

Before exploring this in more detail, let us look at another attempt to justify FH, namely Korsgaard's proposal to ground FH in the constitutive conditions of agency. The difficulties of her account provide additional motivation to turn to a fitting-attitude analysis of value.

4. Korsgaard's Grounding of the Formula of Humanity (FH)

Korsgaard tries to ground the categorical imperative in our conditions of agency, arguing that the categorical imperative amounts to a constitutive condition of agency. The idea is that constitutive conditions cannot coherently be called into question. This way she tries to complete what Kant seeks to achieve by a transcendental argument, namely that the categorical imperative amounts to a principle that is indispensable for human agents. Here I am not discussing the strengths and weaknesses of this methodological program. I have done so elsewhere (Pauer-Studer, 2018).

Korsgaard reconstructs Kant's ethical theory as a form of 'constitutive internalism' (Korsgaard, 1996a, 2009a). The principles of practical reason—the instrumental principle and the categorical imperatives—are constitutive for the person as a rational agent.[7] The self-legislation of an autonomous will is the source of normativity. The capacity to be an autonomous agent—an agent recognizing the force of universality and valuing her or his humanity as an end in itself—is indispensable to having reasons for action at all.

[7] Note that Korsgaard eventually claims the categorical imperative to be the only principle of practical reason. The hypothetical imperative amounts, in her view, to a sub-principle; its normative force derives from the principle that there are objects (i.e., human beings) that are valuable for themselves (see Korsgaard, 2008, Appendix).

The assumption of autonomy is crucial for justifying the categorical imperatives, foremost the Formula of Universal Law (FUL) and the Formula of Humanity (FH). In vindicating FUL, Korsgaard closely follows Kant's reasoning from autonomy to the first formulation of the categorical imperative in the *GMM*. Her argument, in short, goes thus: A free will or an autonomous will acts according to its own principle or norm, that is to say, it is guided by a self-given law. The principle of a free will is henceforth a law, and this condition, of being a law, is exactly fulfilled by the categorical imperative in the universal law formulation.

In her justification of the Formula of Humanity (FH), Korsgaard equally relies on her methodological assumption that the categorical imperative, as morality more generally, is anchored in the constitutive conditions of agency. She presents an argument (she herself calls it a "fancy" reformulation of Kant's own derivation of FH) based on the notion of 'practical identity'. This concept refers to the particular norms and commitments that define you, for instance, whether you are a mother, a father, a teacher, a national citizen, etc.

Korsgaard's argument, in short, goes like this: To be an agent, we need a normative structure, a practical identity. But we cannot develop practical identities unless we attribute value to ourselves, that is, unless we value our own humanity. Yet, to value our own humanity we must equally value the humanity of others. In other words: Behind our particular practical identities stands our identity as human beings who value themselves, that is, our moral identity. To value ourselves and others as human beings constitutes us as moral agents.

Here are the most important steps of her argument:

1) In order to be an agent, you must act on reasons.

2) In order to act on reasons, you must have *some* conception of your practical identity and you must be committed to it. For otherwise you "would lose your grip on yourself" and you would not have "any reason to do one thing rather than another" (Korsgaard, 1996a, 121).

3) The reason to commit yourself to a practical identity does not spring from another contingent practical identity; it springs from your humanity, i.e., your identity as a human being (as someone who needs reasons to act and to live).

4) It is a kind of reason you only have "if you treat your humanity as a practical, normative form of identity, that is, if you value yourself as a human being" (Korsgaard, 1996a, 121).

5) But to value yourself as a human being involves valuing other human beings as well. An agent must therefore value herself and others in exactly the way articulated by the Formula of Humanity, i.e., "always to treat humanity…as an end in itself, never merely as a means" (Korsgaard, 1996a, 121).

A version of the argument can be formulated in terms of valuing ourselves and others: In order to be agents and have a reason to act, we must consider our ends as

important and put value on them. Our ends only have value insofar as we confer value on them. This entails that I can value my ends merely by valuing myself. But if I have reason to value myself, then I have a reason to value all others. The publicity of reasons (reasons are public not private, according to Korsgaard) forces us to consider others as likewise valuable if we consider ourselves as valuable.

How convincing is this argument? What about its underlying assumption that we––qua our value-conferring capacity—are a source of values?

The latter assumption is due to Korsgaard's defense of metaethical constructivism and her rejection of realism. Values, she states, are only in the world insofar as we put value on things. Realists are wrong when they assume that values are ontologically given entities and that it is our task to detect them. It is us who create and construct values.

Korsgaard's argument thus depends on her endorsement of a first-personal normativism: the source of values and, more generally, of morality, lies in our autonomy as agents. And this autonomy consists in both our capacity for self-legislation and our capacity of valuing, i.e., our capacity to confer value on objects.

However, Korsgaard's line of reasoning does not seem successful in grounding FH. Even if we accept her claim that being an agent entails valuing oneself as a person and, because of reasons being public not private, also valuing other persons, that argument by itself does not imply that one must value the persons around one in the specific and demanding way that is prescribed by Kant's idea of treating others and ourselves as ends in themselves.

I can value myself as an agent also by following the principle that my own interests should simply precede those of others—and the publicity condition might lead us in this case to make concessions, but it does not commit us to the deep form of respect for others that Kant had in mind. Moreover, it seems difficult to imagine why I should be denied identity as an agent because of making an egoistic strategy my principle.

The main problem is that a substantive normative principle such as FH cannot merely be justified by reflecting on individuals' capacities of valuing. We need to add additional arguments for why valuing humans as "ends in themselves" is, or should be, the principal standard for our relations to others and to ourselves.

That Kant's principle that humans have value in themselves is intuitively convincing and consistent with our considered moral judgments seems beyond question.[8] It is therefore worthwhile to explore alternative ways of justifying it.

[8] William FitzPatrick, for instance, claims that the principle that humans have special value as ends in themselves is true; it needs, he maintains, metaethical realism to recognize it as a normative truth (FitzPatrick, 2005, 688f.). There is no space here to discuss in detail the strengths and weaknesses of metaethical realism. Let me just mention that I consider it problematic to assume that ethical principles and/or value statements are true. This amounts either to a mere a priori claim or to the claim that ethical principles and/or value statements are true in virtue of objectively given values. Both accounts are problematic; the first one relies on a crude form of rationalism; the latter assumption raises the worry about the ontological status of such entities as objective values (Mackie called them "queer objects").

5. Fittingness and Deontic Normative Reasons: An Alternative Grounding of Kant's Formula of Humanity (FH)

In the following section, I will show that Rønnow-Rasmussen's fitting-attitude analysis of values is helpful for making sense of the principle that humans have value as ends in themselves. However, we need to make some modifications of his account.

Before starting, let us recall the crucial elements of Rønnow-Rasmussen's analysis: Rønnow-Rasmussen assumes that there are two kinds of final value, 'good' and 'good for'. 'Good' is an intrinsic, non-relational, impersonal value, whereas 'good for' is extrinsic and relational. Both values can be explained in terms of a fitting-attitude analysis (FA).

The first modification I suggest concerns Rønnow-Rasmussen's notion of reasons. Fittingness is, as we have seen, the relation between an object and a response, given that the object is worthy of the response or the response is merited. Recall that Rønnow-Rasmussen holds that fittingness alone does not provide sufficient support for an object having value. The move to values, he argues, requires additional backing. And the normative elements that provide this necessary support are reasons. The kinds of reasons doing the work are, as Rønnow-Rasmussen assumes, pro tanto reasons.

Rønnow-Rasmussen is right when he insists that the FA analysis of value needs to be completed by introducing reasons. However, he is wrong when he states that *all* the reasons here at stake are pro tanto reasons. The FA analysis, he writes, "elucidates why we are (at least *pro tanto*) justified in our concern for objects that are valuable for us" (Rønnow-Rasmussen, 2022, 92). This suggests the following picture: we first need to weigh pro tanto reasons in order to determine whether an object is valuable, from which it follows that a certain response is apt and fitting because the object is worthy of that response.

This explanation does not sit well with the idea that humans are valuable as "ends in themselves". There is something wrong with the suggestion that it might be a matter of weighing pro tanto reasons as to whether humans do have special value and are therefore worthy of respect and recognition. Rather, to put the point in terms of Rønnow-Rasmussen's taxonomy of values, that humans as humans do have special moral standing, amounts to a final value, more precisely a morally central final value. Such a value does not depend on mere considerations in its favour. Instead, the apt kind of reasons here are deontic normative reasons, i.e. reasons that are part of a normative argument for why it is appropriate to accept a principle that attributes to human beings special value as "ends in themselves" which entails that the overriding fitting relation between humans is one of respect and recognition.[9]

[9] Note that we are talking here about a basic moral relation. This does not rule out that there are a variety of particular fitting relations in concrete interactions between persons, relations that also

I want to suggest a second modification, which amounts to an extension of Rønnow-Rasmussen's account. It seems obvious to classify the assumption that humans have special value as "ends in themselves" as an intrinsic, non-relational, and impersonal value—a classification Rønnow-Rasmussen seems to endorse. Now, my suggestion is that this idea—humans have special value as "ends in themselves"—should be read in a non-relational sense as well as a relational one. To formulate the point in terms of Rønnow-Rasmussen's value taxonomy: 'good' and 'good for' mark two dimensions of the Kantian principle (FH). That humans are "ends in themselves" is an intrinsic, non-relational, and impersonal value (first dimension). However, that humans have this special standing is also important in a relational sense (to name the second dimension). It is 'good for' human beings to live together in a way that is committed to a principle of mutual respect, acknowledging that respect and recognition are the appropriate fitting attitudes.

As I will now show, together with these two modifications (deontic normative reasons instead of pro tanto reasons; 'good' and 'good for' as two dimensions of final value), Rønnow-Rasmussen's FA analysis of value contributes greatly to our understanding of the meaning and scope of Kant's Formula of Humanity. I will first turn to Kant's own derivation of FH, and then to Korsgaard's grounding of FH.

To return to our discussion of Kant (section 3), recall that he tried to justify the Formula of Humanity (FH) by an argument involving two steps: first, to lay bare the conditions (embraced by the FH) that underpin our valuing of human beings as autonomous rational agents and "ends in themselves"; second, to provide an additional justification (deduction) of the categorical imperative. As mentioned, Kant's deduction is problematic because it rests on an analytic connection between autonomy and morality (a fact that troubled Kant himself).

If, however, we take the relational dimension of FH into account, we get another picture. The idea is simply to ask why it is good for us to introduce and accept the principle that humans have special value.

Kant himself addressed this relational dimension when he stated that FH leads to what he called a "realm of ends". The term refers to a community in which human beings relate to one another in terms of respect and recognition.[10] Although Kant stressed the importance of that ideal (it contains all other aspects mentioned in the

include negative reactions. For instance, in some situations resentment and anger might be a fitting response. But such responses to ordinary moral failures do (at least in general) not question the special normative status of human beings.

[10] The requirement contained in the idea of a realm of ends can, I think, be formulated in the following way: Act according to principles on which you and others can agree to act since they are constitutive of the moral community as a systematic union of human beings who respect one another and each one's autonomy. (For a more detailed discussion see Pauer-Studer, 2016).

Kant's own formulation of the Formula of a Realm of Ends (FRE) holds: "(A)ct in accordance with the maxims of a member giving universal laws for a merely possible realm of ends" (GMS AA 04:439).

different formulas of the categorical imperative[11]), he did not present the full argument for why we should adopt that ideal.

However, a justification in terms of deontic normative reasons can be provided. The idea is to apply Kant's famous argument, developed in his political philosophy, for why we have to move from a state of nature in which no normative regulations hold to a politically rightful condition to the moral domain. The argument then is: Given that we affect each other by our actions, we need moral principles that require us to live together by entertaining relations of mutual respect. Otherwise, we would be in a social and moral state of nature, i.e. a state without any moral principles and regulations, far from a morally rightful condition.[12] That insight yields a compelling deontic normative reason for endorsing the ideal of a realm of ends. Forming our social interactions on the model of a realm of ends grants us the status of being agents who (ought to) relate to each other in terms of respect and recognition.

The argument also yields a justification of the Formula of Humanity, given that FH is part of the idea of a realm of ends. With respect to FH, we can flesh out the argument in the following way: In order to escape an ethical state of nature, we have a deontic normative reason to accept FH. Otherwise, the safeguards for our bodily and psychological integrity would be missing, with possibly grave consequences such as being degraded, humiliated, and misused by others.

The fitting-attitude analysis further supports accepting FH, as well as the associated ideal of a realm of ends. Recall that the fittingness account invites us to reflect on whether an object merits a certain response. Fittingness is not limited to direct factual experiences, but for the most part works via mental representation. When we imagine the full range of humiliating and hurtful, let alone violent, behaviors that would be possible in a moral state of nature, we already react with rejection and disgust to such a hypothetical thought scenario. Confronted with such experiences in real life, these reactions seem fairly evident.[13] Fittingness makes us aware of moral wrongs. Thus, in addition to a deontic reason-based justification, fittingness provides a crucial form of warrant for FH and the ideal of a realm of ends.

---

[11] According to Kant, the idea of a realm of ends includes FUL and FH, but also the Formula of Autonomy. Kant's Formula of Autonomy (FA) states: "(T)he third practical principle of the will" is "*the idea of the will of every rational being as a will giving universal law*" (GMS AA 04:431; italics in the original).

[12] Kant develops this argument in his political philosophy (TP, AA 08:293-297).

The only way to escape the state of nature, which is for Kant a state without normative regulations (no rights exist) is to establish, politically and legally, what Kant calls "a rightful condition". This is a condition in which individuals enjoy the same rights protecting their equal freedom. Interventions by the state are only permitted in case they are necessary to protect, possibly restore, individuals' equal freedom.

[13] Note that there might of course be gravely distorted conditions that pervert individuals' responses (a point I am not going to discuss here).

Let us turn to Korsgaard's argument for FH as outlined in section 4. As I have pointed out, its shortcomings are primarily due to the fact that Korsgaard locates the source of normativity exclusively in the rational will of the person. With respect to FUL, it is the self-legislating will that provides the justification; with respect to FH, the warrant is provided by the will as a value-conferring entity. However, as I have argued, our capacity to confer value on objects does not yet commit us to value others (and ourselves) as ends in themselves. There needs to be an additional argument for why we should do so. Moreover, as I maintained above, such an argument must be couched in terms of deontic normative reasons and the fittingness relation.

Interestingly enough, Korsgaard comes close to a justification that heeds the relational dimension of FH. The relevant part here is a passage in *The Sources of Normativity*, in which Korsgaard introduces the notion of moral law. As she writes (note that Korsgaard uses the term 'Kingdom of Ends' instead of 'realm of ends'):

> The moral law, in the Kantian system, is the law of what Kant calls the Kingdom of Ends, the republic of all rational beings. The moral law tells us to act only on maxims that all rational beings could agree to act on together in a workable cooperative system (Korsgaard (1996a, 98-99).

In the passage cited above, Korsgaard associates the ideal of a realm of ends with Rawls's idea of society as a fair system of social cooperation. Moreover, she points in the direction of an agreement-based justification of the realm of ends, which also provides a justification of the Formula of Humanity that is part of the realm of ends. Such an agreement, to a "workable cooperative system," would be based on deontic normative reasons.

This fits nicely with our reconstruction of a relational argument for accepting the Formula of Humanity, which relies on the deontic normative reasons for escaping a moral state of nature.

Korsgaard, however, does not pursue her own suggestion and does not develop an argument drawing on the relational dimension of the Formula of Humanity. Instead of exploring the resources of her notion of the moral law, she sticks to a first-person standpoint, locating the source of normativity, and eventually morality, in individual agency and rational willing.[14]

Before concluding, let me address a possible objection. Why, one might ask, introduce the notion of fittingness at all? Why not simply work with the categories 'final value' and 'normative reasons'? Why not just rely on the justification normative reasons provide for assuming an object to have final value?

[14] The passage on the moral law remains an isolated passage in Korsgaard's work. In a reply to Stephen Darwall's defense of a second-person moral standpoint, Korsgaard stresses again her endorsement of a first-person account of morality. She sticks to her assumption that the source of normativity lies in rational autonomy and self-legislation (Korsgaard, 2007).

The answer, as Rønnow-Rasmussen emphasizes, is that a fitting-attitude analysis of value takes into account the relation between valuable objects and attitudes. More generally, fittingness reminds us that we, as subjects, need to respond in a specific way to the valence of objects and given facts. In other words, valuable objects are not inert entities of our social world. Rather, they appeal to us and demand a certain response and reaction.

Let me end this section with some general remarks on the merits of a fittingness account. Fittingness is important in the following respects: It heightens our moral awareness, and it enriches our moral thinking. Reflecting on the features of objects and their demands on us in terms of fitting responses enhances our moral knowledge and understanding. It deepens our sense of owing others, and also ourselves an attitude that takes other persons' moral standing and needs seriously. By representing in thought the relations to others that go wrong and arouse our disapproval, we are motivated to re-think our responses and attitudes toward others, as well as to ourselves.[15]

6. Concluding Remarks

This paper sought to show that Rønnow-Rasmussen's fitting-attitude analysis of value offers important clues for an interpretation of the idea that human beings are valuable in themselves, as expressed in Kant's Formula of Humanity. I argued that Rønnow-Rasmussen's assumption of two final values, i.e., a non-relational final value ('good') and an extrinsic and relational final value ('good for'), allows us to determine two dimensions of the principle that humans are valuable as ends in themselves, namely to flesh out its importance as a basic intrinsic value, but also its importance for a relational understanding of morality.[16] I also argued that the fitting-attitude analysis of value, as presented by Rønnow-Rasmussen, contributes to our understanding of the Kantian idea and thus to a central standard of morality.

The paper also indicated why the combination of deontic normative reasons and fittingness is important for a justification of normative principles. A rather intricate question is whether the two notions, deontic normative reasons on the one hand,

[15] In seeing the relevance of a fittingness account for moral theory, I am deeply indebted to Fabienne Peter's work (see Peter, 2022). In a preceding published paper (Peter, 2019), Peter also discusses an entitlement-based form of warrant that is distinct from a reason-based normative justification. Although Peter interprets the entitlement-based warrant in terms of direct support by facts, her insightful analysis of this form of justification can be applied, I think, to the kind of warrant fittingness provides.

[16] There is no space here for discussing the relevance of relational accounts of morality. Just let me note that first-personal justifications of morality often run into the dilemma of relying on a rather limited first-person perspective that does not sufficiently consider the interpersonal perspective and the moral claims of others.

fittingness on the other, amount to two distinct forms of warrant for normative principles. Rønnow-Rasmussen, as I understand his account, tends to the view that reasons are part of the FA analysis and do not present a distinct form of justification. The picture I have developed rather suggests a two-track interpretation, according to which deontic normative reasons and fittingness yield two forms of warrant. However, I will leave the discussion of this issue to future work.[17]

References

FitzPatrick, William J. (2005), "The Practical Turn in Ethical Theory: Korsgaard's Constructivism, Realism, and the Nature of Normativity". *Ethics*, 115(4): 651-691.

Kant, Immanuel (1785/1996) "Groundwork of the Metaphysics of Morals". In I. Kant, *Practical Philosophy* (Mary Gregor, Trans. and Ed.) (273-309). Cambridge: Cambridge University Press (cited as GMS AA:04, according to the Academy edition of Kant's works).

Kant, Immanuel (1793/1996) "On the Common Saying: That May Be Correct in Theory, But It Is of No Use in Practice". In I. Kant, *Practical Philosophy* (Mary Gregor, Trans. and Ed.) (37-108). Cambridge: Cambridge University Press (cited as TP AA:08, according to the Academy edition of Kant's works).

Korsgaard, Christine M. (1996a) *The Sources of Normativity*. Cambridge: Cambridge University Press.

Korsgaard, Christine M. (1996b) *Creating the Kingdom of Ends*. Cambridge: Cambridge University Press.

Korsgaard Christine M. (1996c) "Morality as Freedom". In Ch. M. Korsgaard *Creating the kingdom of ends* (159–187). Cambridge: Cambridge University Press.

Korsgaard, Christine M. (1996d) "Kant's Formula of Humanity". In Ch. M. Korsgaard *Creating the Kingdom of Ends* (106–132). Cambridge: Cambridge University Press.

Korsgaard Christine M. (2007) "Autonomy and the Second-Person within: a commentary on Stephen Darwall's The second-person standpoint". *Ethics*, 118(1): 8–23.

Korsgaard, Christine M. (2008) "The Normativity of Instrumental Reason". In Ch. M. Korsgaard, *The Constitution of Agency* (26-68). Oxford: Oxford University Press.

Korsgaard, Christine M. (2009a) *Self-Constitution: Agency, Identity, and Integrity*. Oxford: Oxford University Press.

Korsgaard, Christine M. (2009b) "The Activity of Reason". *Proceedings and Addresses of the APA* 83(2): 23–43.

[17] Acknowledgments: Research for this article has been funded by the European Union's Horizon 2020 research and innovation program under grant agreement No. 740922, ERC Advanced Grant 'The Normative and Moral Foundations of Group Agency'.

Pauer-Studer, Herlinde (2016) "'A Community of Rational Beings'. Kant's Realm of Ends and the Distinction between Internal and External Freedom". *Kant-Studien* 107(1): 125-159.

Pauer-Studer, Herlinde (2018) "Korsgaard's Constitutivism and the Possibility of Bad Action". *Ethical Theory and Moral Practice* 21(1): 37-56.

Peter, Fabienne (2019) "Normative Facts and Reasons". *Proceedings of the Aristotelian Society* 119(1): 53-75.

Peter, Fabienne (2022) "It's Not All About Reasons: The Fittingness of Actions and Attitudes", unpublished manuscript.

Rønnow-Rasmussen, Toni (2011) *Personal Value*. Oxford: Oxford University Press.

Rønnow-Rasmussen, Toni (2017) "Good and Good For". In H. LaFollette (ed.) *The International Encyclopedia of Ethics.* Hoboken, NJ: Wiley Blackwell.

Rønnow-Rasmussen, Toni (2022) *The Value Gap*. Oxford: Oxford University Press.

Street, Sharon (2012) "Coming to Terms with Contingency: Humean Constructivism About Practical Reason". In J Lenman & Y. Shemmer (Eds.) *Constructivism in Practical Philosophy* (40-59). Oxford: Oxford University Press.

Wood, Allen W. (1999) *Kant's Ethical Thought*. Cambridge: Cambridge University Press.

What Does It Mean for a Species to Be Alien – And Why Is It a Bad Thing?

Erik Persson

Abstract. Invasive alien species are frequently discussed in academic literature, by practitioners, government agencies, and popular media, but what does it mean for a species to be alien and why is this seen as a bad thing? To answer these questions, I have analysed texts about invasive alien species in academic journals and in communication from government agencies. The almost totally unanimous answer to the first question was that a species is alien if and only if it is introduced to an area outside its natural range by humans. I found three primary answers to the second question, namely that (1) alien species are more probable to behave invasively or that it is impossible to know for sure if an alien species will behave invasively, (2) being alien is bad in itself or at least that alien species have a lower value than native species and native environments, and (3) being moved by humans is unnatural and being unnatural has negative value. All three answers probably contribute to why being alien is considered a negative property in species but none of them seem like a satisfying answer to why being alien should be seen as a bad thing in a species.

Introduction

When the word ‘alien’ is used about humans, it usually implies that someone is from somewhere else, usually another country but it could also be another village or (maybe someday) another planet. The word is in that sense, similar to ‘foreigner’,

but it seems to go beyond that. 'Alien' can also mean that something or someone is strange in some sense. An idea, for instance, can even be considered alien in a certain context without being foreign. A person can be alienated even if she is not from somewhere else or moving somewhere else physically, if she is in some sense, be it mentally, emotionally, or intellectually, separated from her group or society. When used about humans, both 'alien' and 'foreign' often also have negative connotations. Throughout most of human history, being alien, in any of the senses mentioned, usually implied a lower value or even a lower moral status. This is in fact still a popular opinion in certain circles though it has no support whatsoever in modern ethical theory.

In astrobiology, 'alien' usually means extraterrestrial. That is, coming from somewhere other than Earth. In astrobiology, the value of alien life is clearly positive, and very high. Finding alien life is of course the holy grail of astrobiology. In this case it is a matter of epistemic value, but empirical studies show that many people assign a very high end value to alien life (e.g. Persson et al. 2019). There are also good reasons for assigning a high end value as well as instrumental value to extraterrestrial life (as shown by e.g. Cockell 2011a, b; Persson 2017, 2019).

From an ethical perspective, the discussions about alien life - where alien means extraterrestrial - have just started. If we depart from Hollywood's depictions of alien life, it is usually (though not always), quite simple. If aliens look more or less like us (e.g., the Navi in the film *Avatar*) then we should sympathize with them. If they look like lizards or giant insects (as in the *Alien* films or the TV series *V*), then they are dangerous, and we should fight them. Among people involved in space exploration, the attitudes differ. Those in favour of human settlements on, for example, Mars, show very little concern for extraterrestrial life (e.g., Smith 2009; Zubrin 1996), while others agree with Carl Sagan's statement in his famous TV show Cosmos, that "[i]f there is life on Mars, I believe we should do nothing with Mars. Mars then belongs to the Martians, even if they are only microbes." Philosophers discussing the moral status of extraterrestrial life divide more or less along the same lines as those discussing moral status on Earth. That is, we have advocates of a cosmic version of anthropocentrism, sometimes called ratiocentrism (Smith 2009). We have biocentrism (Cockell 2005, 2011), sentientism (Persson 2012) and ecocentrism (Rolston 1986). In addition, we also have cosmocentrism (Lupisella 2020; MacNiven 1995; McKay 1990) that in practice does not seem to differ substantially from biocentrism but where extraterrestrial life is explicitly included.

Let us, however, go back to Earth and the questions that are the focus of this chapter: What does it mean to call a species alien and why is it seen as a bad thing for a species to be alien?

In addition to being philosophically interesting, these questions have gained practical importance in connection with the ongoing campaigns against invasive alien species (IAS) that recently has got quite a lot of attention in conservation biology, law and policy making, as well as in popular media.

When used about species, the word 'alien' is in fact almost always used as a part of the concept of IAS and it will thus be analysed in this context.

The increased attention towards IAS is beneficial to our investigation since it means there is a lot of material to draw from, both in the form of academic literature and in the form of, for example, information texts from government agencies. There is also a downside, however, namely that the close connection between 'alien' and 'invasive' in these texts makes it difficult to analyse any one of these terms individually. Nevertheless, the fact that 'invasive' and 'alien' are so closely connected in this discussion is interesting in itself and I will try to make good use of this in my analysis.

Academic texts about IAS most often occur in journals about conservation biology in a broad sense. It is surprisingly rare that these texts attempt to define the whole term or any of the included terms, however. They also seem to take it for granted that being an IAS is something negative (e.g., Cuthbert et al. 2021; Latombe et al. 2017; Mgidi et al. 2007; Sax & Gaines 2003). There are some examples, however, where academic texts do present a definition and a few texts that attempt to defend or contest the assumption that IAS is necessarily something negative. I will make use of these texts and also of official statements about IAS on the websites of Swedish authorities. What makes the latter good sources for this investigation is that they tend to be very thorough and have a pedagogical ambition. They therefore typically both define 'IAS' and explain why it is important to fight them, that is, why being an IAS is something negative.

Let us start, however, with what seems to be the most influential definition of 'Invasive alien species', formulated by the International Union for Conservation of Nature (IUCN):

"Invasive alien species are animals, plants or other organisms that are introduced by humans, either intentionally or accidentally, into places outside of their natural range, negatively impacting native biodiversity, ecosystem services or human economy and well-being." (IUCN).

This formulation indicates three distinct answers to the descriptive as well as the normative aspects of what it means to be an invasive alien species:

- Being introduced by humans,
- Being outside its natural range, and
- Negatively impacting native biodiversity, ecosystem services or human economy and well-being.

Can this tell us anything about what it means for a species to be alien as such?

The first two properties listed here seem to focus on the alienness, while the last aspect seems to be about invasiveness, though, if invasive species have these negative impacts independently of them being alien, then why does invasiveness need to be clustered together with alienness?

Alien Species as a Threat

Invasive alien species are usually assigned negative value because they threaten ecosystems (Fantle-Lepczyk et al. 2022; Latombe et al. 2017; Simberloff et al. 2013), ecosystem services (Cuthbert et al. 2021; Latombe et al. 2017) and other natural resources (Simberloff et al. 2013), society (Mgidi et al. 2007), economic values (Cuthbert et al. 2021; Fantle-Lepczyk et al. 2022; Gholizadeh, et al. 2022; Mgidi et al. 2007; Naturvårdsverket 2021), human health (Naturvårdsverket 2021), or wellbeing (Cuthbert et al. 2021; Simberloff et al. 2013), biodiversity (Cuthbert et al. 2021; Gholizadeh, et al. 2022; Latombe et al. 2017; Naturvårdsverket 2021; Simberloff et al. 2013), or native species (Fantle-Lepczyk et al. 2022; Latombe et al. 2017; Sax & Gaines 2003; Simberloff et al. 2013). This seems like a good reason to assign a strong negative instrumental value to invasive species, but what does this have to do with being alien?

A possible answer could be that only alien species behave invasively. We know that this is not true, however. Another answer, that also makes more sense, is that they behave invasively more often than they contribute positively to their new environment. Another answer that has been suggested is that it is a matter of precaution (Simberloff et al. 2013). If a species is moved from one environment to another it is very difficult to say for certain that it will not have any negative effects in the new environment. That is why it is often recommended to always avoid moving species even if they are not shown to be invasive.

In these cases, the negative attitude towards alien species is purely instrumental. Alien species are not despised because they are alien as such but because of the risk they pose to other species (Simberloff et al. 2013).

This seems like a plausible answer to why alienness and invasiveness is connected in the expression ‘invasive alien species’ and also to why alien species are sometimes persecuted even though they are not shown to be a threat.

If this is the only reason for mentioning alienness in the context of IAS maybe discussions about IAS does not tell us a lot about what it means for species to be alien, but it does tell us something about the negative connotations of alienness in species.

Things are not as crystal clear as they may seem, however. Conservation biologists seem to have different opinions on which is most common, that new species become a threat to their new environment or that they provide a valuable addition to it (see Peretti 1998; Sagoff 2009; Sax & Gaines 2003; Simberloff et al. 2013 for different opinions in this matter). This takes away some of the credibility of the precautionary approach. If alien species overall contribute more value than they take away in their new environments, precaution may not be the best approach.

On the other hand, even if it is not correct that alien species overall do more harm than good, it is still possible that the influence of this view is the main or the only explanation to why alienness is connected with invasiveness.

There are other problems, however. If alien species should be stopped or exterminated because they are merely a risk to other values, then why are not native species that behave invasively treated in the same way?

There are well-documented cases where human encroachments (other than importing new species) cause certain indigenous species to start behaving in an invasive way in environments where they are already established and have been established for a long time. Examples of such human induced changes are eutrophication and climate change. Why are not these invasive non-alien species persecuted in the same way as invasive alien species?

We should also ask ourselves, if the relevance of being alien is merely instrumental, why include it in the name 'invasive alien species'? Would it not make more sense to just call it 'invasive species' and motivate the campaigns against alien species with the risk that they will become invasive if we move them?

Finally, we also need to account for the first two aspects of what it means to be an IAS in the IUCN definition. That is, being introduced by humans and being outside of its natural range. How come the definition of 'alien' focuses on human involvement and naturalness? Why not on, for instance, how different its place of origin was from its new environment, how far or how fast the species has travelled or how long ago it came to its new environment?

Aliens vs. Natives

In many places, alien species are not just seen as problematic when they threaten other values, including other species that are, for example, economically valuable or keystone species in the ecosystem, or when they directly or indirectly threaten a larger number of species and thus cause a decrease in biodiversity. Sometimes, one alien species threatens one indigenous species and there will be no other consequences. In these cases, one species is substituting another species without any further effects on, for example, the economy or biodiversity. Why is that a problem? Here it seems that being alien is seen as bad in itself.

This is even more pronounced in cases where introduced species are being exterminated even if they are not invasive and do not threaten any other species. These cases seem to be less common, but they do occur. Simberloff et al. (2013) deny that invasion scientists see alienness as being bad in itself. Several other authors disagree, however. Young & Larson (2011) claim that "invasion biology places a value on existing biodiversity", and Cidrás & González-Hidalgo (2022) states: "This dichotomy, [i.e. between "natives and aliens"] which is essentially a geographical categorization hinging on questions of placement and displacement, has been strongly supported in the last few decades in invasion science."

Especially Peretti and Sagoff are very clear about their views:

> "Nativist trends in Conservation Biology have made environmentalists biased against alien species. This bias is scientifically questionable, and may have roots in xenophobic and racist attitudes." (Peretti 1998).
>
> "...the purism of biological nativism has historically been associated with fascist and apartheid cultures and governments" (Peretti 1998).
>
> "... many environmental scientists are committed to the idea of pure, 'native' nature." (Peretti 1998).
>
> "No matter how species-rich, beautiful, and complex an ecosystem may appear to the average city dweller, the biologist will see it as degraded insofar as alien species invade it." (Sagoff 2009).

It is also easy to find examples of authors in the field who clearly describe the issue of IAS as a conflict between native and alien species or biotas (e.g., Chaffin et al. 2016; Gholizadeh, et al.; Mgidi et al. 2007).

In a questionnaire study aimed at conservation biologists, 37% answered that they agree and 34% that they do not agree with the statement "Exotics are an unnatural, undesirable component of the biota and environment". (Young & Larson 2011).

There has been some speculation about why native species are seen as inherently more valuable than introduced species. As we saw above, Peretti (1998) associates it with xenophobia. He mentions also that seeing alienness in species as a sufficient reason for extermination campaigns has historically been associated with racist regimes. He mentions World War II Germany and South Africa as examples.

Simberloff et al. (2013) contest the claim of xenophobia. They write: "The wish to maintain the global diversity of native communities and ecosystems has nothing to do with xenophobia. On the contrary, it stems from principles similar to those that defend the right for every human society to retain its cultural distinctiveness, as proclaimed by the Council of Europe and UNESCO."

I am not sure this is a good reply. The right to retain one's cultural distinctiveness may be granted by the Council of Europe and UNESCO, but it is not obvious that it stays totally clear of xenophobia.

Nevertheless, it seems implausible that a large majority (as it is) of the world's conservation biology researchers and practitioners are motivated by xenophobia. It is clearly possible to prefer what one has over what one might get and to have a special relation to existing species that results in valuing them as ends in themselves. A complicating factor for the xenophobia explanation is also that campaigns against IAS are prevalent in most countries around the world, not only in countries whose governments have a xenophobic ideology. This is also true for countries where protection of native species against alien species is an explicit part of the motivation for the fight against IAS. This includes South Africa of today, post-apartheid.

The fact that not just South Africa but also USA, another former colony has protection of native species and keeping native environments clean from alien

species as an explicit motivation for its fights against IAS (Chase 1987; Peretti 1998; Sagoff 2009; Wilson 1992), could indicate another possible explanation, namely that alien species are consciously or unconsciously associated with colonialism (Crosby 1986; Heywood 1989; Peretti 1998). New species can thus in themselves be seen as colonisers or are strongly associated with human colonisers. This cannot be the whole truth either, however, since just as alien species are prosecuted just for being alien in countries that are not led by xenophobic governments, it is also true that this happens in countries that do not have a colonial past.

Unnaturalness

The IUCN definition of IAS mentions two properties that do not seem to have anything to do with invasiveness but that seem to be clearly meant as criteria for alienness: Being introduced by humans and being outside of the species' natural range.

These criteria are echoed in many other definitions. Fantle-Lepczyk et al. (2022) define 'Non-native invasive species' as "organisms introduced beyond their natural range by human activity." Antonsich (2021) distinguishes between native and alien species thus: "natives are species occurring within their natural range and whose dispersal is independent of human action, whereas aliens are species which have crossed a biogeographic barrier thanks to human action". Gholizadeh, et al. (2022) use a less categorical characterisation by stating that "IAS are often introduced by humans to habitats outside of their natural range".

In Sweden, the game plan for the campaign against invasive alien species is set up by the Environmental Protection Agency (EPA) together with the Agency for Marine and Water Management (AMWM). On their website, EPA tells us that "Invasive alien species are plants, animals, fungi and microorganisms that have been intentionally or unintentionally moved to a new environment where they spread rapidly and cause damage to biodiversity, the economy and potentially human health." (Naturvårdsverket 2021).

The Agency for Marine and Water Management (AMWM) explains on their website that "Invasive Alien Species, IAS, are animals, plants and organisms that are introduced accidentally or deliberately by humans into an environment where they are not normally found." (Havs- och Vattenmyndigheten 2021).

In a presentation for the members of the research project *The Human Aspect of Invasive Alien Plants*, at the Pufendorf Institute for Advanced Studies (https://portal.research.lu.se/en/projects/the-human-aspect-of-invasive-alien-plants-the-paradox-of-plants-p), an EPA representative defined 'alien species' (in translation from Swedish) as "a species that has been introduced outside of its

natural habitat after the year 1800 and that can survive and reproduce [in its new habitat]" (Persson 2021).

Here, we get an additional clue in the form of a year - 1800. The fact that this answer was not in terms of a time range (for example 200 years) but in terms of a particular year indicates that it does not matter as such how long a species has been in its new environment. A species cannot be "unalienated" just by being a long-time resident in an area. Instead, there is something special with the year 1800. What is that?

On a direct question about what is special with the year 1800 accompanied by the suggestion that the EPA saw the nature in Sweden at this particular year as an ideal state, the presenter emphatically denied the "ideal state" suggestion and explained that the reason for the choice of the year 1800 was that before that year we do not have enough information about how species were introduced. So, the real answer seems to be that it is not the year or the time range as such that is important. Instead, the answer emphasises that it is a matter of how. In combination with the use of the word 'introduced', this answer fits well with the definition on the EPA website that states that a species needs to be moved to its new habitat to be branded as alien. Both the phrase, 'be moved' (rather than 'move by itself') and the word 'introduced' indicate that only species that have been moved by someone or something else - not species that have moved or spread by their own power - are considered alien by EPA.

The Swedish version of the same web page says: "Invasiva främmande är arter som med människans hjälp flyttats från sin ursprungliga miljö och i sin nya omgivning börjar sprida sig snabbt och orsakar allvarlig skada för ekosystem, infrastruktur eller människors hälsa vilket medför stora kostnader för samhälle och enskilda."

In English translation, the first part of the definition says: "Invasive alien [species] are species that by human help has been moved from its original environment …".

It is not clear why humans are mentioned in the Swedish version but not in the English version. Maybe it is meant to be implied in the English version. Nevertheless, this definition clearly follows the trend: A species needs to be moved by humans to be alien.

This answer is somewhat informative but still quite unsatisfying. It differs quite substantially from the way 'alien' is used in other contexts as we saw in the introduction to this chapter. It is also a bit puzzling why the 'how' – or rather, the 'who' – is so important, and why the answer to the 'who' question seems to be humans.

Let us return for a moment to the second criterion of alienness mentioned in the IUCN definition. This criterion mentioned the natural range of the species. The terms 'natural' and 'unnatural' are notoriously illusive. There are almost as many definitions of the terms as there are authors trying to define them (e.g., Soulé and Lease 1995, Bennett and Chaloupka 1993, Cronon 1995). It is even questioned

whether the distinction is meaningful considering the high degree of human influence on nature everywhere on Earth. (e.g., Peretti 1998).

On top of this, the normative content of the words is debated. Is natural necessarily good? Is unnatural necessarily bad?

One thing that phenomena described as unnatural usually have in common is that they are somehow connected with humans. Sometimes categorically so - anything having a certain type of connection with humans is unnatural. Sometimes it seems to be a matter of degree. The more human involvement, the less natural a phenomenon is. If we assume that naturalness is a matter of low or no human interference, the two criteria for alienness mentioned by the IUCN definition – being introduced by humans and being outside their natural range – seem to converge towards an idea that being alien for a species is to be moved by humans outside of the areas where the species occur without human interference and that this is unnatural and therefore bad.

One might imagine that species that have come about due to breeding or synthetic biology can qualify as unnatural, and probably also alien in any environment outside the lab or the farm. Being unnatural, and therefore alien, for an IAS is not a matter of how the species came about, however, but a matter of how it came to turn up in its new location. If it is moved by humans, its occurrence in its new location is unnatural and thus unwanted, and this is what is meant by the word 'alien' in this context.

If naturalness is the key criterion for alienness, it would also explain why only invasive alien species, and not invasive non-alien species need to be exterminated. Behaving in an invasive way in a species "home" environment may have negative consequences but it is not unnatural.

As we noted above, the basis for seeing 'natural' as positive and 'unnatural' as negative, is shaky, but is naturalness a good basis for branding a species as alien?

Peretti (1998) points out that "The words 'native' and 'natural' are closely linked. The Latin 'nascor' is the original root for several English words including native, natural, nation, and natality." So, there might be an etymological basis for associating 'native' and 'natural' and therefore 'alien' and 'unnatural'. We should not rely too heavily on this connection, however. Stating that alien species are unnatural because of the etymological connection would be to commit the genetic fallacy.

There is also a biologically based problem with the connection between 'native' and 'natural', namely the fact that nature is not static, and it is only "natural" that species come and go. It is well-established that nature types have a certain succession order. Some species pave way for other species that in turn supplant the first species.

A practical problem with defining 'alien' in terms of active human interference is that it might cause problems for assisted migration as an answer to climate change (see e.g., Hoegh-Guldberg et al. 2008; McLachlan et al. 2007; Minteer & Collins 2010; Richardson et al. 2009 for information about assisted migration).

Nevertheless, the connection between alienness and naturalness may well be an important explanation for why human introduction is mentioned in so many definitions of IAS and it may well be a prominent explanation for why being alien is considered a negative property for species, both in connection with invasiveness and in its own right.

Conclusions

The aim of this chapter was to identify what it means for a species to be alien, and why this is considered a bad thing. Since being alien when it comes to species is almost only discussed as part of the concept 'invasive alien species', this is the context in which the questions are discussed. The questions were therefore investigated by using academic texts and communication from Swedish government agencies about invasive alien species.

The sources were almost totally unanimous regarding the question of what it means for a species to be alien in the chosen context, namely, to be moved from its natural range by humans.

I identified three primary answers to the question of why being alien is considered a negative property for a species:

1. Invasive species threaten important environmental, economic, and other values. We can never know for sure which species will start behaving invasively but we do know that when species are moved to a new environment, there is a higher probability they will behave invasively and cause problems than that they will be positive additions to the new environment. Therefore, all alien species need to be treated as potentially invasive and be banned from entering and exterminated if they have already been introduced.

2. Native species and native environments are valuable in their own right. Alien species have a negative value in their own right because they are non-native, or at least they have a much lower value than native species. Alien species sometimes threaten native species and even when they do not, they degrade native environments just by being introduced to these environments. I also identified two possible explanations for why being native was considered valuable in itself and why being alien was being seen as negative in itself. One was xenophobia, the other was that alien species are associated with colonialism. None of them can completely explain this answer, however. It is very implausible that everyone who promote or take part in the fight against IAS have xenophobic motives and the fact that native species and environments have been preferred by xenophobic governments does not explain all other cases where non-xenophobic governments outlaw alien species. It is also not the case that only former colonies outlaw alien species.

3. Being moved by humans is unnatural and being unnatural is bad. This answer is philosophically weak. The distinction between 'natural' and 'unnatural' is questionable. 'Alien' is not defined in terms of 'unnatural' in most other contexts, which is not a problem per se since the discussion in this chapter is explicitly set in this particular context, but it does lower the usefulness of the answer. Finally, and maybe most importantly, the connection between being unnatural and negative value is unexplained. Nonetheless, the opinion that being unnatural implies a negative value is very common in all kinds of contexts and references to unnaturalness are common in the literature about IAS. This answer is also closely connected to the overwhelmingly most common answer in the literature as well as among the government agencies to the question of what it means for a species to be alien. It is therefore plausible that this answer is a very common reason for considering alienness a bad thing in a species.

All three answers have some plausibility as answers to why being alien *is considered* to be a bad thing in a species. None of the answers seem to provide a really solid reason for why being alien *should be* considered a negative property for a species, however, and it has in fact been argued that the term 'invasive alien species' should be discarded and substituted with the term 'invasive species'.

References

Antonsich, M. (2021) "Natives and aliens: Who and what belongs in nature and in the nation?". *Area*, 53(2): 303-310.

Bennett, J. & Chaloupka, W. (eds) 1993. *In the Nature of Things: Language, Politics and the Environment*. University of Minnesota Press.

Chaffin, Brian C., et al. (2016) "Biological invasions, ecological resilience and adaptive governance". *Journal of Environmental Management,* 183: 399-407.

Chase, A. (1987) *Playing God in Yellowstone*. Harcourt Brace and Company.

Cidrás, Diego & González-Hidalgo, Marien (2022) " 'De-eucalyptising Brigades' in Galicia, Spain." *Political Geography,* 99: 102746.

Cockell, C. (2011a) 'Ethics and Extraterrestrial Life', in *Humans in Outer Space – Interdisciplinary Perspectives*, ed. Nina-Louisa Remuss, Kai-Uwe Schrogl, Jean-Claude Worms and Ulrike Landfester (New York: Springer), 80-101.

Cockell, Charles S. (2011b) "Microbial Rights?" *SMBO Reports* 12: 181.

Cronon, W. (ed.) (1995) *Uncommon Ground: Toward Reinventing Nature*. W.W. Norton.

Crosby, A. (1986) *Ecological Imperialism: The Biological Expansion of Europe, 900-1900*. Cambridge University Press.

Cuthbert, Ross N. (2021) "Global economic costs of aquatic invasive alien species". *Science of the Total Environment,* 775: 145238.

Fantle-Lepczyk, Jean E., et al. (2022) "Economic costs of biological invasions in the United States". *Science of the Total Environment,* 806: 151318.

Gholizadeh, Hamed, et al. (2022) "Mapping invasive alien species in grassland ecosystems using airborne imaging spectroscopy and remotely observable vegetation functional traits". *Remote Sensing of Environment,* 271: 112887.

Havs- och Vattenmyndigheten (2021) *Invasive alien species*. https://www.havochvatten.se/en/facts-and-leisure/invasive-alien-species.html Accessed 14/10//2022

Heywood, V.H. (1989) "Patterns, extents and modes of invasion by terrestrial plants" in

Hoegh-Guldberg, O., et al. (2008) "Assisted colonization and rapid climate change". *Science,* 321: 345–346.

IUCN *Invasive Alien Species* https://www.iucn.org/our-work/topic/invasive-alien-species Accessed 14/11/2022

Latombe, Guillaume, et al. (2017) "A vision for global monitoring of biological invasions". *Biological Conservation* 213: 295-308.

Mgidi, T.A., et al. (2007) "Alien plant invasions—incorporating emerging invaders in regional prioritization: A pragmatic approach for Southern Africa". *Journal of Environmental Management,* 84: 173-187.

Minteer, Ben A. & Collins, James P. (2010) "Move it or lose it? The ecological ethics of relocating species under climate change". *Ecological Applications,* 20(7): 1801-1804.

McLachlan, J. S., et al. (2007) "A framework for debate of assisted migration in an era of climate change". *Conservation Biology,* 21: 297-302.

Naturvårdsverket (2021) *Avoid spreading invasive alien species* https://www.naturvardsverket.se/en/topics/invasive-alien-species/avoid-spreading-invasive-alien-species/ Accessed 4/4/2022

Peretti, Jonah H. (1998) "Nativism and Nature: Rethinking Biological Invasion". *Environmental Values, 7*: 183-92.

Persson, Erik (2017) "Ethics and the potential conflicts between astrobiology, planetary protection and commercial use of space". *Challenges* 8(1): 12.

Persson, Erik (2021) "Vad gör en växt främmande? – Några olika perspektiv" in Alkan Olsson, Johanna, et al. *Växtvärk - Perspektiv på invasiva främmande växter i svensk natur* (31-42). Palaver förlag.

Persson, Erik (2019) "A philosophical outlook on potential conflicts between planetary protection, astrobiology and commercial use of space" in Lehmann-Imfeld, Z; Losch, A. (eds.) *Our Common Cosmos* (141-160). Bloomsbury Publishing.

Persson, Erik; Čápová, Klara Anna; Li, Yuan (2019) "Attitudes towards the scientific search for extraterrestrial life among Swedish high school and university students International Journal of Astrobiology" 18(3): 280-288.

Richardson, D. M., et al. (2009) "Multidimensional evaluation of managed relocation". *Proceedings of the National Academy of Sciences*, 106: 9721-9724.

Sagoff, Mark (2009) "Who is the invader? Alien species, property rights, and the police power". *Social Philosophy and Policy,* 26(2): 26-52.

Sax, Dov F.; Gaines, Steven D. (2003) "Species diversity: from global decreases to local increases". *Trends in Ecology and Evolution,* 18(11): 561-566.

Simberloff, Daniel (1997) "Nonindigenous Species—A Global Threat to Biodiversity and Stability" in Raven, Peter H. & Williams, T. (eds.), *Nature and Human Society: The Quest for a Sustainable World.*

Simberloff, Daniel (2013) "Impacts of biological invasions: What's what and the way forward". *Trends in Ecology & Evolution* 28(1): 58-66.

Smith, Kelly, C. (2009) "The trouble with intrinsic value: an ethical primer for astrobiology" in: *Exploring the Origin, Extent, and Future of Life*, Bertka, C.M. (ed.), (261–280) Cambridge University Press

Soulé, M.E. & Lease, G. (eds) (1995) *Reinventing Nature? Responses to Postmodern Deconstruction*. Island Press.

Wilson, A. (1992) *The Culture of Nature*. Blackwell.

Zubrin, R. & Wagner, R. (1996) *The Case For Mars.* Simon and Schuster

Denialism Regarding Moral Mega-Problems

Ingmar Persson

> 'There are different degrees in this aversion to truth; but all may perhaps be said to have it in some degree, because it is inseparable from self-love.'
>
> Blaise Pascal, *Penseés*: 100.

1. The Concept of Denialism

In *Unfit for the Future* (2012), Julian Savulescu and I argued that the moral psychology with which evolution has equipped human beings makes them unfit to deal with the contemporary moral mega-problems of anthropogenic climate change (including environmental degradation and loss of biodiversity) and the global inequality of welfare. For humans to be capable of tackling these problems they need to be morally enhanced by all possible means, including means of a biomedical kind. We concentrated on arguing for means of a biomedical kind because they are controversial and opposed by many, but we did not exclude the effectiveness of other means, like traditional moral education. Here the focus will be on moral education, especially the contribution modern moral philosophy could make to it, though there may be more promising forms of moral education.

I shall conduct the discussion in terms of the problem of ameliorating anthropogenic climate change because it involves the other moral mega-problem of rectifying global inequality, since poorer nations need aid from more affluent nations not only to alleviate starvation and diseases but harm that may result from climate change. A reason why the problem of anthropogenic climate change is hard to solve is that it comprises a *double denialism*. Denialism consists in a denial that something is true that is made not because it is supported by reasons for thinking

that it is not true, but because of a desire or wish that it not be true. There are different degrees of this type of denial. In the strongest form, it is denial that something is true in the face of *overwhelming* evidence that it is true. Then the desire or wish that it be false must be quite powerful to surpass the thrust of this evidence. The denial that the global warming that we are currently observing is to a significant extent anthropogenic, despite what almost all climatological experts maintain, is of this sort.

But apart from such climate science skepticism, this issue also involves another kind of denialism, of a more moral kind, a denial that our emissions of CO_2 and other greenhouse gases are not morally wrong. This is not due to a denial of what climate scientists tell us, but rather simply to the fact that they appear so different from acts that are *archetypally* morally wrong. This paper is mainly devoted to spelling out these differences.

With respect to this feature of the relevant acts being so unlike acts that are archetypally morally wrong, the issue of mitigating anthropogenic climate change resembles the other moral mega-problem of today, the issue of the moral obligation of the affluent to reduce the global inequality of wealth. But the latter issue does not exemplify double denialism because it does not comprise anything corresponding to climate science skepticism. This counterpart would in the case of global inequality be a *blanket* denial that aid from affluent countries to poorer countries could be effective to the end of easing poverty. But this denial is highly implausible, though *in some particular cases* there might be unfounded denials that aid is effective to this end because it is wrongly suspected that the aid ends up in the pockets of corrupt politicians. The fact that the problem of global inequality does not involve double denialism makes it reasonable to hypothesize that it is not quite as hard to tackle as the other moral mega-problem of ameliorating anthropogenic climate change.

The topic of denialism has become a 'hot' topic recently, and if anyone in particular is to be 'credited' for having made it so, it is the former US president Donald Trump. A well-known example is his persistent denial that he was fairly defeated by Joe Biden in the 2020 US presidential election in spite of the fact that several recounts confirmed Biden's victory. This is a strong form of denialism, a denial in the face of overwhelming evidence to the contrary. Of course, Trump himself would not acknowledge that this evidence is overwhelming, but he is presumably aware that in the eyes of many knowledgeable people it appears overwhelming. Trump's denial manifests an aversion to truth that is clearly motivated by the excessive self-love of a narcissistic personality, to put it in terms borrowed from Pascal. But it is not just self-love that could motivate such strong denial. It is also the love of an in-group with which the denier identifies (see Bardon, 2020: 23-4). This is what motivates the denial of many of Trump's followers that he was fairly and squarely beaten.

A desire or wish, irrespective of whether it has the marks of self-interest or group-interest, that something, *p*, is not true cannot be a *reason* to believe that *p* is not true

in the sense of supporting the falsity of *p*. Rather, it must work by such devious measures as directing deniers' attention away from evidence confirming *p* and towards evidence disconfirming it, or by making them spend time on trying to undermine the confirming evidence.

Now people might not be interested in being morally enhanced because they deny they need to be morally enhanced, deny they fall short morally. As regards many people, this would not qualify as a strong form of denialism because it is not a denial in the face of overwhelming opposing evidence. For these are not people who commit archetypally immoral acts, like criminal acts that are punishable by prison sentences, which would supply powerful evidence that their agents are immoral. But a denial that their behaviour is as morally wrong as it actually is can still be motivated by a desire or wish about how things should be or be an instance of wishful thinking. For many people exhibit the so-called *overconfidence bias*: the tendency to think that they are better than they actually are in various respects. For instance, a much-quoted study found that 93% of American drivers believe that they are better drivers than the average (see Svenson, 1981). They may be even more strongly inclined to believe that they are *morally* better than the average because this is an asset that is more important than driving for most people. When human beings set themselves apart from other animals and cherish beliefs such as that they are made in the image of a god, it is such properties as their capacity for being moral – along with their capacity for being rational and for creating art, etc – that they are prone to emphasize.

2. Factors Determining the Obviousness of Wrong-Doing

It is obviously easier to believe that you are a good driver if you have not been responsible for any more significant traffic accident. Similarly, if you have not been responsible for any actions that are archetypally morally wrong, it is easier to believe that you are morally good. Acts that are archetypally morally wrong are acts that evidently cause great harm to people without justification. They are acts that 'ordinary decent people' are unlikely to perpetrate, for instance, acts of violence like punching somebody in the face without good reason, such as this person posing a serious threat to someone.[1] Our acts that contribute to harmful climate changes are not *flagrantly* or *obviously* wrong like this. Let us try to sort out the factors that make the moral wrongness of acts flagrant or obvious rather than so discreet or elusive that it is liable to be overlooked or underestimated.[2]

[1] Here the victims will be assumed to be human, though this should not be taken to imply that harming non-human animals cannot be obviously morally wrong.

[2] The following factors, (1)-(6) and (A)-(F), are largely collected from Persson (2017a). See also Persson & Savulescu (2012: chs. 6-7).

(1) *Temporal proximity between the act and the harm*: the pain and damage to the victim's face occur immediately after the punch. This enables us automatically to associate the harm with the punch. If it instead takes a long time for the harm to occur after an act is done, such an association will not be set up automatically, and we shall feel less uncomfortable about performing the harm-causing act. Partly, this reaction can be explained by the fact that *we are biased towards the near future*: we are more concerned about good and bad events that occur in the near future than in the more distant future. This is why we are relieved when an imminent unpleasant event is postponed, and disappointed when an imminent pleasant event is, even though the postponement does not make it much less probable. By contrast, the harm caused by our CO_2 emissions is temporally very remote. CO_2 can accumulate in the atmosphere for hundreds of years, blocking radiation of heat from the Earth's surface, but letting through sunlight, thereby eventually leading to a harmful increase of the global temperature.

(2) *The victim(s) of the act is (are) identifiable*, that is, identifiable in the sense that witnesses of the act of punching can observe who the victim is. It is a familiar fact that we feel more compassion for individuals who suffer before our very eyes. This is much harder for us to bear than suffering that is merely verbally recounted to us, even if it be the suffering of many more individuals. There is a correlation between this factor and temporal proximity: if the harmful effect of an act we perform is temporally proximate to the act, its victim is often within the eyesight of us, whereas if the harmful effect is temporally distant, this is often not the case. In addition, when the harm is temporally very remote as in the case of climate change, we are normally not personally acquainted with the victims harmed.

(3) *The harm caused is caused by a single agent*: it is a single agent who is dealing the harmful punch, no other agent is involved. Contrast this with the harm of global warming where the harm is caused by several agents acting together, either simultaneously or successively. Common sense conceives of moral responsibility *as being heavily based on causation*, so when causation of harm is spread over several agents, the feeling is that each agent involved is correspondingly morally responsible for less harm.

(4) *Concentration of harmful effects to a single victim* rather than diffusion of the harm over several victims, with the result that each suffers merely a fraction of the total harm caused by the agent. Such a diffusion makes each agent feel that he or she has acted less wrongly than they would have if they had caused this quantity of harm to a single victim on one occasion, even though the total sum of the small bits of harm they have caused to many victims is as big. Each agent's contribution to climate harm is typically of this kind: minimal or negligible harm to innumerable victims.

(5) *Perspicuity of the causal process*: the causal connection between a punch in the face and pain and facial injury is so perspicuous that everyone could grasp it. How CO_2 emissions cause harmful climate changes is of course a much more complicated matter. It takes so much of science to understand that it has only

recently been understood by experts, and most of humanity still lacks this understanding. Moreover, a lot of this more precise knowledge is still missing.

(6) *The harmful act is an act out of the ordinary*: acts like punching someone in the face are not a sort of acts that most of us perform regularly or routinely. By contrast, many of us have driven our cars daily for years and years and got accustomed to the idea that there is not anything wrong about it. The fact that we and others around us have got into the habit of doing something routinely and regarding it as permissible makes it hard for us to take to heart an intellectual realization that these acts involve so much harm that they are in fact wrong, and as a result abstain from them. Habit and conformism make us blind to the moral wrongness of status quo.

Along these dimensions, then, our emissions of greenhouse gases are at the opposite end to acts like punches in the face: their harmfulness is discreet or unobtrusive rather than flagrant or obtrusive and, consequently, we are spontaneously inclined to ignore or underrate their harmfulness and, so, their moral wrongness. The overconfidence bias has an easy time persuading us that they are not morally wrong at all, and our self-esteem can remain intact.

It is plausible to hypothesize that evolution has programmed us to adopt moral aversion towards such flagrantly harmful acts as punching people in the face because they are actions that have been elements of our behavioural repertoire throughout the hundreds of thousands of years of our evolution, and their consequences have been invariably the same. But the causation of harm by the emission of greenhouse gases is a recent addition to this repertoire since it presupposes advanced technology. Therefore, it is not surprising that they could be harmful and wrong without this being obvious to us.

With the possible exception of (3) and (4), it seems on reflection uncontroversial that none of the six factors affects how harmful an action *is*; they only affect how harmful it *appears to be*. But (5), the elusiveness of the causal link between our emissions and climate changes facilitates doubt that there *is* such a link and, thus, buttresses climate science skepticism. This lets in a double denialism. Also, (1) the temporal remoteness of climate changes opens the door to wishful thinking to the effect that there will in time be means to prevent any possible harm caused by our emissions. Thus, we can conveniently continue to reap benefits from our use of fossil fuel with a clean conscience.

3. Factors that Make Collective Action to Fight Climate Change Hard

Now, effectively fighting global warming requires coordinated action from people worldwide and for decades into the future. But, unfortunately, the fact that the wrongness of the emission of greenhouse gases is so discreet or unobtrusive makes

such coordination harder to accomplish because people fail to realize the wrongness of their behaviour. A familiar illustration of cooperation problems is *the tragedy of the commons*. It consists in the herdsmen of a village trying to agree on restrictions on the grazing of their cattle to avoid overgrazing of the commons, and subsequent starvation for the herdsmen and their families. There is a problem of establishing cooperation here since, although every one of the herders has a self-interested reason to cut down on the grazing of their own cattle as a means of preventing overgrazing – which will ultimately inflict starvation on them and their families – they are likely to have a stronger self-interested reason not to do so. They might hope that a sufficient number of the other herdsmen will reduce the grazing of their cattle, and free ride on this reduction without making any reduction themselves. This strategy has the additional advantage that in the event that the other herders by and large decide not to cut down, they have not made any useless sacrifices of their own welfare. But, obviously, if all or most of them reason and behave in this way, the collective grazing will not be reduced sufficiently to avoid overgrazing and eventual starvation, which is bad for all of them.

There are however significant disanalogies between this model and the problem of reducing global CO_2 emissions which make the latter a more pernicious cooperation problem.

(A) *Cooperation to reduce effectively CO_2 needs to be more or less world-wide, involving at least bigger nations which are significantly different from each other*. A global agreement is clearly harder to establish than an agreement in a village in which everyone knows everyone else and shares the same ethnicity and culture. This sharing is something that facilitates the growth of altruistic concern and trust among the herders. By contrast, there are deep ethnic, cultural, and political differences between many of the biggest countries of the world, countries like the USA, China, India, and Russia. Some of them also have long histories of war and conflict with each other. As a result, there will be minimal fellow-feeling between them and trust that any costly agreements will be kept.

(B) *The immense differences between the world's nations regarding their level of welfare, or GDP, and their level of CO_2 emissions per capita*. In the tragedy of the commons model, the herdsmen are thought to be roughly equally well off, have a roughly equal number of cattle whose grazing needs to be reduced, and have equally many dependents to feed. This makes it comparatively easy for them to agree on what is required of each and every one: they should divide equally among themselves the cut-downs of the grazing necessary to attain sustainability. The enormous differences in welfare between the world's richest and poorest nations rule out such a simple solution with respect to combatting climate change. These welfare differences make it reasonable to demand that richer nations pay more for measures to reduce the future level of CO_2 in the atmosphere because of their greater ability to pay, and this is likely to generate disagreement about how much more they should pay. This is something that has manifested itself in international negotiations.

A related problem is that the per capita rates of emissions of the big emission countries differ greatly, and this may be so even though their total amount of emissions may be more equal because the size of their populations differs. To illustrate, consider the two countries that emit most CO_2 in the world, China and the USA. The population of China is around four times as large as the population of the US, but its total sum of annual emissions is only roughly twice as large as that of the US, so the per capita emission of the US is roughly twice as high as that of China. This difference is clearly something that might make it tricky for them to agree on what emissions each should be allowed.

(C) *The historic record of CO_2 emissions differs between the more and the less developed nations*. Again, this can be illustrated by a comparison between China and USA: since 1850 USA has emitted more than twice as much of the CO_2 put by human activity in the atmosphere as China has. This might motivate the Chinese to propose that, based on their more modest historical record, they have a right to a per capita rate of emissions in the future that is considerably higher than that of the US.

(D) *The degree to which different countries of the world are harmfully affected by anthropogenic climate change varies widely*. Some countries are likely to suffer devastating damages, while other countries may stand to gain rather than lose by expected climate changes. Great losers are low-lying countries like Bangladesh, the Netherlands, and South Sea Islands – that run serious risks of being inundated by rising sea levels – and regions in Sahel, Australia and the south-west of USA that will probably be exposed to severe droughts and desertification. Geographic regions which may enjoy salutary effects are some northern countries, like Russia. Obviously, the losers have more of an incentive to implement a reduction of emissions of CO_2 than the winners.

Furthermore, it should be noticed that even in countries which are expected to be comparatively severely hit by global warming, the worst effect will not be suffered by the *present* generation, who is making decisions about climate policies, or perhaps even by their children, but by generations further into the future. This is because climate change is such a slow process. Thus, these decision-makers are asked to make sacrifices for people who are to a great extent beyond the range of their limited or parochial altruism. Due to the bias towards the near we are relatively unconcerned about effects in the more remote future even when they affect ourselves – that is why, for instance, smokers find it difficult to quit their hazardous habit. Obviously, we are even less concerned about temporally remote effects if they affect others, especially if they are not near and dear to us, which they will not be if those affected are unknown people in the distant future or in distant countries.

(E) *Controls of compliance are lacking with respect to global treaties to reduce CO_2 emissions*. It is unlikely that there will be an effective surveillance of whether countries over decades will comply fully with treaties to reduce their CO_2 emissions they have entered. And if they are found out to have defected, there will probably be no effective sanctions to apply. Such checks and sanctions are surely necessary for there to be a reasonable guarantee of compliance, since we cannot expect people

all over the world to have much altruistic concern for and trust in each other, for reasons recounted above.

(F) *The effectiveness of current compliance to international agreements to reduce CO_2 emissions relies on the compliance of future agents who are not bound by the agreements*. Cooperation about reducing CO_2 emissions has to extend far into the future in order to be effective in alleviating global warming. But future generations who have not consented to agreements about CO_2 reductions could in virtue of this fact claim that they are not bound by them. Thus, there is a risk that when future generations realize that their standard of living is going down because of the reductions of CO_2 emissions implemented by earlier generations – reductions which may benefit primarily even later generations – they will be inclined to discontinue these reductions. This is especially so, since they may fear that even if they keep them, the following generation will not because they will be subjected to even greater hardships, and they have still greater reason to fear that the generations succeeding them will not keep in line because they will be subjected to yet greater hardships, and so on. Such a chain of growing incentives to defect seems fatal to the possibility of reaching viable agreements. It encourages present decision-makers to 'pass the bill' to future generations who cannot retaliate.

Six dimensions, (1)-(6), have been reviewed along which our greenhouse gas emitting acts are at the opposite pole of acts whose harmfulness is so flagrant or evident that it is hard to deny their wrongness in the absence of justifying factors. Thus, it will be easy for the overconfidence bias to persuade us that it is not morally bad to carry on with these emissions. This is especially so, since discontinuing them would mean a sacrifice of our welfare. Consequently, we are little motivated to enter into agreements about cutbacks of greenhouse gases. And the factors (A)-(F) bring out why effective agreements on such cutbacks present an especially hard coordination problem even in the absence of the factors white-washing emissions.

If we live in democratic societies, we shall be reluctant to give our votes in general elections to political parties that favour cutbacks and, thereby, impose sacrifices of welfare on us. If we are doubtful that it is wrong not to cut down on the emission of greenhouse gases, we are likely to think that others too have such doubts and will be disinclined to agree to cut down on them. Therefore, governments in liberal democracies are unlikely to give priority to efforts to mitigate global warming. The parties that gain and retain power in liberal democracies are more likely to prioritize issues of employment, education, health care, restrictions on immigration, etc. which directly benefit their voters. The realism of these speculations is borne out by the fact that no sufficiently effective action against climate change has hitherto been taken, even though the problem has been on the agenda of organizations like the United Nations for more than twenty years.

4. The Problem of Defeating Denialism

To solve the two moral mega-problems of our time, Savulescu and I have argued in *Unfit for the Future* and many other publications that human beings need to undergo moral enhancement by all potential means, including biomedical ones, in order to make them more altruistic or benevolent. But we have now seen that they are unlikely to recognize that they need moral enhancement because of their overconfidence bias and the discreetness of much of their moral wrongdoing that serves their self-interest. Is it possible to break down the obstacle of denialism?

Reasons for pessimism are provided by the fact that denialism flourishes even with respect to issues where it is up against incontestable evidence to the contrary. For instance, in the USA denialism concerning the spread and the fatality of the covid-19 virus was rampant during the presidency of Trump. Another telling example is that, although evidence that human beings have evolved from animals has accumulated for at least 150 years, more than 70% of US citizens still deny it because it is incompatible with their religious beliefs (see Bardon, 2020: 105). One reason that many of them also find it difficult to take on board the fact of anthropogenic climate change may likewise be that it sits ill with these religious beliefs according to which their God is ultimately in charge of the creation.

Additionally, there is the 'official' denialism typical of totalitarian or authoritarian regimes to contend with. A recent example is the Chinese authorities' denial of the discovery of the covid-19 virus in Wuhan in December 2019. The consequences of this cover-up have been disastrous. Had the Chinese authorities acknowledged the occurrence of this virus and taken action against it as soon as it was detected, a world-wide pandemic might have been nipped in the bud. It should however be noted that when denialist political leaders indoctrinate citizens to believe falsehoods because it serves the leaders' interests, indoctrination may be so effective that the citizens will not be conscious of any evidence contrary to the falsehoods and, if so, they are not guilty of denialism.

These are depressing facts. But despite them and Jonathan Swift's insightful remark about the belief-formation: 'Reasoning will never make a man correct an ill opinion, which by reasoning he never acquired',[3] let us consider whether moral philosophy could be of any assistance in the fight against denialism with respect to anthropogenic climate change. Among the factors listed in section 1 as obscuring or bleaching the moral wrongness of actions, some can surely be seen as morally irrelevant on reflection, namely that the harm is temporally and causally distant, that the victims of it are anonymous or members of a big crowds, and that the harmful acts are routinely performed. This is encouraging, though we have observed that people can easily deny the obvious.

[3] 'Letter to a Young Clergyman', quoted by Bardon (2020: 36).

However, it cannot be denied that controversy is very widespread in moral philosophy. To make moral philosophy fully effective against denialism, it would be necessary for it to establish a *rational consensus* about what, in light of the climatological evidence, we morally ought to do to prevent harmful climate change. Moral philosophy is far from reaching this end with respect to any moral issue of importance. On the contrary, it rather tends to expand our moral disagreement by achieving ever greater precision with sharpened conceptual tools. New distinctions are incessantly drawn and, consequently, moral claims are split up into more precise versions. Since some of these claims will be only marginally different, it will be well-nigh impossible for us to reach a consensus about which alternative version is most plausible. We do have a batch of common intuitions about what is morally right and wrong in many situations, for instance, about it being morally wrong to kill, torture, or rape human beings in most circumstances, and it being right to help if we can those who are without any fault or choice of their own much more needy than we are. But they are intuitions that have been developed to help us navigate in small, close-knit societies with a primitive technology which allows us to affect only our immediate environment. These are the circumstances in which human beings have lived in all but a tiny fraction of their history. The intuitions they have fostered are not sufficiently sophisticated and fine-grained to enable them to choose strategies to deal with novel problems like anthropogenic climate change out of the bewildering variety of theories offered by modern moral philosophy.

Philosophical problems often take the form of a conflict between commonsensical intuitions and philosophical arguments challenging them. An example of such a conflict in normative ethics is the dichotomy between consequentialism and deontology. By 'deontology' I mean a type of morality that includes some version of the *act-omission doctrine* – to the effect that it is harder to justify morally harming than omitting to benefit – and/or *the doctrine of the double effect*, declaring that it is harder to justify morally harming someone as a means, or harmfully using someone as a means, to a good end than to harm them as a foreseen side-effect of the good end. Nobody has found any version of these doctrines that satisfies critics. Still, even these critics – which include the present author[4] – continue to feel the tug of these doctrines, a sign of how firmly entrenched they are.

As noted, the act-omission doctrine pops up in the moral mega-problems of present concern. If this doctrine is untenable, affluent countries will be morally required to give aid to less well-off countries to an extent that would be morally supererogatory if this doctrine is sound.

Another moral area in which this opposition to the intuitive and the reflective crops up is *justice or fairness*. An idea that has a good claim to being hard-wired is that there are *property rights* to the effect that we have a moral right to our bodies and what we are the first occupy or appropriate of unowned natural resources, and to what we manufacture out of these resources by our own labour (see Locke, 1690).

[4] For my view of these deontological doctrines, see Persson (2013: chs. 3, 4 & 6).

Property rights support the idea that we are permitted to omit helping others in situations in which we would be obliged to help them if rights were rejected: if the things that could help others are our property, they are by definition something that we are permitted to keep, though others might need them significantly more than we do. Contemporary technology and trade have vastly increased the means to hoard property to the point at which, according to Oxfam, the 2153 richest individuals on Earth owned more than 4.6 billion of the poorest individuals in 2019. Without understanding the strong grip property rights exercise on us, we could not understand how we can put up with a world with such an enormous economic inequality.

Like rights, *desert* is a consideration of justice: it is just that you receive what you deserve in virtue of what you are responsible for, just as it is generally just that you get and keep that to which you have a right, and it would be unjust or unfair to deprive you of it. There are philosophical arguments against the attribution of moral rights and deserts to us which turn on the claim that this attribution presupposes that we are *ultimately responsible* – i.e. responsible for all the features that allegedly make us responsible for anything – and this is something that beings with a finite past like us cannot be. If, on the basis of such arguments, the notions of rights and desert are discarded, a more egalitarian conception of justice will follow, since it can no longer be claimed that it is just that some are better off than others because they deserve to be better off, or have a right to more.[5] But, although such arguments possess enough power to convince many, it must be admitted that the notions of rights and deserts have a firm hold on our minds. Therefore, another ethical impasse appears inevitable.

A third moral area of deep opposition concerns whether in order to be moral our *benevolence or altruism* must be *strictly impartial or permits some partiality*. We are disposed to be concerned about the welfare of individuals roughly in proportion to how well we know them and cooperate with them in mutually beneficial ways. People to whom we are closest to next to ourselves include family and friends; thus, they belong to the inner circle of those for whom we care most. We are comparatively indifferent towards strangers. In between these extremes, there are various strata of individuals with whom we associate in some circumstances and for whose welfare we have somewhat greater concern. There is a plausible evolutionary explanation for such a stratification of our altruism. It would be risky for us to extend altruism and invitation to cooperation beyond those with whom we are well acquainted to strangers who for all we know might be inclined to free-riding and even hostility.

The question how far we are morally permitted to be partial to ourselves and other individuals whose welfare we are spontaneously more concerned about is then another great divide in normative ethics. Such partiality should not be confused with

[5] For such an argument, see Persson (2017: ch. 7). For another type of argument against rights, see Persson (2013: ch. 2).

the special moral obligations we have to those to whom we have made promises, have brought into existence, etc., and who have corresponding moral rights against us; or those who deserve favours in return for favours they have done to us. This is a matter of justice, whereas the partiality now at issue is a matter of benevolence, for instance, the more intense compassion we feel for the suffering of family and friends than the suffering of strangers, or for those who suffer before our eyes than those who suffer faraway and whose identity is unknown to us.

Everyone agrees that the partiality of our spontaneous concern about our own well-being and the well-being of those who are near and dear often goes beyond what is morally permissible. The pejorative force of such terms as 'egoism', 'nepotism' and 'cronyism' is clear evidence of this. But this does not show that it is not morally desirable that our altruism or benevolence towards others is strengthened; it just goes to show that there are features that wrongly block or filter it: features such as that someone is foreign, temporally or causally distant, anonymous or one among many. These are features that, on reflection, most of us would agree are morally irrelevant; the trick is to implement this insight in practice.

Precisely what such shutters of altruism are morally acceptable is however controversial, but those who accept some shutters can concur with utilitarians, who accept none and demand impartiality, that stronger altruism is morally good thing, albeit not required. Those who endorse some deontological morality or a theory of justice which includes rights and desert assent to this as well. For instance, advocates of the act-omission doctrine will claim that we are not morally required to do as much to help the needy as those who reject it, but they should not deny that we are morally good if we do more than is required. The same goes for those who champion rights, for it is often praiseworthy to give away some things to which you have rights. And desert-theorists should concede that it could be praiseworthy to show mercy and punish people less than they deserve and give them more of the good than they deserve.

However, in the discussion of how to extend common-sense morality to cater for mega-problems, attention will be drawn to the grounds of morality, whether they are solid enough to possess enough authority to prop up the sterner demands this extension apparently generates. Human beings tend to be conformists, that is, they tend to act and react as most people around them do. If they have been brought up to act and react in certain ways because these are ways in which the majority of the citizens of their societies have acted and reacted for as long as anyone can remember, they are inclined to be highly respectful of these ways. Historically, this respect has often manifested itself in the attribution of moral norms to gods or deified ancestors who are thought to watch over their observance.

The authority with which these norms have been imbued will however be undercut if they are questioned. It is improbable that meta-ethicists could deliver a replacement for this loss of authority that could put a revised morality on a foundation seemingly as solid as a supernatural one because there is in meta-ethics a divide as deep as the divides we have come across in normative ethics. To be sure,

there are accounts that present moral norms as resting on *objective* grounds, grounds that are external to our subjective states, but these accounts are, and are likely to remain, contested by meta-ethicists who take morality to be something *subjective*, expressive of our attitudes. This debate, too, seems destined to be inconclusive by sparking ever more subtle distinctions that promote confusion and dissension rather than clarity and convergence of opinion. In this way, moral philosophy could erode the authority of morality and, thus, drain our incentive to be more altruistically motivated – unless the meta-ethical debate is too esoteric to seep out of the seminar rooms and affect the general public.

All in all, moral philosophy can contribute little if anything to the solution of moral mega-problems. Still, there is hope of a general agreement that it is morally desirable that human altruism is amplified and extended. Such an agreement is a reason why Savulescu and I have made altruism the main target of our argument for biomedical means of moral enhancement; another reason is it has been the focus of much experimental research into such means. If, additionally, some of the more obviously unacceptable 'shutters' blocking altruism are pulled up and climate science denialism is defeated, then the more devastating effects of climate change may be avoided.

But these conditions are unlikely to be realized in time. The problem of counteracting anthropogenic climate change is probably the hardest moral problem humanity has ever faced. In the words of Tony Leiserowitz, of the Yale Project on Climate Change Communication: 'You almost couldn't design a problem that is a worse fit with our underlying psychology', and Daniel Gilbert, professor of psychology at Harvard, joins in: 'A psychologist could barely dream up a better scenario for paralysis'(quoted from Marshall, 2014: 91). Despite its greater foresight, humanity will probably behave as other reproductively successful species and multiply and consume until their natural resources are exhausted.

References

Bardon, Adrian (2020) *The Truth about Denial.* Oxford: Oxford University Press.

Locke, John (1690) *Two Treatises of Government.* Many editions.

Marshall, George (2014) *Don't Even Think about It.* London: Bloomsbury.

Pascal, Blaise (2006/1669) *Pensées.* The Project Gutenberg E-book.

Persson, Ingmar (2013) *From Morality to the End of Reason.* Oxford: Oxford University Press.

Persson, Ingmar (2017) *Inclusive Ethics*. Oxford: Oxford University Press.

Persson, Ingmar (2017a) "Climate Change – The Hardest Moral Challenge?". *Public Reason*, 8: 3-13.

Persson, Ingmar & Julian Savulescu (2012) *Unfit for the Future.* Oxford: Oxford University Press.

Svenson, Ola (1981) "Are we all less risky and more skillful than our fellow drivers?". *Acta Psychologica*, 47: 143-8.

Goodness and Numbers

Wlodek Rabinowicz[1]

A**bstract**. You can save either David or Peter and Mary. Is there a compelling reason for saving the greater number? Taurek (1977) (in)famously denied it. In providing such reason one might attempt to establish that it is better if more people survive rather than fewer. This would settle the issue for consequentialists, but even non-consequentialists might find it relevant to the question at hand. The standard worry, however, is that such an axiological claim can only be established by aggregating gains and losses of different persons. As opposed to intrapersonal aggregation, interpersonal aggregation might seem illegitimate. Frances Kamm's Aggregation Argument is meant to overcome this difficulty. I consider how her argument is dealt with by Iwao Hirose, Weyma Lubbe and Rob Lawlor, and what is wrong with it from Taurek's own perspective. But then I suggest that this perspective is untenable: While Taurek correctly analyses the concept of 'better' in terms of fitting preferences, he accunts for fittingness appealing to the wrong kind of reasons. Still, even so, Kamm's argument fails, but a closely related argument may well be acceptable. Unlike the former, that argument recognizes that different ordinary lives typically are on a par; they seldom are equally good.

[1] This essay is an offering to Björn, Dan, and Toni. Toni, I trust, will find it stimulating, as it touches on several value-theoretical themes he has written about. But Björn and Dan might also find it interesting. My primary target is John Taurek, and we have all, I think, found his paper provocative, outrageous, and yet fascinating. I hope these Taurekian qualities are contagious enough to give some borrowed color to my own reflections.

1. Introduction

Consider this choice situation: You can save either David alone, or both Peter and Mary. Those you don't save won't survive. You have no special ties or special obligations to any of those three; they are strangers to you and to each other. Saving Peter and Mary is not more costly than saving David. And, anyway, the costs of saving are modest, let us assume, negligible in comparison with what is at stake. But you can't save all three of them; you must choose.

For definiteness, suppose Peter and Mary are stranded on one desert island, and David on another. The islands are isolated, the food reserves are dwindling, and the stranded individuals have no means of escape on their own. If they stay, they will soon die. You can send a rescue ship to one of the islands, but not to both – not in time to save them all. The islands are too far apart.

Is there a compelling reason for saving the greater number in a situation like this? Should you save Peter and Mary rather than David? Is it what morality requires? In his much-discussed paper, "Should the Numbers Count?" (1977), John Taurek (in)famously denied it. What he would do instead would be to give each person an equal chance of survival, by flipping a coin: If it falls Heads, he saves David; if it falls Tails, he saves Peter and Mary. This would best express "my equal concern and respect for each person" (Taurek, 1977, p. 303)

Actually, Taurek considered a somewhat different set-up: one in which you can save either David or *five* others. And those involved aren't stranded on desert islands. Instead, they are seriously ill and will die unless administered a drug that is at your disposal. One of them, David, needs all of your drug to survive; the other five only need one-fifth of the drug each. These differences between the two choice situations don't matter, I take it, for my discussion. I will continue to focus on the two-islands case, though quotes from Taurek will mention five individuals (instead of just two) whose lives will be lost if David is saved.

To support the claim that you ought to save the greater number, one might argue that it is *better* if more people survive rather than fewer. In the case at hand, this would mean that it is better if Peter and Mary survive than if David alone does. An argument for the former outcome being better would suffice for consequentialists: for them, what you ought to do is to bring about the better outcome. But such an argument would also be of interest to many non-consequentialists. For the latter, deontological considerations, such as the requirement of fairness, might dictate bringing about an outcome that is less than optimal. But many of them might view value differences between alternative outcomes as an important factor that needs to be balanced against other factors that are morally relevant to one's choice. For such non-consequentialists it still is important to determine whether it is better if more people survive, even if this factor need not be decisive. It needs to be weighed against demands of fairness that favor giving each individual an equal chance of continued life.

How then can one argue for it being better if more people survive? A standard worry is that this would require aggregating benefits and/or losses of different persons – that such aggregation is needed for an assessment of the overall value of an outcome. Unlike *intra*personal aggregation, *inter*personal aggregation of benefits and losses might seem problematic. This worry comes to the fore in John Rawls's famous insistence on *the separateness of persons* (Rawls, 1999 [1971], p. 167). Taurek has similar qualms. To begin with, he questions the attempts to aggregate the (purported) objective values of persons or persons' lives to some sort of combined objective value:

> […] when I am moved to rescue human beings from harm […], I cannot bring myself to think of them in just this way. I empathize with [each of] them. […] It is not my way to think of them as each having a certain *objective* value, determined however it is we determine the objective value of things, and then to make some estimate of the combined value of the five as against the one.
>
> (Taurek, 1977, pp. 306f)

He also objects to interpersonal aggregation of personal losses (or gains):

> It is the loss to the individual that matters to me, not the loss of the individual. […] Five individuals each losing his life does not add up to anyone's experiencing a loss five times greater than the loss suffered by any one of the five.
>
> (ibid., p. 307)[2]

Given these worries about the legitimacy of interpersonal aggregation, it is not easy to see how it can be shown that it is better if more people survive rather than fewer. Nevertheless, an attempt to provide such an argument was made by Frances Kamm in the first volume of her *Morality, Mortality*. She called it "The Aggregation Argument." Adjusting the names of the persons involved to our two-islands example, it went like this:

> If (1) it is worse if [Peter] and [Mary] die than if [Peter] alone dies […]; and (2) it is equally bad if [David] alone dies or if [Peter] alone dies […]; then (3) by substitution, it should also be worse if [Peter] and [Mary] die than if [David] alone dies.
>
> (Kamm, 1993, p. 85)

To call it the Aggregation Argument may be somewhat misleading if the argument is meant to establish the conclusion without relying on problematic aggregative premisses. And indeed, in her later work, Kamm began to refer to it as "The Argument for Best Outcomes" (Kamm, 2005, and 2007, pp. 32, 51.). I will, however, continue to use its original name, as it is one under which it has become

[2] Similarly, in his posthumously published reply to Parfit (1978), Taurek questions interpersonal aggregation of pains and sufferings: "pains of […] many cannot be meaningfully summed in a way that has moral significance for preference and choice." (Taurek 2021, p. 315)

widely known, and also because its conclusion (if not its premisses) does appear to involve interpersonal aggregation: the loss suffered by one person is taken to be outweighed by the combined gains of two others.

In this paper, I am going to consider the Aggregation Argument in more detail. I will suggest that while this argument is problematic as it stands, it can be modified to make it defensible. But that modified version, as we shall see, has a somewhat limited reach; it does not extend to all cases in which more people can be saved or fewer.

2. The Aggregation Argument

Kamm's argument was given a more precise form in Hirose (2001) (see also Hirose, 2004, and 2015, ch. 7). What follows is Hirose's reconstruction, with some changes of my own. The argument may be understood as depending on two general principles:

> **Pareto**: An outcome *x* is better than an outcome *y* if *x* is better than *y* for some individuals and equally as good as *y* for everyone else.[3]
>
> **Impartiality**: Two outcomes, *x* and *y*, are equally good if there is a permutation *p* on individuals such that, for every individual *i*, *x* is equally good for *i* as *y* is for *p*(*i*).

This formulation of Impartiality differs from Hirose's. In Hirose (201), Impartiality is a principle according to which two outcomes (two "alternatives") are equally good "if they differ only with regard to the identities of the people." And in Hirose (2004), where he refers to Impartiality as "Symmetry", outcomes ("states of affairs") are taken to be equally good if they are transformable into each other by "permutation of personal identities." These formulations work well for outcomes that are fully specified (i.e., for possible worlds) but are problematic if, as in the argument to follow, some of the relevant details are left out from outcome specification. For example, if one outcome is that *I drink wine and you drink beer* while the other is that *you drink wine and I drink beer*, then the former might be better (or worse) than the latter for each of us because of our drinking preferences. Given Pareto, this implies that the two outcomes won't be equally good: the first will be better (worse) than the second. This problem is avoided by the formulation I have suggested. If both for you and for me drinking wine were equally good as drinking beer, then, plausibly, the two outcomes would be equally good. However, relying on my formulation of Impartiality will require adding some supplementary assumptions in the Aggregation Argument.

[3] The name of this principle, "Pareto", might be misleading: In economics, the Pareto condition takes as inputs the outcome preferences of individuals and not how good these outcomes are for the individuals. What an individual prefers may, but need not, coincide with what is good for her.

As here stated, both Pareto and Impartiality ground the impersonal value relations between outcomes – the relations of betterness, period, and of equal goodness, period – in comparisons of how good the outcomes are for different individuals.

In this context, we assume (i) that the compared outcomes involve the same individuals, and (ii) that in these outcome comparisons we bracket non-welfarist considerations. We only consider how good the outcomes are for individuals and either disregard other considerations or take them to be irrelevant for outcome values.[4]

The Aggregation Argument itself has two premisses. The first is implied by Pareto and the second by Impartiality, given some supplementary assumptions. The premisses state impersonal value relations between different outcomes, where an outcome specifies, for each of the three individuals involved, whether that individual survives (+) or dies (-).

Premiss 1: {Peter +, Mary +, David -} is <u>better</u> than {Peter +, Mary -, David -}.

That is, it is better if Peter and Mary survive, while David dies, than if Peter alone survives.

This premiss follows from Pareto, given the supplementary assumptions that it is better for each person to survive, and that a person's survival is at least as good as her death for everyone else, whatever their own fate might be.

Premiss 2: {Peter+, Mary -, David -} is <u>equally as good</u> as {Peter -, Mary -, David +}.

That is, it is equally as good (or bad) that Peter alone survives as that David alone survives.

This premiss follows from Impartiality, given the supplementary assumptions that it is equally good (or bad) for David to be a lone survivor as it is for Peter, and that it is equally bad (or good) for Mary that Peter alone survives as that David alone survives. (Remember that all three individuals are strangers to each other.)[5]

The conclusion of the Aggregation Argument is that it is better if Peter and Mary survive while David dies than if David survives while Peter and Mary die.

Conclusion: {Peter +, Mary +, David -} is <u>better</u> than {Peter -, Mary -, David +}.

---

[4] This welfarist restriction on value comparisons between outcomes will be retained until the last section, where I will very briefly broach the possibility that outcome values might partly depend on considerations of fairness.

[5] The first assumption (as indeed Impartiality itself) presupposes that interpersonal comparisons of how good or bad an outcome is for different individuals are meaningful. But even if we accept this, as I think we should, we might well wonder whether survival has the same value for different individuals. Doesn't this depend on how good their continued lives would be for them? I will return to this issue later, in the final section.

To derive Conclusion from Premisses 1 and 2, we only need to assume that

> **Betterness is transitive across equal goodness:** For all outcomes *x*, *y*, and *z*, if *x* is better than *y*, and *y* is equally as good as *z*, then *x* is better than *z*.[6]

Suppose we let at least as good as be our primitive relation. Intuitively, we understand it disjunctively: to be at least as good is to be better or equally good. These two disjuncts can then be defined in terms of our primitive relation as, respectively, its asymmetric and symmetric parts:

> *x* is better than *y* iff *x* is at least as good as *y*, but *y* is not at least as good as *x*.
> *x* and *y* are equally good iff *x* is at least as good as *y*, and *y* is at least as good as *x*.

If we now assume that at least as good is a transitive relation, then it follows that both betterness and equal goodness are transitive and that betterness is transitive across equal goodness.[7]

Using suitable supplementary assumptions together with Impartiality and Pareto, the Aggregation Argument can be generalized. We can obtain the conclusion that for any two unequally large disjoint groups of people, it is better if everyone in the larger group survives, while everyone in the smaller group dies, than vice versa. It is better if more people survive rather than fewer.

As suggested above, the conclusion of the Aggregation Argument appears to involve interpersonal aggregation. But what about the argument's premisses? Hirose denies that any of the premisses is "aggregative". The principles on which they are based are not aggregative either:

> Neither Pareto nor Impartiality aggregates the claims of […] different people.
> (Hirose, 2001, p. 341; cf. also Hirose 2004, p. 68)[8]

---

[6] Hirose (2001) doesn't explain how the argument's conclusion follows from its premisses. Probably he thinks it is obvious. Kamm (1993) is a little more explicit: she states that the conclusion follows "by substitution".

[7] Hirose's formulation of the Aggregation Argument differs from Kamm's. In the latter, the outcomes were only partially specified. For example, Kamm's first premiss stated that it is worse if Peter and Mary die than if Peter alone dies; David's fate wasn't mentioned. Also, Hirose's first premiss is not about an additional person dying being worse, but about an additional person surviving being better. But these differences don't matter much. We can re-formulate Kamm's original argument in terms of more fully specified outcomes, with one premiss based on Pareto and the other on Impartiality. Kamm's Pareto-based premiss then states that it is worse if Peter and Mary die, while David survives, than if only Peter dies, while Mary and David survive. The Impartiality-based premiss states that it is equally good (or bad) that Peter dies, while David and Mary survive, as that David dies, while Peter and Mary survive. Since 'worse', like 'better', is transitive across equal goodness (given the transitivity of 'at least as good'), we can draw the conclusion equivalent to Hirose's: It is worse if Peter and Mary die, while David survives, than if David dies, while Peter and Mary survive.

[8] While Hirose denies that Impartiality and Pareto are aggregative principles, he considers them both as necessary for aggregation – as conditions that every aggregative moral theory must satisfy (Hirose,

What he claims isn't obvious. He may be right about Impartiality and the Impartiality-based Premiss 2. If we accept this premiss, it is because we accept that David's survival has the same value as for David as Peter's survival has for Peter, and that David's death has the same disvalue for David as Peter's death has for Peter. Premiss 2 thus doesn't seem to require aggregation of benefits and losses of different persons. But what about the Paretian Premiss 1? According to it, that both Peter and Mary survive is better, period, than that Peter alone survives. This betterness judgment, while eminently plausible, seems to depend on the interpersonal aggregation of the survival benefits that go to Peter and Mary: the survival of both is worth more than the survival of just one of them.

I imagine Hirose could reply as follows. Consider the comparison between the two outcomes in question. Since Peter survives in both, they don't differ as far as he is concerned. But then, in this outcome comparison, we may disregard Peter (just as we disregard David, who dies in both outcomes), and focus on Mary.[9] In one outcome she survives, in the other she dies. The former outcome is thus better for Mary, the only person with regard to which the two outcomes differ, and this is why the former outcome is better, period. If we justify our 'better, period' judgment in this way, we don't seem to rely on interpersonal aggregation.[10]

On this interpretation, then, the Aggregation Argument arrives at what looks like an aggregative conclusion from non-aggregative premisses. But then the conclusion perhaps is not aggregative either, appearances notwithstanding?

3. Lübbe's Critique of the Aggregation Argument

Weyma Lübbe subjected the Aggregation Argument to a scathing critique (Lübbe, 2008; cf. also Lübbe, 2015). This argument compares outcomes in terms of betterness, period, and equal goodness, period. But these notions, Lübbe insists, are not applicable to outcomes ("states of affairs") at all. As opposed to betterness or equal goodness *for* a person, there are no relations of betterness/equal goodness,

2015, ch. 2). This necessity claim is criticized by Gustafsson (2017), who provides examples of intuitively aggregative theories that violate Impartiality and Pareto, respectively.

[9] Indeed, when arguing that the Impartiality-based Premiss 2 is not aggregative, we have already done something similar: we have disregarded Mary, whose fate is the same in the two outcomes that this premiss compares.

[10] This non-aggregative way of arguing for Premiss 1 can be extended to outcomes in which more than one additional person benefits. We can proceed in steps, each time adding a benefit to yet another person. Since by the argument above, each step is an improvement, the transitivity of betterness implies that the last outcome in this sequence is better than the first.

As Iwao Hirose has pointed out (private communication), the two principles that underlie the premisses (Pareto and Impartiality) are both derivable from Leximin, which intuitively is not an aggregative principle.

period, between states of affairs. Likewise, as opposed to goodness-for, there is no property of goodness, period, that accrues to states of affairs. On her view, judgments of betterness, period, or goodness, period, are essentially *moral* in nature and thus only apply to what can be morally evaluated. In particular, instead of states of affairs, they may apply to *choices* that bring these states about.

> When we go beyond 'better for' judgments [to judgments of betterness, period], we do not in fact evaluate states of affairs but choices that bring about states of affairs, the choices of a hypothetical decision maker. We make, more precisely, moral evaluations. States of affairs […] are not the proper objects of moral judgments. They are not to blame, even if they are very bad for the people involved.
>
> (Lübbe, 2008, p. 74)

Lübbe takes Taurek to be her ally in this. A "Taurekian" will be able to say that a state of affairs is better than another *for a person*, but not, Lübbe claims, that it is better, period.

> […] an impartial observer, a Taurekian one […] would not, of course, answer that [a better choice] brings about a better state of affairs. […] 'better, period' judgments evaluate choices, not states of affairs […]
>
> (Lübbe, 2008, p. 75)

It is debatable whether she is right in her reading of Taurek. The next section will consider this issue in more detail.

So, how does the Aggregation Argument fare if betterness and equal goodness, period, can be predicated of choices, but not of states of affairs? To answer this question, the argument needs to be re-formulated to make choices the objects of evaluations:

> **Premiss 1*:** To choose that Peter and Mary survive (while David dies) is <u>better</u> than to choose that Peter alone survives.
>
> **Premiss 2*:** To choose that Peter alone survives is <u>equally as good</u> as to choose that David alone survives.
>
> Therefore, because betterness is transitive across equal goodness,
>
> **Conclusion*:** To choose that Peter and Mary survive (while David dies) is <u>better</u> than to choose that David alone survives.

What should we say about this 'choice version' of the Aggregation Argument? Lübbe considers Premiss 1* to be highly plausible, and she thinks Taurek would concur. But what about Premiss 2*? Evaluative comparisons of choices may well depend on what other choices are at the agent's disposal. This has repercussions for Premiss 2*. A moment's reflection suffices to realize that this premiss is glaringly

false. It would be outrageous to choose to save Peter alone, letting Mary die, when we can save her along with Peter. But there is no such outrageous omission if we choose to save David alone. There is no one whom we can save along with David.

> […] in choosing Y [= the survival of Peter alone] we decide deliberately to watch Mary die and waste a resource that could have been used to save her, while in choosing X [= the survival of David alone] we do no such thing […] For Y there is a Pareto improvement open to choice […]. For X there is none.
>
> (Lübbe 2008, p. 80)

Thus, to choose that Peter alone survives is not equally as good as to choose that David alone survives. It is *worse*. But then, in this choice version, the argument crumbles. While choosing that both Peter and Mary survive (*z*) is better than choosing that only Peter survives (*y*), choosing that only David survives (*x*) is also better than *y*. Therefore, we cannot reach the conclusion that choice *z* is better than choice *y*.

Lübbe writes:

> I conclude that [… a] Taurekian will deny the second premise. Note that he would not deny it if X and Y were the only open choices. (ibid.)[11]

At this point, one might demur: Isn't it rather natural to interpret choice comparisons as implicitly presupposing that the hypothetical decision maker must make one of these choices? That they are the only ones that are open?

I am not sure that we do presuppose it, but suppose we do. Then Premiss 1* would still be plausible: if we can either save Peter and Mary or just Peter, choosing to save both would be better. And Premiss 2* would now also be plausible: if we can only save one person, Peter or David, Lübbe believes that either choice would be equally good.

Would this suggestion save the choice version of the Aggregation Argument? It would not; its conclusion would no longer follow from the premisses. The reason is that, on this interpretation, the two premisses refer to different choice situations. When applied to choices, betterness must be transitive across equal goodness only if the choice situation is held *constant*. If *z* is a better choice than *y* in one choice situation, and *x* is equally as good as *y* in another, in which *z* is unavailable, then there is nothing to guarantee that *z* is better than *x* when both *x* and *z* are open to the agent.

[11] But what if another person, Ellen, were stranded on the same island as David and we could save her along with David? Then Premiss 2* would become plausible. Choosing that only David survives would be equally outrageous, equally as bad, as choosing that only Peter survives. Lübbe should be willing to accept the choice version of the Aggregation Argument for this case and conclude that it would then be better to choose to save the greater number: better to save Peter and Mary than David alone.

Thus, this attempt to save the choice version of the Aggregation Argument fails. Lübbe is right; if judgments of betterness and equal goodness, period, are applicable to choices but not to the states of affairs these choices bring about, the Aggregation Argument doesn't go through; either one of its premisses is false, or its conclusion doesn't follow from the premisses.

4. Taurek's view

But did Taurek hold the view ascribed to him by Lübbe? Did he think that judgments of betterness, worseness, or equal goodness, period, do not apply to states of affairs?

It might seem that he did. Consider the following quotes (remember that Taurek considers a situation in which we can save either David or five other persons):

> The claim that one ought to save the many instead of the few was made to rest on the claim that, other things being equal, it is a worse thing that these five persons should die than that this one should. It is this evaluative judgement that I cannot accept. […] I do not wish to say this unless I am prepared to qualify it by explaining to whom or for whom or relative to what purpose it is or would be a worse thing.
>
> (Taurek, 1977, pp. 303f)[12]

> Some will be impatient with all this. […] They will insist that I say what would be a worse (or a better) thing, period. It seems obvious to them that from the moral point of view, since there is nothing special about any of these six persons, it is a worse thing that these five should die while this one continues to live than for this one to die while these five continue to live. It is a worse thing, not necessarily for anyone in particular, or relative to anyone's particular ends, but just a worse thing in itself.
>
> (Taurek, 1977, p. 304)

To this Taurek responds:

> I cannot give a satisfactory account of the meaning of judgments of this kind. (ibid.)

In his paper on Taurek, Rob Lawlor suggests that Taurek for this reason would also "simply deny" the Pareto-based premiss of Kamm's Aggregation Argument: "he will deny that [Peter] and [Mary] dying is worse than [Peter] alone dying." (Lawlor, 2006, p. 152)

[12] Christian Piller, who is sympathetic to Taurek's views, puts these words in David's mouth (paraphrasing Taurek, 1977, p. 299): "What do you mean when you say that it would be worse if the many died than if I died? It would be *worse for me* if you saved them and it would be *worse for each of them* if you saved me. That's all there is to it." (Piller, 2014, p. 184) And Piller continues: "If goodness-for points in different directions, there is no resolution of this conflict by any simple appeal to goodness." (ibid.)

Lawlor explains how he thinks Taurek would deal with Pareto-improvements, without falling back on 'better, period' or 'worse, period' judgments – the kind of judgments that "Taurek cannot make sense of" (Lawlor, ibid.):

> […] if it is a choice between [Peter] alone dying or [Peter] and [Mary] both dying, then Taurek will agree that the latter outcome is worse for [Mary], but he will deny that it is worse, period. Taurek will agree that we should save [Mary], even if we can't save [Peter]. But – for Taurek – this is not because two dying is worse than one dying. Rather, we should save [Mary] simply because it is better for [Mary] if [she] lives.
>
> (Lawlor, ibid.)[13]

But do Lübbe and Lawlor interpret Taurek correctly? Does Taurek really question the meaningfulness of 'better, period'-judgments regarding outcomes? Or does he only question whether the judgments such as his *opponents* want him to accept can be given a meaning that would make them plausible?

What strongly supports this second interpretation is that, immediately after declaring that he "cannot give a satisfactory account of the meaning of judgments of this kind," Taurek proceeds, in the same paragraph, to give a general account of what it involves to judge that one outcome is worse, or better, period, than another and how it differs from judging it to be better for a person or a group:

> When I judge of two possible outcomes that the one would be worse (or better) for this person or this group, I do not, typically, thereby express a preference between these outcomes. Typically, I do not feel constrained to admit that I or anyone *should* prefer the one outcome to the other.[14] But when I evaluate outcomes from an impersonal perspective (perhaps we may say from a moral perspective), matters are importantly different. When I judge that it would be a worse thing, period, were this to happen than were that to happen, then I do, typically, thereby express a preference between these outcomes. Moreover, at the very least, I feel constrained to admit that

[13] Lawlor here refers to Kamm's original formulation of the Aggregation Argument. Undoubtedly, he would say the same about the Paretian Premiss 1 in Hirose's version. On Lawlor's view, Taurek would deny that Peter and Mary surviving is better than Peter alone surviving. Rather, we should also save Mary if we save Peter simply because it is better for Mary if she lives.

[14] Does this mean that, for Taurek, 'better for' judgments are purely descriptive – devoid of any normative force? Possibly. Alternatively, he might treat them as conditionally normative, on the lines later developed by Stephen Darwall. Darwall suggests that what is good or better for a person is what one ought to want for her *if* one cares for her. (Cf. Darwall, 2002, pp. 4, 8, 6-7, 8-9, 26 and 48.) This account may be contrasted with a more categorically normative analysis proposed by Toni Rønnow-Rasmussen. On that analysis, *x* is good for a person *i* iff one ought to favor *x* for *i*'s sake. This normative requirement is not conditioned on one's concern for *i*. (Cf. Rønnow-Rasmussen, 2004, 2007, 2011.) Unlike Darwall's, Rönnow-Rasmussen's proposal is definitely in conflict with Taurek's view.

> I *should* have such a preference, even if I do not. It is a moral shortcoming not to prefer what is admittedly in itself a better thing to what is in itself a worse thing.
> (Taurek, 1977, pp. 304f)[15]

If one reads this passage as an analysis (complete or at least partial) of judgments that one outcome is better, period, than another, it suggests the following hybrid account, with an expressivist and a cognitivist part:

> A judgment that an outcome *x* is better, period, than an outcome *y*
>
> (i) typically *expresses* a preference for *x* over *y*;
>
> and
>
> (ii) *states or implies* that one ought to prefer *x* to *y*.

In clause (ii), "one" is to be understood as "everyone", even though Taurek in the quoted passage only considers what he who makes the judgment should prefer. But he considers it a moral shortcoming not to have this preference with respect to what is "in itself a better thing". Which suggests that it is a preference that everyone should have. This reading is fully confirmed by Taurek's subsequent discussion, which I will consider in section 6.[16]

5. Fitting-Attitudes Analysis

The cognitivist part in Taurek's account of 'better, period' judgments (clause (ii) above) is in line with the *fitting-attitudes analysis of value* (*FA-analysis*, for short),

---

[15] Lawlor, who quotes the same passage, feels compelled to admit that "[i]f we interpret Taurek in this way, he can no longer resist the claim that five dying is worse than one dying by simply insisting that such statements don't make sense." (Lawlor, 2006, p. 304)

See also the following passage from Taurek's posthumously published reply to Parfit (1978):

> For example, I really do think it would be, morally speaking, a better thing if this one person were to suffer some substantial pain if these many others could each thereby be spared an agonizing pain. The one alternative is to be preferred to the other. […] It doesn't matter what role I imagine myself to occupy in this situation. (Taurek, 2021, p. 319)

[16] Cf. also these quotes from Taurek (2021), pp. 319 and 320, respectively:

> "In 'Should the Numbers Count?' I traded on what for me is an inextricable connection between the thought of one thing's being, from a moral point of view, preferable to its alternative, and the thought that one should prefer it. Anyone should, for it is, morally speaking, preferable."
>
> "[…] when one alternative, seen from this impersonal perspective, is judged morally preferable to another, then that is what we should prefer; that's what anyone contemplating these same alternatives should prefer."

according to which to be valuable is to be a fitting target of a pro-attitude. FA-analysis has an attitudinal component (pro-attitude) and a normative component. "Fitting" is the standard term used for the latter, but, in some versions of the analysis, it is replaced by a more generic normative expression, such as "ought", "should", or "has reasons to".[17]

For betterness, the pro-attitude that typically figures in the analysis is preference. This suggestion goes back to Brentano, one of the founding fathers of FA-analysis (see Brentano, 1969 [1889], p. 26). Thus, using "ought" for the normative component, we get:

> *x* is *better*, period, than *y* iff one ought to prefer *x* to *y*.[18]

Equal goodness is accounted for correspondingly, in terms of indifference (equi-preference):

> *x* is *equally as good*, period, as *y* iff one ought to be indifferent between *x* and *y*.

In Rabinowicz (2008), I developed an FA-modeling and taxonomy of binary value relations (see also Rabinowicz. 2012). Exploiting the idea that, for the normative component of FA-analysis, two levels of normativity are available – strong (ought, required, fitting) and weak (may, permissible, not unfitting) – I showed that there are several types of binary value relations that fall outside the standard trichotomy of better, worse, and equally good. Here is the definition of one such type of relation, *parity*:

> *x* and *y* are *on a par* iff one may prefer *x* to *y*, and one may prefer *y* to *x*.[19]

In other words, preferring *x* to *y* is permissible, and so is the opposite preference. The notion of parity will prove helpful in the final section.

Cases of parity typically arise in multi-dimensional comparisons. One item, *x*, might be superior to the other, *y*, on some dimensions and inferior on others. To reach an overall comparative assessment of such items, the relevant dimensions need to be weighed against each other. If there is some latitude in such weighing, *x* might come higher than *y* in the overall preferential assessment given one admissible

[17] For a short history of FA-analysis, see Dancy (2000) and Rabinowicz & Rønnow-Rasmussen (2004).

[18] Does Taurek accept this equivalence? It is not clear (and, even so, he might not accept it as an account of the *meaning* of "better"). In any case, as we have seen in the preceding section, he does accept that it holds from left to right. This will suffice for the discussion to follow.

[19] For this definition, see Rabinowicz (2008), but the idea of parity as an independent value relation is due to Ruth Chang (see Chang, 2002a, 2002b, 2005).

assignment of weights to dimensions and lower given another.[20] We will then have a case of parity: it is permissible, all things considered, to prefer *x* to *y*, but it also is permissible, all things considered, to prefer *y* to *x*.[21] It will then typically also be permissible to be indifferent between *x* and *y*, although I haven't included it in my definition of parity.[22]

Even if I were to include it, it might well be questioned whether my definition conforms to the ordinary usage of "on a par". Hugh Barrett has recently suggested modifications to my definitions of betterness and parity (Barrett, 2022, ch. 1).

To begin with, he suggests a weakening of the definition of betterness. For *x* to be better than *y* it is enough, in his view, if one may prefer *x* to *y* and ought to prefer it or be indifferent. To put it differently, *x* is better than *y* iff one ought to favor *x* at least as much as *y*, and may favor it more than *y*. Unlike mine, this definition allows for cases in which it is permissible to be indifferent between a better item and one that is worse. While weaker than mine, Barrett's definition still implies, as it should, that betterness is an asymmetric relation.

Secondly, he proposes a very weak definition of parity: two items are on a par iff one may be indifferent between them. This definition makes parity a kind of 'rough equality'. Two items that are on a par might be equally good, or one of them might be (somewhat) better than the other. (Remember that on Barrett's weak definition of betterness, indifference between a better and a worse item might be permissible.) As I have defined parity, none of this is possible. On my account, if two items are on a par, then none of them is better than the other, nor are they equally good. Barrett suggests that what I and others (especially Ruth Chang) have been after is not "on a par", but "merely on a par" (i.e., on a par, but neither equally good nor better or worse). He may be right, but for the discussion that follows it doesn't matter whether the way I use "on a par" conforms to the ordinary usage. What does matter is that it is clear what kind of value relation I have in mind.[23]

---

[20] An overall assessment of an item need not be literally a weighted average of its scores on different dimensions. Still, it is a function (whose parameters admit of some latitude) of how well the item does on each dimension.

[21] It need not be like this in all cases of multi-dimensional comparisons. In some, one item might be ranked higher than the other on every admissible assignment of weights to dimensions. (Either because it is superior on each dimension or because it is inadmissible to give much weight to the dimensions on which it is inferior.) Then preferring it would be required. Which would imply that it is better than the other item.

[22] For all I know, there might exist cases in which it is permissible to prefer one item to the other and permissible to have the opposite preference, but impermissible to be indifferent. Still, even if such cases might exist, they certainly aren't typical.

[23] In what follows I will also allow for parity in personal values. How such parity should be understood depends on how we decide to analyze personal value (goodness-for) in the first place. If we find Darwall's account of 'good for' attractive, we could say that the value of *x* for *i* is on a par with the value of *y* for *j* iff those who care for *i* and *j* may, for *i* and *j*'s sake, prefer *x* to *y*, but they may also prefer *y* to *x*.

How can an FA-account of value relations guarantee that betterness and equal goodness are transitive and that betterness is transitive across equal goodness? These transitivity properties of value relations will follow if we impose transitivity as a rationality constraint on permissible preferences and indifferences (cf. Rabinowicz, 2008), or – alternatively – if we re-interpret the concept of preferences in such a way that transitivity will fall out as their conceptually necessary feature (cf. Rabinowicz, 2012). For details, I refer the reader to these two papers.

6. Back to Taurek

If, contrary to Lübbe's suggestion, 'better, period' judgments regarding outcomes are meaningful on Taurek's view, then why does he deny that it is better, period, that more people survive rather than fewer? Why does he deny that it is better, period, if David dies but five others (or two others, or fifty others) survive than if David survives but the others die?

He denies this 'better, period' judgment because he accepts the FA-connection between what is better, period, and what everyone ought to prefer, but denies that everyone ought to prefer that more people survive rather than fewer. In particular, it is permissible for *David* to have the opposite preference:

> I do not think [David] morally deficient in any way because he prefers the outcome in which he survives and the others die to the outcome in which they survive and he dies.
> (Taurek 1977, p. 305)[24]

The same goes for anyone who has close ties to David:

> In a situation where the one person, David, is a friend of mine and the others strangers to me, I do have a preference for the one outcome as against the other, to me a natural and acceptable preference. […] His survival is more important to me than theirs. I would expect them to understand this, provided they were members of a moral community acceptable to me, just as I would were our roles reversed.
> (ibid.)

Let us apply this to the two-islands case. If preferring David's survival, even if it means that Peter and Mary will die, is permissible for some of us (for David and his friends), then it follows that it is *not* the case that

> (every)one ought to prefer {Peter +, Mary +, David -} to {Peter -, Mary -, David +}.

[24] Cf. also Taurek (2021, p. 320: "That [David] prefers the alternative in which he survives manifests to me no moral deficiency in him. I cannot expect that he should prefer the alternative in which they survive and he dies. Hence I cannot give as my reason for sparing the five instead of him the impersonal evaluational comparison that the one outcome is, from a moral point of view, preferable to the other."

But then, given FA-analysis, it is *not* the case that

{Peter +, Mary +, David -} is better than {Peter -, Mary -, David +}.

Thus, Taurek would reject the Conclusion of the Aggregation Argument. He can do it because he would reject the Impartiality-based Premiss 2:

{Peter +, Mary -, David -} is equally as good as {Peter -, Mary -, David +}.

His objection here would be predictable: It is permissible for David (and David's friends) to prefer that David alone survives rather than that Peter alone survives. Consequently, it is not the case that everyone ought to be indifferent between these two outcomes. Therefore, on the FA-account, it is not the case that these two outcomes are equally good. Premiss 2 is false.

The falsity of one of the premisses suffices for the Aggregation Argument to crumble. But it might still be of interest to consider what Taurek would want to say about the argument's other premiss, Premiss 1:

{Peter +, Mary +, David -} is better than {Peter +, Mary -, David -}.

I am not sure what he would say about this Pareto-based claim. He might say that for people for whom Mary is a stranger, it is permissible to be indifferent between Peter and Mary surviving and Peter alone surviving. This would imply, on his FA-account of betterness, that the former outcome is not better than the latter.[25]

But Taurek might instead say that it would be a moral shortcoming to be indifferent to Mary's plight even if she is a stranger. Pro tanto, one should prefer outcomes that are better for others, especially if they are so dramatically better for them as when it is a matter of life and death. Arguably, this is required by the fundamental respect we owe to other persons.

Would this suffice to guarantee the truth of Premiss 1? What about the *enemies* of Mary? Is it impermissible also for them not to prefer that she survives along with Peter? Would it be a moral shortcoming on their part not to have this preference? As long as this is unclear, Taurek's assessment of the Pareto-based Premiss 1 also remains unclear.[26]

In this section, I have sketched what I take to be Taurek's view regarding the Aggregation Argument. Now let me consider how we can resist it.

[25]Though, note that on Barrett's weakened definition of betterness, the permissibility of such indifference would not yet falsify Premiss 1. The latter would still be correct if it also were permissible to prefer that Mary survives, along with Peter but impermissible to have the opposite preference.

[26] Kamm (1993, p. 97, fn. 12) reports that in a conversation she had with Taurek the latter said he would go so far as to accept Pareto. But to derive Premiss 1 from Pareto we had to assume that there is no one for whom it is better that Mary dies.

7. Reasons of the Wrong Kind

In Rabinowicz & Rønnow-Rasmussen (2004), we posed and (unsuccessfully) attempted to solve a worrisome problem for FA-analysis: Some of the reasons for a pro-attitude (or a con-attitude) towards an object might have no bearing on that object's value (disvalue). Such reasons are of the 'wrong kind' as far as the FA-analysis is concerned: their presence doesn't make the object valuable (disvaluable). Here is a dramatic example: A powerful demon demands that we admire him; if we don't, he will destroy the world. His determination (and capacity) to destroy the world unless we comply is a very strong reason for admiring him, but it does *not* make him admirable. Thus, it may be that one ought to have a pro-attitude towards an object (in this case, the demon) even though the object itself lacks value.

Some philosophers have argued that examples like this don't pose a serious obstacle for FA-analysis. One line of resistance is to deny that we do have a reason to admire the demon. What we have is a reason to *desire* that we admire him and reasons to bring it about.[27] This only shows that admiring the demon has value, which is uncontroversial given his threat. Another line of resistance is different: Even if we might have a reason to admire the demon, to admire such an evil creature would not be *fitting*. And on the FA-analysis, an object x is valuable only insofar it is fitting to have a pro-attitude towards x. It isn't enough, on this interpretation of the FA-account, that one ought to have a pro-attitude towards x for x to be valuable. Fittingness is taken to be a distinctive deontic concept that differs from a mere Ought.[28]

While each of these two maneuvers would help to disarm the demon example, I doubt whether they could get FA-analysis entirely off the hook. Unless an attitude's fittingness is understood as its adequacy to the object's value, which would make the analysis of value in terms of fittingness viciously circular, it is not difficult to provide examples of cases in which it intuitively *is* fitting to have a pro-attitude towards an object that is not valuable or a con-attitude towards an object that is not disvaluable. In such cases, it also seems perfectly intuitive to say that we have reasons for these attitudes and not merely reasons for desiring to have them (and for bringing them about. Thus, think of a rather indifferent poem composed by your teenage daughter. It may well be fitting for you as a parent to admire it; not just to pretend admiration. You would be a worse parent otherwise. And the reason to admire this poem is that it is your daughter's creation. Or think of Bernard Williams's example of a lorry driver who runs over a child, despite taking all sorts of precautions. It is fitting for the driver to feel guilt, even though he knows he isn't blameworthy; the child's death

[27] For this view, see, for example, Gibbard (1990), p.37, Parfit (2001, 2011, Appendix A), Skorupski (2007, 2010).

[28] For such a "fittingness first" approach to FA-analysis, see McHugh & Way (2016, 2022) and Howard (2019).

wasn't his fault. Still, there would be something morally wrong with him if he didn't feel guilt and remorse. There is a reason for him to feel guilt: he caused the child's death, however inadvertently. (Cf. Williams, 1981.)

One might put this worry about the effectiveness of the appeal to fittingness as follows: In some cases, what makes attitudes fitting are not the value-making features of their *objects*, but the deontological constraints on the *subjects* of the attitudes: constraints that require the subjects to have these attitudes. From the point of view of FA-analysis, such deontological constraints give rise to reasons of the wrong kind.[29, 30]

The difficulty I have just sketched is a problem for the FA-analysis of value properties, but a similar problem arises for value relations. Thus, it might be that an item *x* is better than another item *y* but we still ought to prefer *y* to *x*: we might have strong reasons for doing so and it might be fitting. Thus, I ought to prefer my daughter's poem to a better poem of her classmate. As a father, I have a reason to prefer it, and it may well be fitting.

In other words, some reasons for preference might be of the wrong kind from the point of view of FA-analysis. They are the kind of reasons that this analysis needs to bracket – exclude from consideration. Which doesn't hinder, of course, that they might be perfectly good reasons, as such, and that the attitudes grounded in such reasons might be required and fitting. But such reasons don't make the preferred outcome better or the dispreferred outcome worse.

In what follows, I refer to them as "WK-reasons" (short for "reasons of the wrong kind"). Since such reasons must be bracketed in the FA-account, we need to add an appropriate proviso to the analysis. As applied to the value relations, this analysis should be framed roughly along the following lines:

> *x* is <u>better</u> than *y* iff, with WK-reasons bracketed, there are conclusive reasons for preferring *x* to *y*.
>
> *x* and *y* are <u>equally good</u> iff, with WK-reasons bracketed, there are conclusive reasons for being indifferent between *x* and *y*.

---

[29] Cf. Crisp (2008), pp. 260f. Crisp emphasizes the role of deontological constraints in generating some of the reasons of the wrong kind.

[30] This problem might not arise if one instead of fittingness appeals to the concept of c*orrectness*. Fitting attitudes might still be incorrect if what makes them fitting are deontological constraints. (Think of my admiration for my daughter's indifferent poem, or of the lorry driver's feelings of guilt.) And indeed, Brentano's original formulation of FA-analysis was framed in terms of correct ("richtig") pro-attitudes (Brentano, 1969 [1889], p. 18). His approach has been recently revived in Danielsson & Olson (2007). The worry is, though, that correctness does better than fittingness only because it is so natural to understand the correctness of a pro- or con- attitude as its adequacy to the object's value. As this would make the correctness version of FA-analysis circular, Danielsson and Olson declare that correctness is a primitive, unanalyzable concept. How plausible is this? To seriously assess their proposal would, however, take us too far from the main topic of this paper.

The 'wrong kind of reasons' problem (the WKR problem, for short) is the problem of defining WK-reasons, in a non-circular way. Obviously, if they need to be explicitly excluded from consideration in the FA-analysis of value, it wouldn't do to define them as reasons for an attitude that do not bear on the value of its object. To provide a satisfactory definition of WK-reasons has proven to be difficult. (Cf. Rabinowicz & Rønnow-Rasmussen, 2004.) But certain types of WK-reasons are easier to identify. When arguing that the death of five persons is not worse than the death of one, David, Taurek appeals to David's *personal* reasons for preferring his own survival to the survival of others: reasons of self-interest. Likewise, he appeals to the personal reasons that David's friend would have for the same preference: reasons of friendship. Such personal, "agent-relative" reasons involve an essential reference to the person who has the reason in question: to her interests, projects, social and institutional roles, personal ties and attachments. (Cf Nagel 1970, 1986.)[31, 32]

It is arguable that all agent-relative reasons are WK from the point of view of FA-analysis of *impersonal* value (goodness, period, and betterness, period). Such an analysis must focus on reasons that are there for *everyone*, and not merely for this or that person. Thus, agent-relative reasons should be bracketed in the analysis.

This allows us to resist Taurek's argument. As we have just seen, in defense of the claim that the death of five persons is not worse than the death of one, David, Taurek appeals to the agent-relative reasons on the part of David and David's friend: reasons of self-interest for the former and reasons of friendship for the latter. As I have suggested in the preceding section, he would also appeal to these reasons when confronted with the Impartiality-based Premiss 2 of the Aggregation Argument. But, if I am right, the appeal to agent-relative reasons is illicit in this context. Such reasons must be excluded from consideration when it comes to judgments of betterness and equal goodness, period. With respect to such impersonal evaluations, they are irrelevant.[33, 34]

---

[31] "If a reason can be given a general form which does not include an essential reference to the person who has it, it is an *agent-neutral* reason […] If on the other hand the general form of a reason does include an essential reference to the person who has it, it is an *agent-relative* reason." (Nagel, 1986, pp. 152f) Nagel's way of drawing this distinction is not unproblematic. For some worries, see Rønnow-Rasmussen (2009).

[32] I will continue to talk about "agent-relative" reasons, since this label is so well-established, even though it is slightly misleading when it comes to reasons for attitudes (rather than actions). Perhaps "subject-relative" would be a more adequate characterization.

[33] In the rescue case, these reasons make it *permissible* for David and his friends to prefer David's survival to the survival of others: these reasons override agent-neutral reasons that would otherwise rule out this preference. They differ from the father's reason for admiring his daughter's poor poem in that the latter reason arguably makes his admiration not simply permissible but positively commendable. Still, in both cases, it is a matter of reasons that need to be bracketed when it comes to impersonal evaluations.

[34] Olson (2009) discusses "the Partiality Challenge" to FA-analysis: "there are circumstances in which some agents have reasons to favour or disfavour some object – due to the personal relations in which

It should be obvious, though, that Taurek himself would reject this suggestion. On his view, David's preference for his own survival, even if it means that others must die, is not morally objectionable. And yet it is based on agent-relative reasons. Consequently, this preference cannot be legitimately disregarded when we evaluate outcomes from the impersonal point of view. But Taurek's critics may retort that the moral permissibility of an attitude does not guarantee its relevance for value judgments. They can here take their lead from Parfit (1978) and agree that there is an "agent-relative permission" for David and his friends to have such preference. While this preference is not morally objectionable, it is justified by reasons that don't bear on impersonal value relations.

8. Back to the Aggregation Argument – Final Treatment

If agent-relative reasons don't count when it comes to impersonal evaluations, if Taurek was wrong on this point, does it mean that the Aggregation Argument now is in the clear? One might well think so. Both premisses of the argument appear to be plausible from the impersonal point of view and its conclusion follows from the premisses. Nevertheless, I find this diagnosis much too quick. The Paretian Premiss 1 is indeed very plausible. But Premiss 2, according to which it is equally as good (or bad) if Peter alone survives as if David alone survives, may well be questioned even if we accept the Impartiality principle on which this premiss is supposed to be based. To derive it from Impartiality, we had to assume that Peter's survival is equally good for Peter as David's survival is good for David. But why should it be so? Surely, how good survival is for a person depends on how his or her continued life would unfold. And David's continued life might well be different from Peter's, perhaps better for David than Peter's continued life would be for Peter. Under these circumstances, Premiss 2 would not follow from Impartiality and the Aggregation Argument would not get off the ground.

I expect protests at this point: The kind of situation we have envisaged in the two-islands case is not meant to involve dramatic differences in how the individuals' continued lives would look like. David and Peter are thought of as people whose prospects do not dramatically differ. That one of them would have a wonderful life if he survived while the other's life would be wretched is simply not in the cards.

---

they stand to the object – without this having any bearing on the value of the object." (ibid., p. 365) Olson considers different ways in which this challenge might be met, one of being close to my proposal: agent-relative reasons for attitudes must be disregarded when it comes to impersonal value. His favorite solution is different; it is based on his joint work with Danielsson (Danielsson & Olson, 2007) and their Brentano-inspired analysis of value in terms of correct attitudes. (The idea being that agent-relative reasons don't bear on the correctness of attitudes.) Yet another solution is proposed in Zimmerman (2011). Here, I abstain from discussing these alternative solutions.

What we envisage is a case in which the continued lives of David and Peter would be as ordinary lives use to be.

I agree. Nevertheless, the assumption that their continued lives would be *equally good* for them seems extremely unrealistic. Ordinary lives differ from each other in a multitude of ways. One life has highs and lows that the other lacks, and vice versa. It is superior to the other life in some respects and inferior in others. And, typically, these comparative advantages and disadvantages do not exactly balance off. Rather than being equally good, different ordinary lives are *on a par* when it comes to their personal value. David's continued life would have value for David that is on a par with the value Peter's continued life would have for Peter. If this is right, then it would be reasonable to suppose that from the impersonal point of view we also have a case of parity here.[35]

Premiss 2 should therefore be rejected and replaced by

New Premiss 2: {Peter+, Mary -, David -} is on a par with {Peter -, Mary -, David +}.

In other words, Peter' surviving alone is on a par with David's. One may prefer one outcome to the other or have the opposite preference. And, I suppose, one may be indifferent.[36]

What does this replacement in the second premiss imply for the validity of the Aggregation Argument? The short answer is that Conclusion no longer follows. While betterness is transitive across equal goodness, it is not transitive across parity. Thus, while it is better if both Peter and Mary survive than if Peter alone survives, if the latter outcome is on a par with David alone surviving, we cannot conclude that it is better if Peter and Mary survive than if David alone survives.

That betterness isn't transitive across parity is well known. If two items, x and y, are on a par, then slightly improving (or worsening) one of them, say, replacing x with x^+, typically preserves parity. x^+ is better than x but still on a par with y. Here is an example: Consider two holiday trips to very attractive but also very different locations: one to Peru, with its fascinating remains of ancient civilizations, while the other to Galapagos, with its abundant and unique animal life. Intuitively, the two trips are on a par, and, just as intuitively, this parity would still be preserved if one of the trips, say the one to Peru, were slightly improved (maybe extended with an extra day in Cuzco).

[35] This move from parity in personal values to parity in impersonal value is plausible in the case we consider, but we shouldn't accept, as a general principle, that any two outcomes must be on a par if they are on a par for every individual for whom they differ. This general principle might seem intuitive, but it is vulnerable to an objection related to the well-known problem of 'opaque sweetening' (originally posed in Hare, 2010; see also Rabinowicz, 2021). Here, I abstain from presenting this objection, but see Nebel (2020).

[36] Here, I interpret "on a par" in accordance with my definition. But the argument that follows would also go through if parity were interpreted as rough equality, as suggested, for example, by Qizilbash (2007) and Barrett (2022).

Does it mean that we are stuck – that we cannot reach the desired conclusion in the Aggregation Argument? No, I don't think so. Note that the examples such as the one with two holiday trips involve *small* improvements: a small improvement of one of one item typically preserves parity. But it is natural to expect that a *large* improvement will not preserve parity; it will make the improved item better than the other item in the pair. If, instead of the original trip to Peru, we were offered an extended trip to both Peru and Patagonia, this would beat Galapagos! It thus seems that

> if x and y are on a par, and x^{++} is much better than x, then x^{++} is better than y.
>
> (Whether x^{++} is much better than y is another matter. Perhaps not.)

This restricted form of transitivity of betterness across parity saves the Aggregation Argument. That both Peter and Mary survive is much better than that only Peter does. The latter is on a par with David alone surviving. But then we can draw the conclusion that it is better (though perhaps not much better) if both Peter and Mary survive than if David alone survives. We are home![37]

Note that this solution can also help in explaining our intuitions regarding some large-scale rescue cases. Consider a two-islands case in which there are thousand people stranded on one island and slightly more than a thousand on the other. In this

---

[37] Mozaffar Qizilbash calls the following feature "the mark of parity":

> "if two states of affairs are on a par, a significant improvement (worsening) of one makes it better (worse) than the other, while small changes in value do not make one better than the other." (Qizilbash, 2005, p. 423. See also Qizilbash, 2007.)

By contrast, Christian Piller questions (in private communication) the validity of this restricted form of transitivity of betterness across parity. I think this issue is complicated. Parity is gradable; it can be closer to or further away from equal goodness. The further two items on a par are from being equally good, the larger improvement it takes for the improved item to become better than the other item in the pair. And vice versa: the closer they are to being equally good, the smaller improvement is needed. (For a suggestion how to interpret such closeness and its degrees, see Hájek and Rabinowicz, 2022. For another interpretation, see Chang, 2016, last section.) In the case at hand, the argument goes through if the continued lives of David and Peter, while being on a par, are relatively close to being equally good: close enough for the addition of Mary's survival to that of Peter to be a sufficiently large improvement.

There's another issue I'd like to mention here. In a paper with Toby Handfield, I have argued that spectrum arguments for various counter-intuitive conclusions can be blocked by positing incommensurability in some intervening steps of the spectrum sequence, but that this incommensurability must then be *persistent*: it must persist when we continually improve the next item in the sequence in the dimension on which it has an advantage over its predecessor. If the kind of incommensurability that needs to be posited is parity, then such persistency would appear to be a counterexample to the view that parity is not preserved if one of the items that are on a par is considerably improved. (Cf. Handfield and Rabinowicz, 2018, Herlitz, 2020, and Rabinowicz, 2022.) However, I am not convinced that this persistency phenomenon is a genuine counterexample to the "mark of parity". It is such a counterexample only if these continual improvements of the next item in the spectrum sequence can eventually add up to a sufficiently large total improvement. But whether they can or cannot is unclear.

case, it is no longer intuitively obvious that we should opt for saving the (slightly) greater number. We might instead consider tossing a coin to decide which group to save. Our intuitions can be explained along the lines sketched above: If there were equally many people on both islands, thousand people on each, the survival of one group would be on a par with the survival of the other. There are in fact a few more people on the other island. The survival of that surplus, along with the thousand others, makes things better. But not much better. When a thousand people survive, adding a few more survivors is, relatively speaking, a small improvement. And we already know that small improvements typically preserve parity. Consequently, we can conclude that in this large-scale case the Aggregation Argument fails: if the groups are large and don't differ much in size, the survival of one group is on a par with the survival of the other. But then we can just as well toss a coin when deciding which group to save.[38]

Indeed, tossing a coin might well be what we ought to do in this large-scale case. There are strong deontological considerations – considerations of fairness – that favor giving each individual an equal chance of survival. Which tossing a coin, when deciding which group to save, will achieve. If we instead simply send the rescue ship to the island on which the slightly larger group is stranded, we treat the individuals on the other island unfairly: we don't even give them a chance to survive. In a small-scale rescue case, in which only one individual, David, is treated unfairly if we decide to save Peter and Mary, this unfairness is outweighed by the significant improvement in the outcome: 2 lives are saved while the expected number of lives saved if we toss a coin is 1,5 (0,5 x 1 + 0,5 x 2). (Cf. Broome, 1998.) But, as Lawlor (2007) points out, in the large-scale case, if we instead of tossing a coin decide to save the slightly larger group, there will be a thousand individuals who are left to die on the other island without being given a chance to survive. They are all treated unfairly. At the same time, the gain in saved lives, as compared with the expected number of lives saved if we instead were to toss a coin, is very limited, possibly not larger than the corresponding gain in the small-scale rescue case. Consequently, in this large-scale rescue case, fairness wins: randomization is required.[39]

An argument similar to Lawlor's was earlier presented by Hirose (2004).[40] But Hirose, unlike Lawlor, takes considerations of fairness to affect the very value of an

[38] A coin toss may be seen as a lottery on outcomes among which it decides. And given my analysis of parity, the expected outcome of a lottery on outcomes that are on a par must itself be on a par with these outcomes.

[39] Lawlor assumes that the relatively small gain in the expected number of saved lives would still make the outcome better in this large-scale case, but that this improvement in the outcome value is outweighed by so many individuals being treated unfairly. Clearly, if the outcomes instead are taken to be on a par, whatever we decide, this fairness argument in favor of randomization is strengthened even further.

[40] He presented it earlier that year, in the doctoral dissertation he defended at St. Andrews. I had the pleasure of being his external examiner.

outcome: an outcome is worse if it involves an unfair treatment. And the more unfairness it involves, the more individuals are unfairly treated, the worse it is. Lawlor, on the other hand, separates the value of an outcome from the assessment of the action that brings it about, with fairness considerations being relevant to the latter but not to the former. So do I. Without this separation, Lübbe's persuasive critique of the choice version of the Aggregation Argument would be applicable to the outcome version as well. For Peter alone to survive is particularly unfair to Mary, given that she could have been saved along with him. Thus, with unfairness counted in, this outcome would be worse (and not merely on a par) than if David alone were to survive. While the conclusion of the Aggregation Argument would still be plausible, the argument itself would be beyond repair.

To sum up: Taurek believed that in rescue cases it is justified to make a decision by a coin toss or some other random process. He was wrong. In small-scale rescue cases, it is better if more people survive rather than fewer. And this has implications for what we ought to do. But Taurek's recommendation seems right in some large-scale rescue cases – cases in which there are many people in each group, all with a claim to fair treatment. If these disjoint groups, only one of which can be saved, don't differ much in size, then denying all the members of the slightly smaller group a chance of survival is not compensated by the relatively small increase in the number of survivors, as compared with the expected number of survivors if the decision is made by a random process. In such large-scale cases, the Aggregation Argument doesn't go through: the outcomes will all be on a par, whatever we decide to do and thus what we ought to do is to toss a coin: considerations of fairness tip the scale in favor of randomization.

Acknowledgements

This paper was long in the making. Its first draft was prepared in 2012, for a colloquium on aggregationism at GAP 8 in Konstanz. The colloquium was organized by Kirsten Meyer and Thomas Schmidt, and the other two speakers were Iwao Hirose and Weyma Lübbe. Then it took ten years before I came round to write the final version. In the meantime, I presented the draft at seminars in Lund, Oxford, Glasgow, York, and Canberra. I am indebted to the participants in these events. Especially, I want to thank my colleagues in Lund, Toni, Björn, Dan, and David Alm, and philosophers in York: Christian Piller, Mozaffar Qizilbash, Thomas Baldwin and Johan Gustafsson. I also wish to thank Iwao Hirose, Weyma Lübbe, Christian Piller, and Thomas Schmidt, who have all sent me challenging and helpful comments.

References

Barrett, Hugh (2022), *Normative Judgement, Rationality, and Reflective Agency*, doctoral thesis (submitted 2021, passed 2022), Canberra: Australian National University, School of Philosophy.

Brentano, Franz (1969 [1889]), *The Origin of Our Knowledge of Right and Wrong*, translated by Roderick Chisholm, London: Routledge & Kegan Paul.

Broome, John 1998), "Kamm on Fairness", *Philosophy and Phenomenological Research* 58: 955-961.

Chang, Ruth (2002a), "The Possibility of Parity", *Ethics* 112: 659-688.

Chang, Ruth (2002b), *Making Comparisons Count*, New York: Routledge.

Chang, Ruth (2005), "Parity, Interval Value, and Choice", *Ethics* 115: 331-350.

Chang, Ruth (2016), "Parity: The Intuitive Case", *Ratio* 29:395-411.

Crisp, Roger (2008), "Goodness and Reasons: Accentuating the Negative", Mind 117: 257-265.

Dancy, Jonathan (2000), "Should we pass the Buck+", in Anthony O'Hear (ed.), *Philosophy, the Good, the True and The Beautiful*, Cambridge: Cambridge University Press.

Danielsson, Sven, and Olson, Jonas (2007), Brentano and the Buck-Passers, *Mind* 116: 511-522.

Darwall, Stephen (3002), *Welfare and Rational Care*, Princeton: Princeton University Press.

Gibbard, Allan (1990), *Wise Choices, Apt Feelings. A Theory of Normative Judgment*, Harvard: Harvard University Press.

Gustafsson, Johan (2017), "Review of Moral Aggregation by Iwao Hirose", *Mind* 126: 964-967.

Hájek, Alan, and Rabinowicz, Wlodek (2022), "Degrees of Commensurability and the Repugnant Conclusion", *Noûs* 56: 897-919.

Handfield, Toby, and Rabinowicz, Wlodek (2018), "Incommensurability and Vagueness in Spectrum Arguments: Options for Saving Transitivity of Betterness", *Philosophical Studies* 175: 2373-2387.

Hare, Caspar (2010), "Take the Sugar", *Analysis* 70: 237-247

Herlitz, Anders (2020), "Spectrum Arguments, Parity and Persistency", *Theoria* 86: 463-481.

Hirose, Iwao (2001), "Saving the Greater Number without Combining Claims", A*nalysis* 61: 341-342.

Hirose, Iwao (2004), "Aggregation and Numbers", *Utilitas* 16: 62-79.

Hirose, Iwao (2015), *Moral Aggregation*, Oxford and New York: Oxford University Press.

Howard, Christopher (2019), "The fundamentality of fit", in R. Shafer-Landau (ed.), *Oxford studies in metaethics*, vol. 14, ch. 10, Oxford: Oxford University Press: 216-236.

Kearns, S., & Star, D. (2008). Reasons: Explanations or evidence

Kamm, Frances M. (1993), *Morality, Mortality*, Volume 1: *Death and Whom to Save From It*, New York: Oxford University Press.

Kamm, Frances M. (2005), "Aggregation and Two Moral Methods", *Utilitas*17: 1-23.

Kamm, Frances M. (2007), *Intricate Ethics: Rights, Responsibilities, and Permissible Harm*, New York: Oxford University Press.

Lawlor, Rob (2006, "Taurek, Numbers, and probabilities", *Ethical Theory and Moral Practice* 9: 149-166.

Lübbe, Weyma (2008), "Taurek's No Worse Claim", *Philosophy and Public Affairs* 36: 69-85.

Lübbe, Weyma (2015), *Nonaggregationismus: Grundlagen der Allokationsethik*, Münster: Mentis.

McHugh, Conor, and Way, Jonathan (2016), "Fittingness First", *Ethics* 126: 575-606.

McHugh, Conor, and Way, Jonathan (2022), *Getting Things Right: Fittingness, Reasons, and Value*, Oxford: Oxford University Press.

Nagel, Thomas (1970), *The Possibility of Altruism*, Princeton, N.J: Oxford University Press.

Nagel, Thomas (1986), *The View from Nowhere*, New York: Oxford University Press.

Nebel, Jacob M. (2020), "A Fixed-population Problem for the Person-affecting Restriction", *Philosophical Studies* 177:2779-2787.

Olson, Jonas (2009), "Fitting Attitude Analyses of value and the Partiality Challenge", *Ethical Theory and the Moral Practice* 12: 365-78.

Parfit, Derek (1978), "Innumerate Ethics", *Philosophy and Public Affairs* 7: 285-301.

Parfit, Derek (2001), "Rationality and Reasons," in *Exploring Practical Philosophy: From Action to Values*, ed. By D. Egonsson, B.Petersson, J. Josefsson, and T. Rønnow-Rasmussen, Aldershot: Ashgate: 17–41.

Parfit, Derek (2011), *On What Matters*, vol. 1, Oxford: Oxford University Press.

Piller, Christian (2014), "What Is Goodness Good For", *Oxford Studies in Normative Ethics*, vol. 4, ed. By Mark Timmons, Oxford: Oxford University Press: 179-209.

Qizilbash, Mozaffar (2005), "The Mere Addition Paradox, Parity and Critical-level utilitarianism", *Social Choice and Welfare* 24: 413-441.

Qizilbash, Mozaffar (2007), "The Mere Addition Paradox, Parity and Vagueness", *Philosophy and Phenomenological Research* 75: 129-151.

Rabinowicz, Wlodek (2008), "Value Relations", *Theoria* 108: 18-49.

Rabinowicz, Wlodek (2012), "Value Relations Revisited", *Economics and Philosophy* 28: 133-164.

Rabinowicz, Wlodek (2021), "Incommensurability Meets Risk", in *Incommensurability: Vagueness, Parity and Other Non-conventional Relations*, ed. by H. Anderson and A. Herlitz, New York and London: Routledge: 201-230.

Rabinowicz, Wlodek (2022), "Can Parfit′s Appeal to Incommensurabilities Block the Continuum Argument for the Repugnant Conclusion?", in *Ethics and Existence: The Legacy of Derek Parfit*, ed. by J. McMahan, T. Campbell, J. Goodrich and K. Ramakrishnan, Oxford: Oxford University Press: 430-460.

Rabinowicz, Wlodek, and Rønnow-Rasmussen, Toni (2004), "The Strike of the Demon: On Fitting Pro-attitudes and Value", *Ethics* 114: 391-423.

Rabinowicz, Wlodek, and Rønnow-Rasmussen, Toni (2006), "Buck-Passing and The Right Kind of Reasons", *Philosophical Quarterly* 56: 114-120.

Rawls, John (1999 [1971]), *A Theory of Justice*, revised edition, Cambridge, Mass.: Harvard University Press.

Rønnow-Rasmussen, Toni (2004), "Buck-passing Personal Values", in W. Rabinowicz and T. Rønnow-Rasmussen (eds.), *Patterns of Value – Essays on Formal Axiology and Value Analysis,* vol. 2, Lund: Lund Philosophy Reports 2004:1: PAGES?

Rønnow-Rasmussen, Toni (2009), "Normative Reasons and the Agent-neutral/Relative Dichotomy", *Philosophia* 37: 227-243.

Rønnow-Rasmussen, Toni (2007), "Analysing Personal Value", *The Journal of Ethics* 11: 405-435.

Rønnow-Rasmussen, Toni (2011), *Personal Value*, Oxford: Oxford University Press.

Skorupski, John (2007), "Buck-Passing about Goodness", in *Hommage à Wlodek: Philosophical Essays Dedicated to Wlodek Rabinowicz*, ed. D. Egonsson, J. Josefsson, B. Petersson and T. Rønnow-Rasmussen, www.fil.lu.se/hommageawlodek.

Skorupski, John (2010), *The Domain of Reasons*, Oxford: Oxford University Press.

Taurek, John (1977), "Should the Numbers Count?", *Philosophy and Public Affairs* 6: 293-316.

Taurek, John (2021), "Reply to Parfit's 'Innumerate Ethics'", in *Principles and Persons: The Legacy of Derek Parfit*, ed. By J. McMahan, T. Campbell, J. Goodrich, and K. Ramakrishnan, Oxford: Oxford University Press, ch. 14: 311-322.

Zimmerman, Michael J. (2011), "Partiality and Intrinsic Value." *Mind* 120: 447-483.

Against the 'First' Views

Why None of Reasons, Fittingness, or Values are First

Andrew Reisner

0. Introduction and Overview

Toni Rønnow-Rasmussen has been one of philosophy's most important contributors to our understanding of the nature of value, not least of all with respect to questions of whether different types of value are reducible to each other and whether value in general is basic. In this paper I make my own modest effort to follow in Toni's long shadow. The aim of this paper is to argue that there are at least two categories of normative or 'non-descriptive'[1] properties (in the terminology used in this paper) that cannot be reduced to other more basic non-descriptive properties and that one of those categories is that of value properties. Although the emphasis in the paper is on reductionist views, most of the arguments work equally well against a weaker category of view about the relations amongst non-descriptive properties, namely those that require different categories of non-descriptive properties to be linked by a necessary bi-conditional.

The arguments in this paper are incomplete in at least one rather obvious way. There are a number of candidates for basic non-descriptive properties,[2] but here I focus only on the three non-descriptive properties that receive the most attention in

[1] I recognise the problem with this terminology, insofar as 'non-descriptive' suggests that I am taking a linguistic or metaphysical stance on the nature of normative properties broadly construed, whereas I do not mean to do so. There are problems with other alternatives. Using 'normative' in this context makes it difficult to distinguish between the kinds of properties which are normative in a stricter sense, like oughts and reasons, and those which are not, such as evaluative properties. That same problem arises for using 'evaluative' to describe the category. I apologise to the reader for not finding a better term to use.

[2] I refer the reader to chapter 1 of Nils Sylvan's (2021) recent doctoral thesis for an excellent catalogue of candidate properties.

the literature: reasons, fittingness, and value. It is my conjecture, one for which I have no general argument at present, that with some work the arguments concerning the relations amongst those three properties can be adapted for use against reductionist programmes employing other combinations of non-descriptive properties. With that limitation in mind, I shall argue that no reductions are possible amongst these properties and thus that views that fall under the heading of 'reasons-first', 'fittingness-first', and 'value-first' – views that hold that there is a single most basic non-descriptive property – are false.

The arguments in this paper are directed in the main at metaphysical reductions, where one property is reduced to one or several more basic properties. There are at least two ways to argue against putative reductions of this kind. One is to show that the *analysandum* and the *analysans* have different necessary extensions. In work on value, this strategy is perhaps most familiar in the form of the *wrong kind of reasons problem* (WKR) for the fitting-attitude analysis of value.[3] WKR arguments are intended to show that there are instances in which there is a reason to favour *x* when *x* is not valuable. Likewise, there is the less commonly discussed *wrong kind of value problem* (WKV),[4] which aims to show the reverse, namely that there are instances where *x* is valuable, but where there is no reason to favour *x* or it is not fitting to favour *x*.

A second way to argue against attempted reductions is to show that despite the necessary extensional adequacy of the proposed analysis, the *analysans* lacks essential characteristics possessed by the *analysandum*, or alternatively adds features that in the relevant sense cannot be part of the *analysandum*. Needless to say, these two strategies do not exhaust the possibilities for arguing that an attempted reduction fails, but they are the two approaches that will be used in this text. These approaches differ in force in one important respect. Necessary extensional inadequacy is not only sufficient for showing that a reduction fails, but it is also sufficient to defeat a weaker claim, namely that there is a necessary bi-conditional equivalence between two or more categories of properties. The first strategy may thus be used to show that the *correctness conditions*[5] for one class of non-descriptive property cannot be given in terms of another, insofar as they are not necessarily extensionally equivalent. The second strategy does not show this directly. As at least some *-first* authors take *-first* claims to be about correctness conditions, they are only committed to necessary bi-conditional equivalence.

[3] It is undoubtedly fitting in this context to note that this problem was given life by the two classic Rabinowicz and Rønnow-Rasmussen (2004 & 2006) papers. The literature on this topic is now extensive. For some important examples, see Danielsson and Olson (2007), Lang (2008), and Olson (2009).

[4] See Bykvist (2009 & 2015), Dancy (2000), Heathwood (2008), Hurtig (2019), and Reisner (2015).

[5] Correctness conditions in this sense give the criteria for the conditions under which an object has a particular property. Understood this way, the fitting attitude analysis of good would say that an object is good only and always under the condition that it is fitting to favour that object without positing that what it is for the object to be good is for it to be fitting to favour.

Because of this, the extensional arguments tell against a wider understanding of what *-first* views amount to being.

In this text, I shall offer what I take to be decisive examples showing that neither fittingness nor reasons is necessarily extensionally equivalent to value, which suffices to show that an overarching *-first* project that aims to reduce two of reasons, fittingness, and value to the third property must fail. However, I shall also argue, using the second strategy, that reasons cannot be reduced to fittingness or to value, which tells us that at minimum reasons[6] and value are not less fundamental than fittingness.

1. Unalike Variance Conditions for Reasons and Value

Reasons and value have unalike variance conditions, or so I shall argue. And if they have unalike variance conditions, then that is enough to show, assuming that reasons and value are two of the three candidate non-descriptive property categories, that *-first* theories are false.

Any *-first* view with the ambition of being an analysis or a reduction must be built on a core bi-conditional that contains one of the non-descriptive properties on the lefthand side and another non-descriptive property of a different kind on the righthand side. These bi-conditionals are in general stronger than simple bi-conditionals, for example they may include determination and must in any case be necessary to play a role in an analysis. But since the present concern is with extensional inadequacy (from under-generation), it will suffice to work with simple bi-conditionals; if the relevant simple bi-conditional is false, then *a fortiori* so is a strengthened bi-conditional. Let us begin by focusing on the reasons version of the fitting-attitude analysis:

> 2. *The reasons version of the fitting-attitude analysis of value (RFAV):*
> *x* is valuable if and only if there is a reason to favour *x*.
>
> F2. RFAV: *x* is valuable ⟷ there is a reason to favour *x*.

The target is to develop a schema for creating examples in which *x* is valuable, but there is no reason to favour *x*. One may start by considering the structure of reason relations:

> 3. *The simple reason relation*: Fact *f* is a reason for agent *A* to *ψ* to degree *d*[7]

[6] I suspect that in the final reckoning, one may need to treat oughts as irreducible to reasons. See Gjelsvik (2020) for a defense of the view that reasons and oughts cannot be reduced to each other.

[7] Some contemporary writers omit the final place in this relation. John Skorupski (2002, 2010) was careful to avoid this mistake in his pioneering work on the metaphysics of reasons.

In the simple reason relation, '[f]act *f*' should be interpreted liberally so as to include conjunctions of facts or sets of facts.[8] The schematic variable ψ simply stands for anything for which there can be a reason (i.e. an action, belief, emotion, pro-attitude, etc). Crucially, reasons are indexed to agents.

1.1 The Under-generation Argument for Reasons and Value

With the essentials of the reason relation and RFAV having been set out, it is now possible to develop a schema for creating cases in which the lefthand side of the bi-conditional is true but the righthand side is false, thus showing that an analysis of value in terms of reasons to favour under-generates.

The simplest structure for such examples relies on descriptive, or if one prefers, non-normative *entanglement*.[9] One needs to generate examples in which favouring *x* makes *x* not be valuable.[10] I shall focus for now on *good* as a paradigm type of value. Here is a generic counterexample:

> 4. *The generic counterexample*: *x* is valuable at t_1 if and only if nobody ever has, does, or will favour *x*.

It is not difficult to fill out the details of this schema by making an appeal to sufficiently knowledgeable and powerful agents. Imagine that the demiurge has created a powerful entity whose nature is such that she relieves pain and suffering around the world anytime she waves her left arm, so long as nobody ever has, does, or will favour her waving her left arm. Her nature is also such that if anyone ever has, does, or will favour her waving her left arm, the effect of her doing so will instead be that she causes pain and suffering around the world. One may treat the effect of her waving her arm in both circumstances as necessary[11] due to her nature.[12]

---

[8] In Skorupski's (2002) explication of the reason relation, *f* stands for a set of facts.

[9] See Reisner (2015) and Risberg (2018) for detailed discussions of entanglement. The 'descriptive' qualifier is important; as Haim Gaifman argued as far back as the 1983, normative entanglement is highly problematic. I take this observation from Wlodek Rabinowicz's opposition at Olle Risberg's disputation.

[10] Strict covariance is also sufficient.

[11] An anonymous reader in another context pointed out to me that if one accepts S5, then this example is impossible, unless the entity in question exists. Given that philosophers often rely on possibly (but not in fact) necessary examples, one will have to take one's chosen solution to understanding this and other examples of this kind. Nothing about the example itself hinges on accepting S5. Only accepting K is required. I thank Jonathan Shaheen for a valuable discussion about this worry.

[12] In the past (2009 and 2015) I have treated FA as concerning final value. I assumed, too, that the final value of an action was the value of its consequences. I shall dispense with that assumption here for reasons that will soon be apparent.

An example of this form entangles favouring *x* (descriptive) with *x*'s value (non-descriptive), or lack thereof. One can construct other such examples, of course, based on the same schema. Implicit in using an example of this form is the assumption that there is no reason to favour *x* if *x* will be bad, should one favour it. This underlying assumption seems highly plausible to me on its face.[13] Favouring *x* effaces the reasons for favouring *x* and thus defeats even the weakest guidingness constraints on reasons.[14]

Now we are in a position to see why value and reasons may have unalike variance conditions in the arm-waving example. So long as nobody ever favours the powerful entity's waving her left arm, it is good (valuable) that she waves her left arm. If somebody ever favours her waving her left arm, then it is bad (has disvalue) that she waves her left arm. Thus, whether or not somebody favours her waving her left arm changes the value valance of her waving her left arm. If we accept the argument about self-effacing reasons not being reasons at all, then there is never a reason to favour her waving her left arm. While the value valence of her waving her arm changes depending on whether or not anyone favours it, the valence of the reason to favour (i.e. a reason not to favour) never changes. And thus we have under-generation.

More needs to be said about this example, as I have as yet not specified what sort of value is at stake.[15] I shall consider three possibilities: intrinsic final value, extrinsic final value, and instrumental value.[16] It is at best unclear whether the entity's waving her left arm has intrinsic final value. The act itself, at least under that description, appears to be neutral. Perhaps the case could be reconfigured such that it has intrinsic final value, but I am unsure, so I shall assume for the moment that it does not. A second possibility is that the case has extrinsic final value. This seems more plausible to me. One might hold the view, for example, that the final value of an action is a function of the amount and distribution of wellbeing of its consequences.[17]

With respect to this case and others structured like it, whether something is extrinsically finally valuable will depend first on whether there is in fact such a thing as extrinsic final value and then on how one divides up the value bearers and background conditions. So perhaps the arm-waving example concerns extrinsic final value. It should be much less controversial to say that the arm-waving example

[13] I argue for this claim in §2.1.

[14] See Risberg (2020) and Rosenqvist (2020) for further discussion on guidingness. As Bruno Guindon pointed out to me, guidingness constraints are often understood in some sort of deliberative internalist terms, i.e. that one can do what there is a reason to do by including the reason in one's deliberation. The guidingness constraints that are relevant here are extremely weak and fully consistent with rejecting all forms of deliberative constraints.

[15] The importance of clarifying what sort of value applies in this example was pointed out to me by Antti Kauppinen, who also provided advice I have followed here in structuring the discussion.

[16] I have left out a discussion of possible differences between *value for someone* and *value simpliciter*. For a discussion of the latter in the context of FA, see Rønnow-Rasmussen (2007, 2011 & 2021).

[17] This is perhaps John Broome's (2004) view in *Weighing Lives*.

is a case of instrumental value. The arm-waving case thus creates clear difficulties for a reason-to-favour analysis of instrumental value. It may create difficulties for an analysis of final value that includes extrinsic final value, and it does not yet pose a straightforward difficulty for analysing intrinsic final value.

A second example is required to create clear difficulties for an analysis of intrinsic final value.[18] Let us suppose, as many philosophers have, that it is intrinsically finally valuable to love another person unconditionally.

This example also involves a demiurge who decides this time that if anyone ever favours a particular instance of Xenophon's unconditionally loving any particular person, he will never unconditionally love that person. The demiurge's decision has the peculiar effect that it is impossible to favour a particular (actual) instance of Xenophon's unconditionally loving another person, because the existence of the pair {Xenophon loves *x* unconditionally at t_1, anybody ever favours that Xenophon loves *x* unconditionally at t_1} is impossible. No instance of Xenophon's loving another person can be favoured while there is a reason to favour it, because if it is favoured, there will be no such instance. Put another way, the demiurge's condition makes favouring particular (actual) instances of Xenophon's unconditionally loving another person metaphysically impossible.

One may find parallel cases when it comes to reasons for action and value. Suppose that one offers the following bi-conditional claim about beauty:

> 5. *The beauty bi-conditional*: *x* is beautiful if and only if there is a reason to have an aesthetic experience of *x*.

We should understand 'have an aesthetic experience of *x*' as encompassing actions such as viewing paintings, listening attentively to symphonies, watching films, etc. Now consider a delicate sandstone rock formation whose unique beauty can only be experienced from the changing perspectives given by climbing its face. Regrettably the rock is delicate enough that even the lightest touch of its surface destroys those natural features that make it beautiful, rendering its beauty impossible for anyone to experience.[19]

In this case, presumably the features that make the rock formation beautiful do so whether or not they can be experienced.[20] Thus so long as one does not climb the formation, it remains beautiful. But if one is climbing or has climbed the formation,

---

[18] This example was proposed to me by Jaakko Kuorikoski. I am grateful for his suggestion.

[19] Randall Harp expressed to me the worry that there are no beautiful objects that could only be experienced in this way, as perhaps an object that is beautiful, but that cannot be experienced, is not in fact beautiful. I do not share this intuition, but I have no argument against it that does not rely on one's already sharing my intuition that there are such objects. Bruno Guindon expressed concern that the example itself suggests the implausibility of the beauty bi-conditional.

[20] Objectivism of this sort about beauty is controversial. Nonetheless, I follow Elisabeth Schellekens (2006) in accepting an adequate degree of objectivity for the purposes of this example.

then the formation is not beautiful, due to the destructive effects of climbing it. There is no reason for one to experience the formation, because doing so effaces the physical features of the formation that provide reasons to experience it; one has no (aesthetic) reason to climb the formation once one is climbing it. Here again, we see that there is no reason for one to climb the formation, irrespective of whether one climbs it or not, but the formation is beautiful if one does not climb it and is not beautiful if one does.[21]

It bears noting at this point that although RFAV is formulated as a simple bi-conditional, the counterexamples would also hold for a counterfactual version of the principle. In all relevantly similar worlds, the same entanglements would exist.

1.2 Objections to the Counterexample Schema

It is of course fair to ask whether the assumption that self-effacing (putative) reasons to favour are not actual reasons to favour is correct. I believe it is, but I would like to look at two possible objections against the force of cases built on the entanglement schema.

The first objection posits that there is a reason for someone in another possible world to favour the entity's waving her left arm, since that person would sit outside the actual world's past, present, and future. I find this proposal very odd, but a parallel proposal has been suggested to me with respect to fittingness. There are a number of technical issues that arise with respect to this proposal, many of which I have discussed in depth in an earlier paper.[22] However, I am now convinced that there is a (somewhat) more straightforward way to reply to this objection than my previous attempt, at least with respect to reasons.

Note that this objection is describing a possible reason to favour the entity's waving her left arm, not an actual (in the modal sense) reason to favour it. This would mean that RFAV would have to be modified:

> 2a. *Possible reasons fitting-attitude analysis of value* (PRFAV): *x* is good in the actual world if and only if there is a possible reason to favour *x*'s occurrence in the actual world.[23]

Although the arm-waving case is stated in general terms, it has specific implications. If it is generally good for the entity to wave her left arm, so long as it is never favoured, then each specific existentially quantifiable occurrence of her waving her

[21] Simon-Pierre Chevarie-Cossette suggested another example of a beautiful painting that blinds anybody who looks at it before they can see it.

[22] Reisner (2015).

[23] I have not noticed any commitments specifically to this view in writing. Despite that, it has often been suggested to me in correspondence and conversation as a way to solve the sorts of difficulties raised by WKV.

left arm (when nobody favours her doing so generally) is also good. A successful analysis of *good*, or of any sort of value, and the bi-conditional on which it is built, will entail that each specific instance of the entity's waving her left arm is good under the condition that (eternally) nobody favours it.

PRFAV implies that there is someone in another possible world who has a reason to favour one or more specific occurrences in the actual world in which the entity waves her left arm. This is because reasons are indexed to individuals, or sets of individuals. A reason needs to be a reason for at least one particular individual to be a reason at all. It is doubtful that individuals in other possible worlds can favour an entity in the actual (from our perspective) world's doing anything at all, because favouring that occurrence would require having that occurrence in mind. And it is itself doubtful that we can have singular thoughts about individuals or specific events in other possible worlds,[24] which is what would be required to get a particular individual (situated in a particular world) in mind. If nobody *can* have the reason, then nobody *does* have the reason; therefore, it is not the case that there is a reason for *x* to favour that such-and-such occurs in another possible world.

But suppose that it is possible to have singular thoughts about individuals or events in other possible worlds. In that case, PRFAV itself seems like a bad principle, in part because it would over-generate in a peculiar way.

Suppose that a powerful being will improve life in another possible world (which is not the actual world) each time someone in the actual world[25] performs a cruel act that causes only pain. Someone in that other world has a reason to favour the performance of those cruel acts in the actual world, namely that they reduce suffering in her world. According to PRFAV, the fact that she has a reason to favour their occurrence in the actual (from our perspective) world also makes them good in the actual world, when it instead is right to say that they are bad in the actual world, although their occurrence in the actual (from our perspective) world is good in her world.

Of course talk about what merely possible rather than actual individuals have reason to favour in the actual (from our perspective) world is strange in numerous ways, not least of all because it is difficult to understand the idea at all without accepting modal realism. Otherwise, it is not clear that there are in an interesting sense individuals in other possible worlds.[26] The very claim that *x* is good if a merely possible person favours it sounds false. Strangeness aside, PRFAV is extensionally inadequate, which is enough to reject it without complaining about the metaphysics.

---

[24] *Ibid.* and see Soames (2002).

[25] The *actual* operator indexes to this world, whereas 'another possible world' should be taken to indicate the use of a different indexical operator *W*, which functions like the *actual* operator but localises to the world in which it is being used. I discuss how this operator works in Reisner (2015).

[26] On a modal anti-realist view, one might wish to treat worlds as logically consistent complete state descriptions. Such descriptions would include descriptions or representations of individuals, but not actual individuals.

The second objection concerns the 'eternity' condition in the counterexample to RFAV, namely that it is implausible to say that the entity in the example's actions could be affected by what occurs in future, perhaps because of an assumption that the future is open and thus non-determinate. I do not have very much to say about this objection, because it clearly hinges on the difficult question of whether the future is determinate, or perhaps knowable. I suspect that if the future is non-determinate or non-knowable, complications will arise, too, for versions of RFAV that rely on the possibility or existence of reasons in future to favour the entity's waving her arm. I shall simply concede for the time being this remains an unaddressed potential objection.

2. The Argument Extended to Fittingness

If the argument in §1 is correct, then reasons-first is ruled out, because it is extensionally inadequate on any interpretation. This still leaves the possibility that a fittingness-first view is correct. In this section, I argue that fittingness-first is false, most importantly because the fitting-attitude analysis of value is extensionally inadequate, under-generating in some circumstances and perhaps over-generating in others.

However, I shall begin by looking at another potential problem, one astutely identified by Christopher Howard.[27] The problem is that fittingness on traditional views seems to under-generate with respect to reasons, at least if one accepts that there are state-given reasons for propositional attitudes. Howard's account is cleverly constructed so as to avoid cases in which fittingness under-generates with respect to reasons.

My presentation of Howard's view is not entirely faithful to the original, but the changes affect small details that are distracting to include in this context and not the central extensional adequacy concerns.[28] His account is built on two main claims:

> 6. *Value as fittingness (VAF)*: *x* is non-instrumentally good if and only if it is fitting to favour *x*.

And

> 7. *Reasons as fittingness (RAF)*: There is a reason to favour *x* if and only if: 1) it is fitting to favour *x*, or 2) it is fitting to favour that one favour *x*.

VAF is just FA. RAF, read with the first disjunct alone, says that there is a reason to favour *x* if and only if it is fitting to favour *x*. That would appear to rule out state-

[27] Howard (2019).

[28] I thank Christopher Howard for checking to make sure I have not misrepresented his view in a way that does violence to it.

given reasons. For example, it would be ruled out that one ought to desire to listen to Vogon poetry[29] to avoid being thrown off of a Vogon ship, although Vogon poetry itself lacks desirable qualities.[30] Intuitively, it is good to desire to listen to Vogon poetry, because it is good to avoid being cast out into the vacuum of space. According to VAF, it would follow that it is fitting to desire that one desires to listen to Vogon poetry. Howard stipulates that when a second-order desire is fitting, then there is a reason to have the first-order desire. This resolves the under-generation problem for state-given reasons.

However, notice that Howard's view still entails that *x* is good only if one has a reason to favour *x*. That is because the righthand side of VAF and the first disjunct on the righthand side of RAF specify the same condition, namely that it is fitting to favour *x*. Thus when it is fitting to favour *x*, *x* is good and there is a reason to favour *x*.

Yet this is problematic in light of the arguments in §1. They show that reasons under-generate with respect to value, i.e. that there are some cases in which *x* is good, but there is no reason to favour *x*. That conclusion is inconsistent with Howard's view:

1)	*x* is good iff it is fitting to favour *x*	(Ass. VAF)
2)	If it is fitting to favour *x*, then there is a reason to favour *x*	(Ass. sufficient cond. in RAF)
3)	If *x* is good, then there is a reason to favour *x*	(from 1, 2)
4)	Not: If *x* is good, then there is a reason to favour *x*	(Ass. from §1)
5)	Conclusion: Either 1, 2, or 4 is false	(from 1-4)[31]

Assuming we do not reject premise 4, then this raises a problem for Howard's view: namely that either VAF is false or that RAF is false and consequently that all-in his view is false. If nothing else, this points to the difficulty of constructing a fittingness-first account that implies that there are state-given reasons for propositional attitudes.

Nonetheless, for now I want to focus on FA/VAF and show that it is false. To do so, I shall introduce a new version of WKV for fittingness. I shall take up the question of whether one of either reasons or fittingness might be first relative to the other in §3.

---

[29] For more on Vogon poetry, including some examples, see Adams (1981).

[30] For purposes of the example, I assume that Vogons can tell whether one has a desire to listen to their poetry or whether one is merely acting as though one does.

[31] I thank Jens Johansson for pointing out a problem, now remedied, with an earlier version of this argument.

2.1 Some New Arguments against the Fitting-attitude Analysis of Value

There are, as far as I can see, two strategies for showing that fittingness and value have unalike variance conditions. One strategy is the strict argumentative analogue of the arm-raising or unconditional love argument presented in §1 against RFAV. One need only swap in 'fittingness' for reasons and fix the grammar accordingly to see how such an argument would look.

However, there is a complication. The argument in §1 relied on adopting what I shall call the 'realisability condition for reasons' (RCR):

> 8. *Realisability condition for reasons (RCR)*: Fact *f* is a reason for agent *A* to ψ to degree *d* only if *A* can (metaphysically) ψ whilst there is (still) a reason for *A* to ψ to degree *d.*

As I noted when the idea was presented informally in §1, it is difficult to doubt this condition, which may be understood as an extremely weak guidingness constraint.[32] A parallel condition would be required to transfer the same argumentative structure to fittingness. That would give us a *realisability condition for fittingness (RCF)*:

> 9. *Realisability condition for fittingness (RCF)*: It is fitting for *S* to favour *A*'s ψ-ing only if *S* can favour *A*'s ψ-ing whilst it is (still) fitting for *S* to favour that *A* ψs.

Intuitions about this principle may be less clear than they are for RCR. However, I suspect that most people will find RCF difficult to doubt on reflection.

It may help to begin by thinking about fittingness outside the context of FA. Consider these fittingness claims, some with synonyms for 'fitting' to avoid leaning too heavily on a single word for evidence:

> F1: It is fitting to feel gratitude towards Sophia, but not if you feel gratitude towards her.
>
> F2: It is appropriate to be angry at Harvey, but not if you are or become angry at Harvey.
>
> F3: It is correct to hold your fork in your left hand, but not if you hold your fork in your left hand.
>
> F4: It is meet to honour Achilles, but not if you honour Achilles.[33]

Each of F1-F4 would be a pretty odd thing to say. Presumably, they are odd to say, because they each imply a conditional claim of the form: If you will feel/do *x*

[32] I thank Bruno Guindon for pointing out to me that I ought to say this explicitly.

[33] I thank Jimmy Goodrich for suggesting a valuable revision to these examples.

towards *A*, it will not be fitting/appropriate/correct/meet to feel/act that way. Or perhaps it implies a counterfactual version of the same claim. It would be bemusing, if not vexing, to be told that it is appropriate to hold one's fork in one's left hand, only then to be told that holding one's fork in one's left hand is inappropriate on account of the fact that one is holding one's fork in one's left hand. One might be forgiven for worrying that one has fallen through the looking glass. These examples are, of course, not dispositive. Perhaps the relevant intuitions rest on social factors that are not indicative of the nature of fittingness itself. However, they are at least suggestive.

Let me offer what may be a stronger consideration in favour of RCF. The entanglement cases I have been discussing are instances of the following general schema:

> C1F: It is fitting that *S* favour *A*'s *ψ*-ing only if *S* does not favour *A*'s *ψ*-ing.[34]

Particular events can be fitting to favour, too:

> C1Fp: It is fitting that *S* favour that instance of *A*'s *ψ*-ing only if *S* does not favour that instance of *A*'s *ψ*-ing.

If RCF is correct, then no cases for which either C1F or C1Fp are true. Conversely, if there are cases for which C1F or C1Fp are true, we must reject RCF. If one could find a reading of C1F or C1Fp where there were cases that seemed intuitively correct, then assuming that other reasonable conditions are met, we could reject RCF. As I shall argue briefly here, it is difficult to see what kind of reading would do the trick.

One way to try to find acceptable cases of C1F and C1Fp is to see if we can find a helpful interpretation of 'favouring *A*'s *ψ*-ing'. Both the most natural reading and what strikes me as one promising-seeming alternative interpretation are problematic. One way to read the phrase is with a universal quantifier: all favourings of *A*'s *ψ*-ing are fitting for *S*. But C1F and C1Fp entail that no favourings of *A*'s *ψ*-ing are fitting. Consequently, this reading is simply false if there are any favourings of *A*'s *ψ*-ing.

Another possible reading of C1F and C1Fp is that 'favouring *A*'s *ψ*-ing' should be understood as expressing an event (or mental state) type. Since the existence of a type does not entail the existence of tokens of that type, it seems open in principle that it could be fitting for *S* to favour *A*'s *ψ*-ing, *qua* type, without *S*'s ever favouring a token instance of *A*'s *ψ*-ing. This reading is better, but still problematic, because the type features in a relation in which none of its tokens can feature. Of course, there are some relations in which types can feature in which their tokens cannot due to category problems, e.g. those relations in which the relevant *relatum* must be an abstract object and the type's tokens are concrete objects.

[34] *S* and *A* need not be different individuals, but they of course may be.

In this case, however, it is difficult to see why the fittingness relation could not take an individual instance of favouring as a *relatum*. Thus the situation remains odd. Consider a parallel case. The type, Charles Maturin's *Melmoth the wanderer*, contains a greater number of nested narratives than either the type or a complete token of Edgar Allan Poe's 'The cask of amontillado'. It is impossible that a complete token of *Melmoth the wanderer* contains fewer nested narratives than either the type or a complete token of 'The cask of amontillado'.

It is generally, but not universally the case that tokens share the relevant properties of their types. Given that there is no difficulty with the existence of complete tokens of favouring event/state types, it seems to me that interpreting 'favouring *A*'s ψ-ing' as being about an event or mental state type does not render C1F or C1Fp true, at least not without further argument. In order for the use of types to work, one would have to be happy with the existence of types with complete tokens that do not share in principle shareable properties and relations with the type itself, where the failure to share in those properties is not due to category problems.[35] To the best of my knowledge, there has been very little work done on spelling out the conditions under which tokens inherit properties or roles in relations from their types, and thus I make the foregoing comments with all due caution.

A final interpretation of C1F and C1Fp is that 'favouring *A*'s ψ-ing' expresses an existentially quantified claim about actual or possible favourings. C1F and C1Fp remain false on this interpretation, as no actual or possible instances of *S*'s favouring that *A* ψs make them come out as true. One can make the modal point explicit:

> C1F*: It is, or would be, fitting that *S* favour *A*'s ψ-ing only if *S* does not, or would not, favour *A*'s ψ-ing.

Someone who wishes to deny RCF must offer another interpretation of 'fitting to favour' that is consistent with the fact that there are no possible instances of favouring, actually or counterfactually, that have the property of being fitting.

Thus far I have been discussing these cases with the assumption that *S* and *A* are in the same world. As far as I can see, the remaining option is to allow that *S* and *A* exist in different worlds. I have already mentioned some difficulties with doing this,[36] but I shall set those aside. The arm-raising example poses no problem for FA, if we allow trans-world fittingness – its being fitting for an individual in one world to favour events or states-of-affairs in another – into the analysis.

However, trans-world fittingness has its own difficulties. In particular, it over-generates for value. I can offer two kinds of example of over-generation. The first is the example of attitudes that are fitting on comparative grounds:

---

[35] I thank Louis deRosset and Matti Eklund for very helpful correspondence on the question of the inheritance of properties and relations between types and tokens.

[36] See §1 and Reisner (2015).

> 10. *Comparative admiration*: It is fitting to admire individuals, the moral character of whom is substantially higher than our own and than that of those around us.

In the actual world, this is at least a plausible fittingness principle. In a scene in the movie, *Rocky*, Rocky Balboa is watching a fight on TV at a local bar. Apollo Creed wins, but the bartender dismisses Creed as a chump. Rocky is appalled and criticises the bartender, saying that at least Creed took his best shot, remarking that the bartender has not done anything remotely so worthy with his life. Rocky is of course impressed that Creed won, but he also admires his dedication to developing his talents.[37] The admonition and the admiration would be out of place if Creed's efforts were merely typical of those made (up to that time) by Rocky himself, the bartender, and the other 'bums[38] from the neighbourhood', even if many other top boxers train equally as hard.

If we accept *comparative admiration*, or any other fittingness claim with a similar structure, we end up with the following problem. Suppose that *S* lives in a possible world occupied only by people of low moral character. *S* (somehow) comes to learn about *A*, who exists in a different possible world. Although *A* is in fact a pretty awful person by the standards of *A*'s world, he is a paragon of virtue compared to those who inhabit *S*'s world. It is fitting for *S* to favour *A*, but it is clearly not the case that *A* has the property of being admirable in *A*'s own world. That is the first example of over-generation.

Here is a second. If we accept the strange picture on which people in one world can get those in other worlds in mind, the following is a possible case. Individuals in *S*'s world take the greatest pleasure from the existence of feats of daring-do in other worlds. In her own (different possible) world, *A* sets out to climb its tallest mountain. It is fitting for *S* to favour that *A* climb the mountain, because *A*'s doing so is good in *S*'s world due to the pleasure that her doing so causes there. But let us suppose that *A*'s climbing the mountain in her own world will lead her to install the relay that will bring Skynet online. Her climbing the mountain is bad in her own world. It is fitting for *S* to favour that *A* climb the mountain, but it is not good in *A*'s world that she do so, violating the core bi-conditional of FA.[39]

Therefore, I conclude that value cannot be reduced to fittingness, and I have likewise argued that value cannot be reduced to reasons. This entails that value is

---

[37] For people concerned about *Rocky* interpretation, this point is made explicit in the temporally distant sequel, *Creed*.

[38] Henry Hill expresses a similar sentiment, although in his case about being a 'schnook', in *Goodfellas*.

[39] Peter Fritz pointed out to me the extreme bizarreness of the metaphysics required to make sense of this example, and I can only agree. However, it seems to me that someone who wished to use trans-world fittingness as a way to resolve the worries I have raised about FA would have to accept similarly bizarre metaphysics. I should certainly be content to see the entire approach of using trans-world fittingness ruled out as beyond the pale of reasonable metaphysics. I am regrettably not in a position to make that judgement or the required argument myself.

not subject to analysis or necessary bi-conditional equivalence in the manner required for fittingness-first and reasons-first theories. This rules out any understanding of -first views that entail at least as much a necessary bi-conditional equivalence.

3. Reasons and Fittingness

We are now left with a final question: is one of reasons or fittingness first relative to the other? I believe the answer to this question is 'no', but I have no conclusive argument to offer to that effect. Instead of offering a conclusive argument, I wish to turn to Danielsson & Olson's influential 2007 paper on FA.

When Danielsson & Olson set out to solve the wrong kind of reason problem, they did so by importing a non-descriptive notion, *correctness,* that appeared to be in some important way distinct from *being a reason*. Correctness is fittingness. Their strategy was initially to divide reasons into two kinds: those that arise directly from correctness ('content reasons') and those that do not ('holding reasons' that are not also content reasons). The former are suitable for FA, and the latter are not.

Importantly for the present discussion, Danielsson & Olson then pursue a reductive project in the later part of the paper, developing a Ewing-inspired account of how to reduce all holding reasons to content reasons. Because content reasons are nothing more than facts about its being correct or fitting to hold certain attitudes, Danielsson & Olson's project is in the final analysis an early version of fittingness-first.

We can see the same general idea if we look back to Howard's reasons-as-fittingness condition. He offers a way of accounting for non-correctness reasons in terms of fittingness. I have already introduced Howard's account in some detail, and it is worth considering again in this context.

According to Howard, there is a reason to have a pro-attitude with contents *c* if it is fitting to favour *c* or if it is fitting to favour favouring *c*. This second condition is perhaps necessarily co-extensional with Danielsson & Olson's holding-but-not-content reasons. Let us suppose that it is. A proposed advantage of Howard's view is that it offers conceptual gain.[40] But conceptual gain comes at the cost of theoretical unity. The relationship between fittingness and reasons looks *ad hoc*, with the second disjunct of the bi-conditional introduced only to ensure extensional adequacy (to preserve the existence of state-given reasons for propositional attitudes).

Perhaps one might want to defend the introduction of the second disjunct by pointing out that on Howard's view, this makes sense of reasons' being sensitive to (changes in) value. Reasons' sensitivity to value is explained by the underlying

[40] A term I borrow from Rabinowicz (2008 & 2012) to describe a reduction in the number of categories of concepts or properties in a particular (e.g. normative) domain.

relation between fittingness and value on the one hand and fittingness and reasons on the other. If it is fitting to favour *x*, then *x* is good, according to Howard. And if it is fitting to favour favouring *x*, then favouring *x* is good. Correspondingly, there is a reason to favour *x*, namely that *x* is good. And favouring *x* itself turns out to be good when there is a reason to favour favouring *x*.

However, if, as I have argued, there is no bi-conditional equivalence between its being fitting to favour *x* and *x*'s being good, then the relation between reasons and fittingness, if there is one, does nothing to explain whatever relation there is between reasons and value. The loss of theoretical unity and explanatory unity seems to sap the independent motivation for accepting reasons as fittingness, making it look like it is an *ad hoc* principle designed to ensure extensional adequacy alone.

To this end, I am more strongly inclined to think that a view like that offered by Conor McHugh and Jonathan Way[41] is well supported by considerations of theoretical unity, despite still being false. On their view, one has a reason to desire *x* only if it is fitting to desire *x*, excluding Howard's additional disjunct that there is a reason to desire *x* if there is a reason to desire to desire *x*. They stand with philosophers such as Derek Parfit and John Skorupski in suggesting that all reasons are reasons of the right kind for the fitting-attitude analysis.[42] And according to McHugh & Way, this fact is meant to be explained by the primacy of fittingness.

Whether one favours the Howard-style approach or the McHugh & Way-style approach to fittingness-first, there is a basic problem that neither account can avoid. Fittingness does not do the work of reasons. The central feature that underlies reasons, oughts, and other properly normative properties is that they are guiding in some loose sense. This sense is loose enough that it need not include any link between being (potentially) motivated by a consideration and that consideration being a reason, but not so loose that the realisability condition is violated. Note that I am not assuming that not violating the realisability is sufficient for possessing guidingness. This seems to put a fittingness-before-reasons view onto the horns of a dilemma. If fittingness is not a properly normative property, then there is more to something's being a reason than its being fitting: a new feature, guidingness, is added. On the other hand, if fittingness is as guiding as reasons, fittingness then looks rather like a normative property, perhaps so much so that one doubts that there is anything more to being fitting than being a reason that obtains in virtue of certain kinds of relations between an attitude and its contents. In that case, it is most natural to interpret reasons as being prior to fittingness, perhaps making fittingness reducible to reasons.

If the arguments in the rest of this chapter are correct, and fittingness is not prior to value, then there seems to be no special reason to believe that fittingness is in general more basic in the relevant sense than other non-descriptive properties.

---

[41] McHugh & Way (2016 & 2022).

[42] Parfit (2001) and Skorupski (2002 & 2010) take these reasons to be object given reasons. However, the spirit of their views and that of McHugh and Way are much the same.

This is clearly not a conclusive argument against the claim that fittingness is prior to reasons. However, properly normative notions are central to much of our ethical and even epistemological theorising, and if we are not willing to abandon the weak guidingness that I claim is the characteristic feature of the normative, then it is difficult to see how fittingness will in any interesting sense be prior to reasons. Perhaps the reverse is true as well, but I shall let the matter rest there.

4. Conclusion

In this paper I have argued that value is not analysable in terms of reasons or fittingness, due to the extensional inadequacy of such analyses. The fact that value over-generates for fittingness also means that fittingness cannot be reduced to value. This is sufficient to show that *-first* views that have the ambition to reduce two of fittingness, reasons, and value to the remaining third property category are false. I have not taken up the interesting question of the right aims or ambitions of *-first* projects. If the arguments here are correct, that is unnecessary. The least ambitious version of the *-first* projects is to provide adequacy conditions for all non-descriptive properties in terms of just one non-descriptive property, even when there are no analytic or reductive ambitions in play. Even this least ambitious project cannot survive the falsification of the relevant bi-conditional claims. More ambitious projects will necessarily imply more, and are *a fortiori* also false.

The arguments in §3 are incomplete, but perhaps suggestive of the claim that reasons cannot be analysed in terms of fittingness. Whether the reverse is true is uncertain, but I see no special grounds for optimism that such an analysis is possible.[43]

Works Cited

Adams, Douglas (1981). *The Hitchhiker's Guide to the Galaxy*. New York: Pocket Books.

Broome, John (2004). *Weighing Lives*. Oxford, Oxford: Oxford University Press.

Bykvist, Krister (2009). No Good Fit: Why the Fitting Attitude Analysis of Value Fails. *Mind* 118 (469): 1-30.

Bykvist, Krister (2015). Reply to Orsi. *Mind* 124 (496): 1201-1205.

[43] I would like to thank Krister Bykvist, Jonas Olson, and Toni Rønnow-Rasmussen individually for their invaluable comments on earlier versions of this paper. The paper has been improved significantly due to comments from audiences at Lund University, Stockholm University, Uppsala University, and the University of Neuchâtel and from two anonymous referees for this volume. This paper was written with the generous support of Vetenskapsrådet for the project Pragmatism, Pluralism, and Reasons for Belief.

Dancy, Jonathan. (2000). Should We Pass the Buck? *Royal Institute of Philosophy Supplement* 47: 159-173.

Danielsson, S. and Olson, J. (2007). Brentano and the Buck-Passers. *Mind* 116 (463): 511-22.

Gaifman, Haim. (1983). Paradoxes of Infinity and Self-Applications, I. *Erkenntnis* 20 (2):131-155.

Gjelsvik, Olav (2020). Reason and Oughts: Fundamentals with the Normative. *Acta Philosophica Fennica* 96: 155-177.

Heathwood, Christopher. (2008). Fitting Attitudes and Welfare. *Oxford Studies in Metaethics* 3: 47-73.

Howard, Christopher. (2019). The Fundamentality of Fit. *Oxford Studies in Metaethics* 14: 216-235.

Kiesewetter, B. (2018). Contrary-to-Duty Scenarios, Deontic Dilemmas, and Transmission Principles. *Ethics* 129 (1): 98-115.

Lang, Gerald. (2008). The Right Kind of Solution to the Wrong Kind of Reason Problem. *Utilitas*, 20 (4): 472-89.

McHugh, C. & Way, J. (2016). Fittingness First. *Ethics* 126 (3):575-606.

McHugh, C. & Way, J. (forthcoming). *Getting Things Right: Fittingness, Reasons, and Value*. Oxford: OUP.

Olson, Jonas (2009a). The Wrong Kind of Solution to the Wrong Kind of Reason Problem. *Utilitas* 21 (2):225-232.

Parfit, Derek (2001). Rationality and Reasons. In Egonsson, Dan, et al. (eds.) *Exploring Practical Philosophy: From Action to Values.* Aldershot: Ashgate: 17-39.

Rabinowicz, Wlodek (2008). Value Relations. *Theoria* 74 (1): 18-49.

Rabinowicz, Wlodek (2012). Value Relations Revisited. *Economics and Philosophy* 28 (2): 133-164.

Rabinowicz, W. and Rønnow-Rasmussen, T. (2004). The Strike of the Demon: On Fitting Pro-attitudes and Value. *Ethics*, 114 (3): 391-423.

Rabinowicz, W. & Rønnow-Rasmussen, T. (2006). Buck-passing and the right kind of reasons. *Philosophical Quarterly* 56 (222):114-120.

Reisner, Andrew (2009). Abandoning the Buck Passing Analysis of Final Value. *Ethical Theory and Moral Practice* 12 (4): 379-395.

Reisner, Andrew (2015). Fittingness, Value, and Trans-World Attitudes. *Philosophical Quarterly* 260: 1-22.

Risberg, Olle (2018). The Entanglement Problem and Idealization in Moral Philosophy. *Philosophical Quarterly* 68 (272): 542-559.

Risberg, Olle (2020). *Guiding Concepts: Essays on Normative Concepts, Knowledge, and Deliberation*. Thesis, Uppsala University.

Rosenqvist, Simon (2020). *Hedonic Act Utilitarianism: Action Guidance and Moral Intuitions*. Thesis, Uppsala University.

Rønnow-Rasmussen, Toni (2007). Analysing Personal Value. *The Journal of Ethics* 11 (4):405-435.

Rønnow-Rasmussen, Toni (2011). *Personal Value*. Oxford: Oxford University Press.

Rønnow-Rasmussen, Toni (2021). *The Value Gap*. Oxford: Oxford University Press.

Schellekens, E. (2006). Towards a Reasonable Objectivism for Aesthetic Judgements. *British Journal for Aesthetics* 46 (2): 163-177.

Skorupski, J.M. (2002). The Ontology of Reasons. *Topoi* 21 (1-2): 113-124.

Skorupski, J.M. (2010). *The Domain of Reasons*. Oxford: Oxford University Press.

Soames, Scott (2002). *Beyond Rigidity: The Unfinished Semantic Agenda of" Naming and Necessity"*. Oxford: OUP.

Sylvan, Nils (2021). *Fittingness and Partiality: On the Partiality Problem for the Fitting Attitude Account of Value*. Doctoral thesis, Stockholm University.

Tonicing Moral Supervenience

Caj Strandberg

Abstract. Inspired by a novel remark by Toni Rønnow Rasmussen, I compare realist and non-cognitivist accounts of the modality of moral supervenience. Toni's remark suggests that non-cognitivists face a critical choice: If they opt for weak supervenience, they have difficulties to account for the dependence of the moral on the non-moral. If they opt for strong supervenience, they have difficulties to account for the inner 'necessary'. However, non-cognitivism seems to have an important advantage: It can explain *why* the outer 'necessary' is analytical by reference to the function of moral language to influence behaviour. According to the preferred realist account of moral supervenience, it amounts to strong supervenience where the outer 'necessary' is analytical and the inner metaphysical. Most importantly, I argue that realism can explain *why* 'necessary' in moral supervenience needs to be understood in accordance with this view by reference to the connection between moral properties and normative reasons. Moreover, I argue that the realist account can be generalized to other normative properties and that it is part of an explanation of why moral language can have the function to influence behaviour. Thus, realism provides a superior account of the modality of moral supervenience as compared to non-cognitivism.

1. Introduction

In a wonderful metaphor, Toni Rønnow-Rasmussen observes that 'Values are not like butterflies that happen to settle on a flower' (Rønnow-Rasmussen (2006): 2).

The moral depends on the non-moral: It is necessarily the case that moral terms apply to objects because, or in virtue of, their having non-moral properties. It is generally agreed that a necessary condition for this dependency relation is that the moral supervenes on the non-moral. There agreement ends, however. In particular, it is commonly assumed that realism has difficulties to account for the modality of moral supervenience, whereas non-cognitivism is able to do so. In this paper, I compare realist and non-cognitivist views of moral supervenience. The result of the discussion is that the converse is the case: Realism provides a superior account of the modality of moral supervenience. Thus, the paper provides an argument for realism and against non-cognitivism based on supervenience.

2. Realism and Non-cognitivism

As I will understand *moral realism*, it amounts to three claims: (**i**) Cognitivism: Moral judgments consist in beliefs that ascribe moral properties to objects. (**ii**) Moral properties are instantiated such that some moral judgments are true. (**iii**) Moral properties are mind independent: Their nature is not counterfactually dependent merely on the mental attitudes of individual agents.[1]

Thus understood, there are different versions of realism. On *reductionist realism*, moral properties are identical to non-moral properties, i.e. properties that can be fully defined without employing moral terms.[2] On *non-reductionist realism*, moral properties are not identical to non-moral properties thus understood. On *naturalist realism*, moral properties consist in natural properties. There are both reductionist and non-reductionist versions of naturalist realism.[3] On *non-naturalist realism*, moral properties are *sui generis* and not identical to any other type of properties. In this paper, I use 'realism' to refer to the generic sense of realism rather than any particular version of it. Thus, my discussion applies, *mutatis mutandis*, to all the mentioned versions of realism.

As I will understand *moral non-cognitivism*, it amounts to two claims: (**i**): Moral judgments do not consist in beliefs that ascribe moral properties to objects. (**ii**) Instead, moral judgments consist, wholly or partly, in non-cognitive attitudes, such as desires.

[1] I provide an account of mind independence in Strandberg (Forthcoming).

[2] For this understanding of non-moral properties, see e.g. Hare (1997: 64); Railton (1989: 160); Sayre-McCord (1997a; 281), and Timmons (1999: 48).

[3] On *reductionist naturalism*, moral properties consist in natural properties that can be fully defined without employing moral terms. On *non*-reductionist *naturalism*, they consist in natural properties that cannot be thus defined.

3. The Importance of Moral Dependence

It is plausible to assume that a condition for being competent with the meaning of moral terms is to acknowledge that they apply to objects in virtue of their having non-moral properties. Assume that an agent makes statements indicating that she does not believe that an action being morally right depends on some of the action's non-moral properties. She is then committed to admitting that it would be correct to judge that the action is right even if it does not have any non-moral properties.[4] More importantly, she is committed to admitting that two actions can differ as regards rightness in spite of not differing in any non-moral properties. We would presumably regard her statements as an indication that she is not linguistically competent with 'right'.

The competence with the meaning of moral terms involves other aspects than recognition of the mentioned dependence relation. However, some of these aspects are presumably to be explained with reference to it. Assume that an agent justifies her judgment that an action is right by citing some of its non-moral properties. She can then be understood as pointing at some non-moral properties in virtue of which 'right' is applicable. Similarly, consider an agent who maintains that the fact that a person is good explains why the person performed a certain action and then justifies the explanation by citing some of the person's non-moral properties. She can then be understood to point at some non-moral properties in virtue of which 'good' applies. These aspects underwrite how important it is for a metaethical view to be able to explain the dependence of the moral on the non-moral.

As indicated, it is common in philosophy to characterize dependence relations in terms of supervenience. However, while supervenience reasonably is a necessary condition for a dependence relation to hold, it might not be sufficient.[5] In particular, supervenience might be insufficient to account for a metaphysically explanatory and asymmetrical aspect of the dependence relation between properties. Accordingly, it has recently been argued that the notion of grounding is needed to account for dependence, at least on non-naturalism.[6] In this paper, I will consider moral supervenience on the assumption that it is necessary to account for moral dependence, but recognize that it might need to be supplemented by further notions so as to fulfil this task. Importantly, if my argument that non-cognitivism is unable to account for the modality of moral supervenience is correct, it follows that this view also is unable to account for moral dependence.

[4] However, it might be objected that it is inconceivable, and hence not analytically possible, that there are objects that lack *any* non-moral properties.

[5] However, see Strandberg (2008: 129–158).

[6] See e.g. Rosen (2010: 109–135) and Leary (2017: 76–105).

4. Realist Supervenience

As realism maintains that moral judgments ascribe moral properties to objects, it can characterize moral supervenience directly by reference to connections between properties. Consider first:

> *Realist Weak Supervenience* (**RWS**)*:* It is necessary that, for any object x, and for any moral property M, if x is M, then there is some set of non-moral properties G (G_1, G_2, G_3,…) such that (**a**) x has G, and (**b**) for any object y, if y has G, then y is M.

The outer 'necessary' binds the formula as a whole.[7] In weak supervenience, there is no inner 'necessary' that prefixes the implication in (**b**). Thus, it does not extend to all possible worlds. Consider next:

> *Realist Strong Supervenience* (**RSS**)*:* It is necessary that, for any object x, and for any moral property M, if x is M, then there is some set of non-moral properties G (G_1, G_2, G_3,…) such that (**a**) x has G, and (**b**) it is necessary that, for any object y, if y has G, then y is M.

As before, the outer 'necessary' binds the formula as a whole. In strong supervenience, there is an inner 'necessary' that prefixes the implication in (**b**). Thus, it says that it holds in all possible worlds that any object which has G has M.

It is plausible to think that realists should opt for strong supervenience rather than weak. The primary reason is that weak supervenience is too weak to be part of an account of the notion that the moral depends on the non-moral.[8] Assume that we want to claim that an action is right because it has a certain set of non-moral properties G. Weak supervenience merely states that *within* a possible world any action that has G is right. This indicates that, on weak supervenience, it would be mistaken to claim that an action is right because it is G, since an action *could* have G and yet not be right. In that case, the co-instantiation of rightness and G does not seem to be a matter of dependence, but rather coincidence: actions that have G *happen* to be right. Strong supervenience provides the required supplement by stating that it holds in all possible worlds that any action that has G is right.

Furthermore, much of our moral thinking is constituted by thought experiments. Assume that one wonders whether the fact that an action causes happiness is relevant as to whether it is right. One might then ask if the action would be right in a possible world where it does not cause happiness. We often trust the results of such thought experiments and let our moral decisions be guided by them. However, if only weak supervenience is the case, thought experiments would not be of any

[7] For ease of exposition, I will refer to 'outer' 'necessary' also when discussing weak supervenience.

[8] Cf. Kim (1993 (1990): 143–144). See also e.g. Blackburn (1993 (1985): 132); Dreier (2015: 275–276), and Franzén (Forthcoming: 7–8).

help, since we would not be justified to hold beliefs about one possible world based on what we believe about other possible worlds. For example, we would not be justified to believe that causing happiness contributes, or fails to contribute, to actions being right in the same way in our world as it does in the possible worlds employed in our thought experiments. However, on strong supervenience we would be justified to trust the result of such thought experiments, since what is the case in one world extends to other worlds.

It might next be asked how the two occurrences of 'necessary' in strong supervenience should be interpreted. As noticed, it is a requirement on being competent with the meaning of moral terms to acknowledge that they apply in virtue of objects having non-moral properties. Thus, it is generally agreed that the outer 'necessary' needs to be understood as analytical necessity. By contrast, it does not seem plausible to understand the inner 'necessary' in this manner. One reason is it does not seem to be a matter of linguistic competence to know about a set of non-moral properties G that if an object has G, it is right. Another reason is that some instances of necessary implications from non-moral properties to moral properties constitute moral principles. However, it might be argued that such principles cannot be analytically necessary, since it would mean that they would lack normativity. Instead, it seems more plausible to think that the inner 'necessary' should be understood as metaphysical necessity. I will return to these points.

5. Non-cognitivist Supervenience

According to non-cognitivism, moral judgments do not ascribe moral properties to objects, but consist in non-cognitive attitudes. As a result, it cannot characterize moral supervenience by reference to any metaphysical relation between properties. Instead, it is accounted for in terms of the connection between an agent's moral attitudes and her beliefs about what non-moral properties objects have. It is maintained that to be competent with the meaning of moral terms, an agent needs to be consistent in having the same moral attitude towards objects that she believes have the same non-moral properties.[9]

It is rarely stated clearly how moral supervenience should be understood according to non-cognitivism.[10] In what follows, I suggest ways of stating moral supervenience on this view. Assume that an agent's moral judgment to the effect that x is right consists in her having moral attitude M towards x. Weak supervenience might be formulated as follows:

[9] See e.g. Hare (1952: 131–134) and Blackburn (1993 (1985): 136–137, 146).

[10] But see Gibbard (2003: 90). However, Gibbard's formulation is concerned with his particular version of non-cognitivism.

> *Non-Cognitivist Weak Supervenience* (**NWS**)*:* It is necessary that, for any object x, and for any agent S, if S has moral attitude M towards x, then there is some set of non-moral properties G (G_1, G_2, G_3,…) such that (**a**) S has attitude M towards x because she believes that x has G, and (**b**) for any object y, if S believes that y has G, then S has attitude M towards y.

Similarly, strong supervenience can be formulated as follows:

> *Non-Cognitivist Strong Supervenience* (**NSS**)*:* It is necessary that, for any object x, and for any agent S, if S has moral attitude M towards x, then there is some set of non-moral properties G (G_1, G_2, G_3,…) such that (**a**) S has attitude M towards x because she believes that x has G, and (**b**) it is necessary that for any object y, if S believes that y has G, then S has attitude M towards y.

It follows on both versions that in case an agent does not comply with them, she does not make a moral judgment, since she does not have a moral attitude of which such a judgment is constituted.

We might now query whether non-cognitivists should adopt weak or strong supervenience. Importantly, in this regard non-cognitivists face a critical choice. As far as I know, Rønnow-Rasmussen was first to pay attention to it:

> The real crux of the matter concerns therefore how a prescriptivist would account for the second necessity operator in the strong supervenience thesis. It is one thing to claim that in endorsing Va [an object a having value V] we commit ourselves by conceptual necessity to subscribe to a principle like the one in premise p [for all x, if Nx, then Vx]. It is quite another thing to say that endorsing Va commits you, by conceptual necessity, to subscribe to a principle that in part expresses that there holds a necessity relation between certain natural properties and a certain value property. The latter claim squares badly with his idea that value terms have no fixed descriptive content. (Rønnow-Rasmussen (2006: 8))[11]

In the frame of the present discussion, we can formulate the choice in the following manner. *On the one hand*, non-cognitivists have reason to adopt strong supervenience. Contemporary non-cognitivists generally concede that realism in many respects seems to be in line with how we talk and think about morality. Consequently, they try to save as much as possible of the appearance of realism while arguing that, ultimately, non-cognitivism is to be preferred.[12] As we have seen, realism should adopt strong supervenience to capture the notion that the moral depends on the non-moral. Thus, insofar as non-cognitivists aim to account for this notion in a way that accords with our conception of it, they should adopt strong

[11] Cf. Dreier (2015: 289–290), and Franzén (Forthcoming: 8). See also Rønnow-Rasmussen (1993: 142–151). Rønnow-Rasmussen's discussion is concerned with Hare's prescriptivism, but applies to non-cognitivism in general.

[12] See e.g. Blackburn (1984: Ch. 6) and (1993 (1988): 166–181).

supervenience. In case they do not, they need to provide a particular argument why weak supervenience should be preferred to strong. *On the other hand*, it is difficult for non-cognitivism to adopt strong supervenience. As we have seen, there are reasons to understand the inner 'necessary' as metaphysical rather than analytical necessity. A metaphysical necessary connection is a connection that holds between properties or facts. For example, an account of why it is metaphysically necessary that if an object has certain properties, it has a certain other property, is provided by reference to the nature of the properties referred to in the antecedent. However, non-cognitivists explain moral supervenience in terms of the connection between attitudes and beliefs about non-moral properties. Consequently, it is difficult to see that they can understand the inner 'necessary' as metaphysical necessity.[13] It appears that the only remaining alternative is to interpret it as analytical necessity, which appears implausible for reasons mentioned above. In that case, it might be more plausible for non-cognitivists to refuse strong supervenience and argue for weak supervenience.

We have already touched on how non-cognitivists understand the outer 'necessary' in moral supervenience. They maintain that to be competent with the meaning of a moral term, an agent needs to be consistent in her moral attitudes in a manner complying with supervenience. Thus, as Rønnow-Rasmussen observes, they understand the outer 'necessary' as analytical necessity.

It was argued above that non-cognitivists face a critical choice as to whether they should opt from weak or strong supervenience. However, non-cognitivism seems to have a crucial advantage over a realist account of moral supervenience: It is able to explain *why* the outer 'necessary' needs to be understood as analytical necessity. A plausible idea motivating non-cognitivism is that an essential function of moral language is to influence attitudes and actions.[14] It is reasonable to argue that in order for moral language to fulfil this function, it needs to be a condition on linguistic competence that we are consistent in our attitudes in a way conforming to supervenience.

To summarize: There are good reasons for realists to adopt strong supervenience where the outer 'necessary' is analytical and the inner metaphysical necessity. Because realists opt for strong supervenience, they are in the position to account for

[13] It might perhaps be argued that non-cognitivists can employ a deflationary view of metaphysical necessity to account for the inner 'necessary'. On this view, it does not involve any 'worldly' metaphysical relation between properties that is incompatible with non-cognitivism. I do not have space to evaluate this important suggestion in the present paper, but will merely make two comments. First, I doubt that a deflationary view of metaphysical necessity is able to capture the contention that moral principles are substantive, and hence non-linguistic, as I will argue in the next section. The reason is that non-cognitivists seem committed to explaining a deflationary notion of metaphysical necessity ultimately in linguistic terms. Second, it would be incompatible with one of the main arguments for non-cognitivism which relies on weak supervenience (Blackburn (1993 (1985): 166–181).

[14] See e.g. Blackburn (1984: 186) and (1993 (1985): 137).

the notion that the moral depends on the non-moral. Non-cognitivists face a crucial choice. If they opt for weak supervenience, they have difficulties to account for the notion that the moral depends on the non-moral. If they opt for strong supervenience, they have difficulties to account for the inner 'necessary'. However, non-cognitivism seems to have an important advantage over realism in that it can explain *why* the outer 'necessary' in moral supervenience needs to be analytical.

In the remainder of the paper, I argue that realism can explain *why* 'necessary' in moral supervenience needs to be read in accordance with this view. Moreover, I suggest that this view is generalizable to other normative properties and that it is part of an explanation why moral language can have the function to influence attitudes and actions.

6. A Realist Explanation of the Modality of Moral Supervenience

According to non-cognitivism, moral judgments consist in non-cognitive attitudes. As we express our moral judgments and thereby our moral attitudes in using moral language, it can have the function to influence attitudes and actions. In Simon Blackburn's view, moral language has the function of influencing other people to have the same attitudes as we do so as to coordinate our attitudes and thereby our actions.[15] In order for it to have that function, we need to be consistent in our moral attitudes.[16] As a result, to be linguistically competent with the meaning of moral terms, an agent needs to adhere to moral supervenience by being consistent in having the same moral attitude towards objects that she believes have the same non-moral properties. Hence, the outer 'necessary' in moral supervenience amounts to analytical necessity.

According to realism, by contrast, moral judgments do not consist in non-cognitive attitudes. It therefore seems that it cannot explain *why* the outer 'necessary' in moral supervenience is analytically necessary by referring to the mentioned function of moral language. Moreover, this explanation seems to square badly with understanding the inner 'necessary' as metaphysical necessity. Thus, it might be argued that while non-cognitivism has a straightforward explanation of their understanding of 'necessary' in moral supervenience, realism lacks such an account.

However, I think realists are in the position to provide an explanation of *why* 'necessary' in moral supervenience is to be understood in the manner they suggest.

[15] See e.g. Blackburn (1998: 68–69). Cf. Hare (1952: 131–134) and Gibbard (2003: 56, 89–94).

[16] However, these assumptions can be questioned. See e.g. Zangwill (1997: 510–511); Sturgeon (2009: 83–88), and Atiq (578–599).

In my view, realists can provide such an explanation by reference to the connection between moral properties and reasons. Moreover, this account is available to all versions of realism. The basic idea is this: The outer 'necessary' is analytical because the meaning of a sentence to the effect that an object has a moral property entails that there is some *moral reason*, where such a reason is constituted by non-moral properties on which the moral property strongly supervenes. The inner 'necessary' is metaphysical because *moral principles*, in the form of implications from non-moral properties constituting moral reasons to moral properties, are *substantive*. The account rests on three claims that I will consider in turn.

First, a normative standard exhibits an analytically necessary connection between normative properties and normative reasons. One instance of this connection is the following:

> *Moral Property→Moral Reason*: It is analytically necessary that if an object x has a moral property M, then there is a moral reason pertaining to x.

There are several instances of this connection. For example: If it is morally right to perform an action, then there is a moral reason to perform that action.

It should be uncontroversial that morality is a normative standard that exemplifies this connection between normative properties and reasons. Indeed, terms denoting moral properties, such as 'right', 'wrong', 'good', and 'bad', are frequently used to entail that there is moral reason (not) to perform certain actions or (not) to have certain attitudes.

Second, on realism it is analytically necessary that moral reasons are constituted by the supervenience base of moral properties:

> *Moral Reason/Supervenience*: It is analytically necessary that a moral reason is constituted by a set of non-moral properties G on which a moral property M strongly supervenes.

There are several instances of this claim. For example: A moral reason to do what is morally right is constituted by a set of non-moral properties on which moral rightness strongly supervenes.

The present claim should not be understood to entail that a moral reason is constituted by *all* the non-properties included in a set of non-moral properties on which a moral property strongly supervenes. For instance, such a set might include properties on which rightness supervenes but that do not make up parts of a moral reason to do what is right. One example is 'enablers': properties that do not constitute reasons but which are relevant for whether an agent has a reason.[17] Hence, the claim is compatible with a moral reason to do what is right being constituted by a *subset* of the non-moral properties on which rightness strongly supervenes.

[17] Cf. Strandberg (2008: 138–147).

It should be uncontroversial that realism is committed to this connection between moral reasons and the supervenience base of moral properties. There are both intuitive and more formal grounds for this contention. The intuitive ground: A reason to do what is right consists in some feature of the action that 'makes' it right or 'in virtue' of which it is right. Similarly, an action is right 'because' there is a reason to perform it. Now, terms like 'make', 'in virtue of', and 'because' denote a dependence relation between moral properties and underlying properties that realists try to capture by strong supervenience. The more formal ground: It is analytically necessary that if an action is morally right, then there is a moral reason to perform the action. On realism, the property of being right depends, and hence strongly supervenes, on a set of non-moral properties. It then seems very plausible to think that a moral reason consists of some set of non-moral properties. Thus, on realism a moral reason to do what is right is constituted by a set of non-moral properties on which rightness strongly supervenes.

There is a further ground to accept the claim above based on the notion of normative explanation: An explanation of why an object has a certain normative property, for instance why an action has the property of being morally right. Pekka Väyrynen has argued, convincingly in my view, that normative explanations are subject to a 'justification condition'. Applied to the present example, this condition implies that a normative explanation of why an action is right needs to identify some feature of the action that provides a normative reason to perform it.[18] Väyrynen does not explicitly comment on moral supervenience. However, it is plausible to assume that a normative explanation of why an action is right refers to non-moral properties on which rightness depends and hence strongly supervenes.

Third, moral principles are substantive:

> *Moral Principles are Substantive*: A moral principle of the form 'If x has a set of non-moral properties G, where G constitutes a moral reason pertaining to x, then x has a moral property M' is not analytically necessary but metaphysically necessary.[19]

A simple example of a moral principle of this form would be the following: 'If an action maximizes happiness, then it is morally right'. In the present paper, I do not commit myself to this or any other moral principle.

As mentioned, it is plausible to think that a moral principle of the relevant type refers to a set of non-moral properties which constitutes a moral reason to do what is morally right. However, it is implausible to think that such a principle is

[18] Väyrynen (2021b: 3–22). See also Väyrynen (2021a: 278–927).

[19] It should be noticed that in the present sense of 'moral principle', the existence of such principles is compatible with particularism. A moral principle, as this notion is used here, merely constitutes a necessary implication from a set of non-moral properties, comprising a moral reason, to a moral property. The set might be very complex and the properties in it interact in ways maintained by particularists. Cf. Strandberg (2008: 129–158).

analytically necessary. First, it does not seem to be part of competence with the meaning of 'right' to know that there is an implication from a given set of non-moral properties to rightness. Second, and more controversially, it might be argued that such principles cannot be analytically necessary, since it would imply that they lack normativity. Assume that moral principles are analytically necessary. In that case, it would be a matter of the meaning of 'right' that such a principle is true: We use 'right' in such a way that if an action has a particular set of non-moral properties, it is correct to apply 'right' to it. However, whether we have a moral reason to perform actions that have certain non-moral properties does not seem to be a matter of meaning of words. Instead, it is a matter of the nature of the non-moral properties in question. In the example above, it is the nature of maximizing happiness which would explain why it is the case that if an action has this non-moral property, there is moral reason to perform it. If this is correct, there are grounds to think that moral principles need to be metaphysically rather than analytically necessary.

Advocates of realism are now in the position to provide an explanation of *why* 'necessary' in moral supervenience should be understood in accordance with this view. That the outer 'necessary' is analytical necessity follows from the connection between moral properties and reasons, and from moral reasons being constituted by non-moral properties on which moral properties strongly supervene. That the inner 'necessary' is metaphysical necessity follows from moral principles being metaphysically necessary. In more detail: According to *Moral Property→Moral Reason*, it is analytically necessary that if an object x has a moral property M, then there is a moral reason pertaining to x. According to *Moral Reason/Supervenience*, it is analytically necessary that a moral reason is constituted by a set of non-moral properties G on which M strongly supervenes. It follows that the outer 'necessary' amounts to analytical necessity. According to *Moral Principles are Substantive*, moral principles are not analytically but metaphysically necessary. A moral principle maintains that if an object has a set of non-moral properties G, which constitutes a moral reason pertaining to x, then x has a moral property M. It follows that the inner 'necessary' amounts to metaphysical necessity.

As indicated, this account of the modality of moral supervenience is available to all forms of realism mentioned above. However, it is worth mentioning that it might have implications for a certain argument against non-naturalism. One objection against this view is that it is unable to account for the supervenience of *sui generis* moral properties on non-moral properties. It is important to distinguish between different versions of this argument. On one version, there cannot be any necessary connection between properties that belong to entirely distinct types. On another version, non-naturalism is unable to account for the modality of the supervenience of moral properties on non-moral properties because moral properties are *sui generis* on this view. The above account does not offer any response to the first argument. However, it might provide a response to the second one. There does not seem to be any reasons to think that the modality of the supervenience of moral properties on non-moral properties is different because the former are conceived of as *sui generis*.

The reasoning above, motivating why the outer 'necessary' is analytical and the inner metaphysical, seems applicable irrespective of whether moral properties are understood as *sui generis* or not.

7. Unifying the Normative Sphere

In the last section, I argued that realism can explain *why* 'necessary' in moral supervenience should be understood in the manner proposed by this view. Now I would like to briefly indicate that the realist account of moral supervenience has a significant advantage over the non-cognitivist: The realist account can, in contrast to the non-cognitivist alternative, unify the normative sphere by being generalized to other normative properties. The same type of problems that non-cognitivists have of accounting for moral supervenience are bound to arise regarding supervenience in relation to other types of normative notions. For example, for basically the same reasons as those indicated above, non-cognitivism about aesthetic value will face difficulties to explain how aesthetic value depends, and hence strongly supervenes, on non-aesthetic properties. Moreover, it is not evident that all uses of normative language have the function to influence attitudes and actions. By contrast, it is plausible to argue that all sentences that ascribe a normative property to an object entail the existence of some reason that is constituted by a set of non-normative properties on which the normative property strongly supervenes. Hence, there are grounds to think that the realist account of moral supervenience is generalizable to other normative properties.

8. Explaining the Function of Moral Language

The non-cognitivist explanation of *why* the outer 'necessary' in moral supervenience amounts to analytical necessity is that it is needed to account for the function of moral language to influence attitudes and actions. It should be clear that realism is not committed to the claim that moral language has such a function. However, it is possible for realists to concede that moral language does have this function and then argue that *their* preferred reading of 'necessary' is needed for moral language to fulfil it. While non-cognitivists explain this function by referring to the meaning of moral sentences, realists can account for it by referring to the pragmatics of moral utterances. I have defended this view in other contexts and will only indicate the contours of it here.[20]

[20] For a full defence, see Strandberg (2012: 87–122).

The realist account rests on three assumptions. First, conversations about moral matters generally have the mutually accepted purpose to communicate moral beliefs about which actions are right and wrong. Second, such conversations generally have the further mutually accepted purpose to influence attitudes and actions. Third, a sentence like 'X is right' entails that there is moral reason to perform the action in question. In view of the two purposes of moral conversations, it is plausible to assume that utterances of the type 'X is right' standardly conversationally implicates that the utterer has a favourable attitude towards the action being performed. The basic explanation is that it does not seem to be any *point* in uttering a sentence which entails that there is moral reason to perform an action in a moral conversation which has as a mutually accepted purpose to influence behaviour unless one has a favourable attitude towards it being carried out.[21] As moral utterances standardly conversationally implicate positive or negative attitudes towards actions, they can have the function to influence attitudes and actions.

It is plausible to argue that for moral language to fulfil this function, 'necessary' in moral supervenience needs to be understood as suggested by realism. First, it needs to be analytically necessary that a sentence like 'X is right' entails that there is a moral reason to perform the action. If this were not the case, an utterance to the effect that an action is right made in a moral conversation which has a mutually accepted purpose to influence behaviour would not entail that there is moral reason to perform it. Second, this moral reason needs to be constituted by a set of non-moral properties on which rightness strongly supervenes. If this were not the case, the fact that an agent performs the action that she, according to the utterance, has moral reason to perform would not guarantee that she performs an action that is right. In case this condition is not fulfilled, the utterance would then not be effective in influencing people to perform such actions in various possible circumstances. Thus, it should be clear that both these conditions need to be fulfilled in order for moral utterances to have the function to influence attitudes and actions. Moreover, there are grounds to think that moral principles need to be synthetically rather than analytically necessary. The crucial point is this: A part of the realist account of why moral language can have the function to influence attitudes and actions is that moral utterances entail the existence of moral reasons in accordance with the realist view of moral supervenience, where the first 'necessary' is analytical and the second metaphysical.

[21] It might be objected that it need not be awkward to utter a sentence which entails that there is a moral reason to perform an action without having any favourable attitude towards it, since it might be merely a *pro tanto* moral reason. However, I think it would be awkward to utter such a sentence in the absence of a favourable attitude unless the utterance is accompanied by an additional utterance which modifies it, in which case the implicature of the original utterance is cancelled. After all, there seems to be little point in uttering a sentence which entails that there is even a *pro tanto* moral reason to perform an action in a context with the mentioned purposes unless one has the relevant attitude.

9. Concluding Remarks

In this paper, I have argued that moral realism provides an account of the modality of moral supervenience that is superior to the one offered by moral non-cognitivism. Realists should maintain that moral supervenience amounts to strong supervenience where the outer 'necessary' is analytical and the inner metaphysical. As it is strong supervenience, it can be part of an account of the notion that the moral depends on the non-moral. Most importantly, realists can explain *why* 'necessary' should be understood in this manner that is available to all versions of this view. Furthermore, this account is generalizable to other normative properties. In addition, it is part of an explanation of why moral language can have the function of influencing attitudes and actions. Non-cognitivism has a difficult choice. If it opts for weak supervenience, it cannot account for the notion that the moral depends on the non-moral. If it opts for strong supervenience, it cannot account for the inner occurrence of 'necessary'. Non-cognitivism might seem to have an advantage in being able to explain *why* the outer 'necessary' is analytical. However, the realist account is superior for the reasons indicated above.[22]

References

Atiq, Emad (2020) "Supervenience, Repeatability, and Expressivism". *Noûs* 54(3): 578–599.

Blackburn, Simon (1984) *Spreading the Word*. Oxford: Oxford University Press.

Blackburn, Simon (1993 (1971)) "Moral Realism". *Essays in Quasi-Realism*. Oxford: Oxford University Press.

Blackburn, Simon (1993 (1985)) "Supervenience Revisited". *Essays in Quasi-Realism*. Oxford: Oxford University Press.

Blackburn, Simon (1993 (1988)) "How to be an Ethical Anti-Realist". *Essays in Quasi-Realism*. Oxford: Oxford University Press.

Blackburn, Simon (1998) *Ruling Passions*. Oxford: Oxford University Press.

Dreier, Jamie (2015) "Explaining the Quasi-Real". *Oxford Studies in Metaethics*, ed. by R. Shafer-Landau, Vol. 10 (273-297). Oxford: Oxford University Press.

Franzén, Nils (Forthcoming) "Non-factualism and Evaluative Supervenience". *Inquiry*.

Gibbard, Allan (2003) *Thinking How to Live*. Cambridge: Harvard University Press.

Hare, R.M. (1952) *The Language of Morals*. Oxford: Oxford University Press.

Hare, R.M. (1997) *Sorting Out Ethics* Oxford: Oxford University Press.

[22] I am indebted to two anonymous referees for value comments on an earlier version of this paper.

Kim, Jaegwon (1993 (1990)) “Supervenience as a Philosophical Concept”. *Supervenience and Mind*, Cambridge: Cambridge University Press.

Leary, Stephanie (2017) “Non-Naturalism and Normative Necessities”. *Oxford Studies in Metaethics*, ed. by R. Shafer-Landau, Vol. 12 (76-105). Oxford: Oxford University Press.

Railton, Peter (1989) “Naturalism and Prescriptivity”. *Social Philosophy and Policy* 7(1): 151-174.

Rønnow-Rasmussen, Toni (1993) *Logic, Facts and Representation. An Examination of R.M. Hare’s Moral Philosophy*. Lund: Lund University Press.

Rønnow-Rasmussen, Toni (2006) “Dislodging Butterflies from the Supervenient”. *Philosophical Anthropology*, Vol. IX, ed. by Stephen Voss.

Rosen, Gideon (2010) “Metaphysical Dependence: Grounding and Reduction”. In *Modality: Metaphysics, Logic and Epistemology*, ed. by B. Hale and A. Hoffman (109–135). Oxford: Oxford University Press.

Sayre-McCord, Geoffrey (1997) “‘Good’ on Twin Earth”. *Philosophical Issues*. ed. by E. Villanueva, Vol. 8, Truth (267-292).

Strandberg, Caj (2008) “Particularism and Supervenience”. *Oxford Studies in Metaethics*, ed. by R. Shafer-Landau, Vol. 3 (129-158). Oxford: Oxford University Press.

Strandberg, Caj (2012) “A Dual Aspect Account of Moral Language”. *Philosophy and Phenomenological Research*, 84(1): 87-122.

Strandberg, Caj (Forthcoming) “Moral Properties”. In *Routledge Handbook of Properties*, ed. by A.R.J. Fisher and A.-S. Maurin. London: Routledge.

Sturgeon, Nicholas (2009) “Doubts about the Supervenience of the Evaluative”. *Oxford Studies in Metaethics*, ed. by R. Shafer-Landau, Vol. 4 (53-90). Oxford: Oxford University Press.

Timmons, Mark (1999) *Morality without Foundations*. Oxford: Oxford University Press.

Väyrynen, Pekka (2021a) “Normative Explanation Unchained”. *Philosophy and Phenomenological Research*, 103(2): 278-297.

Väyrynen, Pekka (2021b) “Normative Explanation and Justification”. *Noûs*, 55(1): 3-22.

Zangwill, Nick (1997) “Explaining Supervenience: Moral and Mental”. *Journal of Philosophical Research*, 22: 509-518.

Do We Have Obligations to Collectives?

András Szigeti[1]

Abstract. I argue that we can have obligations towards collectives that are non-distributive and irreducible to obligations towards individual members. This is because we can discharge obligations towards the collective by treating different configurations of individual members in the required way. This means that the obligation is directed at the collective, not any given individual member. This account respects ontological individualism since we still discharge the obligation towards collectives by treating individuals in certain ways. Two additional burden-of-proof considerations support the main argument. First, if collectives cannot be obligees, then either collectives cannot have obligations, or we must reject the plausible *obligation reciprocity thesis* according to which if *X* can have obligations, then we can have obligations towards *X.* Second, if collectives cannot be obligees, we will have to explain why collectives are like individuals in certain normative domains (e.g., in being fit to be morally responsible), but not in others.

Introduction

One can observe a surprising asymmetry in the literature on groups and obligations. A number of authors have argued that certain collectives can have obligations and that these obligations are not reducible to the sum of obligations incurred by individual members of such collectives (Held 1970; Copp 2007; Isaacs 2011;

[1] While this paper cites only one recent work by Björn, it comes out of projects inspired by him in too many ways to enumerate. My intellectual and personal debt not just to him, but also to Dan and Toni for their friendship and mentorship over the years is enormous.

Lawford-Smith 2012; Wringe 2016; Collins 2017; Tamminga & Hindriks 2019). On the other hand, we find much less discussion of the nature of obligations one can have *towards* collectives. For example, List & Pettit (2011) state that certain collectives operate in the space of *mutual* obligations. They argue that all parties in this space "must acknowledge that others occupy a reciprocal status: they too may address claims, expect compliance, and make compelling complaints about failures" (173). However, for all the talk of mutuality, they then largely limit their discussion to the obligations certain collectives may have. There is hardly any mention of obligations owed *to* collectives. Other collectivists about obligations do not even raise this issue.

At the same time, it would be surprising if groups could have obligations distinct from the obligations of their members but were categorically unqualified to be "obligees". While the mutuality of obligations need not entail *equal* standing in terms of what exactly is owed to any given participant in the regime of obligations, it surely entails that all the participants in this regime do not only qualify as obligation-holders, but also as parties towards whom obligations can be owed under certain circumstances. So, to make sense of mutuality we must show that obligations may be held *vis-à-vis* collectives, whereby the talk of obligations is not just a mere shorthand for the sum of obligations held by individual members of the collective.

This is what I set out to do in this paper. Specifically, I want to defend the claim that we can have irreducible and non-distributive obligations towards collectives. I will also show that this claim does not entail a denial of ontological individualism. This means that collectives to which obligations are owed need not be anything more ontologically speaking than the sum of individuals who constitute them.

These arguments are meant to stand on their own. However, I would also like to show that they are further buttressed by at least two additional burden-of-proof considerations. The first such consideration has to do with what I call the *obligation reciprocity thesis:* if *X* is the kind of thing that can incur obligations, then we can also incur obligations towards *X*. Those resisting the claim that collectives can be obligees would either have to accept that collectives cannot have obligations if they wanted to hold onto the obligation reciprocity thesis, or they would have to give up or significantly limit the scope of the obligation reciprocity thesis. As I will show, there is a theoretical "price" to be paid for either of these options. Second, those who deny that we can have obligations to collectives will have to explain why the moral status of collectives is on a par with other members of the moral community in certain normative domains, but not in others. For example, collectivists about moral responsibility—who think that collectives can be fit to be held responsible in basically the same sense as individuals—need to explain why collectives can be morally responsible but not be "obligees".

Affirmative Action as a (*Pro Tanto*) Obligation

To illustrate these general points, throughout the paper I will be using the example of affirmative action, and specifically, that of affirmative action through the use of quotas.

As regards this example, I will make three points, which I take to be relatively intuitive, and for which I will not argue any further here. The first point is that the underrepresentation of some minority group in a larger group (at a company or university, for example) could be unjust and unfair for a variety of reasons. A meritocratic reason would be that when members of the minority group (say, that of women) are equally qualified, it is not fair that they are significantly outnumbered by the majority group (say, that of men).[2]

The second point is that quotas are at least in some cases effective *and* morally justifiable instruments of remedying numerical underrepresentation of minority groups. While the use of quotas is not morally or practically unproblematic or straightforward, there is strong evidence that purely merit-based, group-neutral (e.g., gender-neutral or race-blind) hiring policies often fail to equalize or even significantly raise the proportion of the minority group (Lippert-Rasmussen 2013; Lippert-Rasmussen 2018; Barabás & Szigeti 2022).

The third point is that if underrepresentation is indeed unjust, then the injustice generates an obligation[3] to do something about the injustice of underrepresentation, provided one is capable of doing so. The combination of the first and second points yields the claim that specifically *the use of quotas* can constitute an obligation. We may be duty-bound to resort to quotas because underrepresentation can be unfair and morally unjust (from the first point), and there may not be available comparably effective ways of remedying the situation (from the second point).

It is worth emphasizing that the obligation to use quotas in such cases is meant merely to illustrate the general claim that obligations towards certain groups may be irreducible and non-distributive. That is, the feasibility of the general claim does not hinge on this specific example. Below, I will offer alternative examples for the benefit of those opposed to affirmative action or the use of quotas.

[2] There can be non-meritocratic reasons as well for the injustice of underrepresentation. For instance, if women make up half of the population of a certain country, then there is a good case to be made that the political system does not adequately represent women as long as only a small proportion of that country's members of parliament are women.

[3] To be precise, this is only a *pro tanto* obligation because preferential hiring can lead to the unfair treatment of individuals who are equally or more disadvantaged than members of the minority group, such as the proverbial gifted and hard-working son of an unemployed miner from a poverty-stricken region (Nunn 1974; Goldman 1975; Goldman 2015). The unfairness to such individuals can be relevant to the all-things-considered justifiability of the use of quotas.

The Irreducibility of Certain Obligations Towards Collectives

For now, let us proceed on the assumption that the use of quotas in certain hiring cases can constitute a moral obligation because it is a necessary means to remedying the injustice of unfair underrepresentation. The question then is: what kind of obligation is this and to whom is it owed?

Observe that the obligation in question can be discharged in a wide variety of configurations. In order to achieve the desired equitable representation, we do not have to recruit specific individuals, but rather a specific number of individuals. If, say, five members of the minority group A have to be recruited to make sure that they are no longer underrepresented, then this goal can be equally well realized, for example, by recruiting $a_1 \ldots a_5 \in A$ or by recruiting $a_6 \ldots a_{10} \in A$.

That is to say, in any given case, the obligation towards the group is discharged by treating certain individuals in certain ways (i.e., in this case by recruiting them). However, the obligation is not held towards specific individuals. It can equally well be discharged by treating other individuals belonging to that group in the required way. This "multiple realizability" of the relevant obligation is, I submit, crucial in establishing that we are indeed dealing with an irreducible and non-distributive collective obligation here.

In what precise sense is the obligation towards the group irreducible? It can be helpful here to distinguish between two senses of reducibility: *logical reduction* and *ontological reduction*.[4] The first type of reduction depends on whether statements about collectives are logically equivalent to conjunctions of statements about individuals. The second type of reduction depends on whether collective entities are equivalent without remainder to sets of individual entities.

The account presented here is consistent with accepting reducibility of the second type for collectives as obligees. This means that for the argument to go through one does not have to violate the strictures of ontological individualism. Groups to which obligations are owed need not be anything more ontologically speaking than the sum of individuals constituting the group. Whenever an obligation is discharged—whether it is an obligation towards a collective or an individual—it will necessarily be discharged through actions undertaken towards one or more separate individuals. In the affirmative action case discussed above, this condition is clearly met as the obligation towards the collective is discharged by treating preferentially one or more individuals.

So, the claim about irreducibility concerns the first, *logical* type of reduction mentioned above. Such irreducibility obtains because the relevant obligation can be discharged in various ways by treating this or that subset of individual members of the minority group preferentially. True, when certain individual members of the

[4] These terms are adopted from Tamminga & Hindriks (2019).

minority group are treated preferentially, then the obligation towards the group is discharged by treating those individuals in certain ways. However, the relevant obligation is still held towards the group. So, when an individual is treated in certain ways, this is because there is an obligation towards the group which the individual in question is a member of.

The Non-Distributivity of Certain Obligations Towards Collectives

One may object, however, that we cannot have an obligation towards a group unless we have corresponding obligations towards members of the group. If, for example, the obligation is to recruit five members of the minority group *A*, and we discharge this obligation by recruiting $a_1 ... a_5 \in A$, then we have (at least) an obligation to a_1, a_2, a_3, a_4, $a_5 \in A$ (and perhaps to other individuals as well, members of A, and perhaps even non-members). I now want to show that this objection fails. The obligation towards the collective is not distributive. That is to say, the obligation to the collective is not the sum of obligations towards individual members of the collective. To show this, I will now briefly review the main theories on the direction of obligations in order to argue that whichever theory one aligns oneself with the obligation in question should not be seen as directed to individual members of the minority group.

Not all obligations are directed towards someone. For example, the obligation not to harm the environment may not be owed to anyone in particular. However, many obligations are directed. Consider a simple case: if you promised to me to do *X*, then you owe it to me to do *X*. Now, what does it mean for your obligation to be directed to me in this way? It means that *I* would be wronged if the duty were not discharged. For example, even if what you promised to me (i.e., the content of *X*) was to visit a friend of ours, it is I (not the friend) who is wronged if you break the promise.[5] There are different theories to explain the directionality of obligations, i.e., why *I* would be wronged (or wronged in a special way) rather than others in such cases. Depending on which theory of the directedness of obligations one accepts, the explanation could be (i) that only I have the standing to release you from the obligation (as well as impose or waive secondary duties of compensation or enforcement), or (ii) that I have special standing to demand that you fulfill your

[5] Though of course the friend may also be wronged for various other reasons in connection with the breaking of the promise. What these reasons may be will depend in part on why the promise was made in the first place.

obligation, or (iii) that I have a special interest in your fulfilling the obligation (May 2015).[6]

Now, plainly, the obligation to treat the minority group preferentially using a quota is a directed obligation. It seems to me that the obligation in question is very different from standard examples of non-directed obligations such as our duty not to destroy great works of art or respect the value of nature (May 2015). For reasons mentioned above, it is obviously not the case that *nobody* is wronged if the obligation of preferential treatment is not discharged. Nor is it the case, I think, that *everybody* (i.e., each member of the moral community) is harmed to an equal extent if that obligation is not discharged. In particular, it would be preposterous for members of the unjustifiably favored majority, who would stand to benefit if the obligation in question was not discharged, to claim that they were wronged to the same extent as members of the relevant minority group.

However, I also want to argue that no matter which of the above three criteria (i)-(iii) for the direction of obligations we consider, the obligation to treat the minority group preferentially does not appear to be an obligation directed at any given individual member (or any given subset of members). First, no individual member of the minority group can exercise normative control over the obligation of preferential treatment. This means that no individual member of A (or a subgroup, say, $a_1 \ldots a_5 \in A$) has the authority to waive the obligation that a certain subgroup of A, say, $a_1 \ldots a_5 \in A$, be treated preferentially. Nor does an individual member of A have standing to impose or waive secondary duties of compensation or enforcement if a certain subgroup of A (say, $a_1 \ldots a_5 \in A$) rather than another (say, $a_6 \ldots a_{10} \in A$) ends up being treated preferentially.[7]

[6] All of these theories have been contested (see May 2015) and some have proposed hybrid theories in response to criticisms. I will ignore these debates and theoretical developments here and will try to show that the obligation to the collective remains non-distributive and irreducible whichever theory we adopt. I think I am entitled to doing so as the account is based on the "multiple realizability" of the obligation to redress the injustice of underrepresentation rather than on some prior theoretical commitment about the direction of obligations. I should add, however, that at a pinch I would rather give up the neutrality of my account with regard to theories of direction than give up the account itself. In particular, for reasons to be explained below, the conclusion that we can have non-distributive and irreducible obligations towards collectives may well be seen as putting additional pressure on the demand theory of directionality (which is subject to quite severe objections anyway).

[7] There is a slight complication here. In some cases, it may indeed be permissible for an individual member of A to decide that s/he does not want to be treated preferentially. However, I want to note that this right is much more limited than it may appear at first sight. For example, if the inclusion of that individual member is necessary for redressing the injustice of underrepresentation by using quotas (because otherwise there would not be enough minority applicants to fill in the quota), then it is arguable that the individual member's right is overridden by the need to redress the injustice. In addition, there is a good case to be made that by applying for a position at an institution that uses affirmative action in hiring, individual applicants forfeit the said right. In any case, even if an individual member has the right to refuse to be treated preferentially, I do not think this shows that the obligation to be treated preferentially was owed to her as an obligation to an individual.

Second, an individual member of A, say a_6 has no standing to insist that any given subset be treated preferentially, e.g., $a_6 \ldots a_{10} \in A$ rather than $a_1 \ldots a_5 \in A$. This also means, *a fortiori,* that no individual member of A has standing to insist that s/he be treated preferentially either.

And finally, third, the obligation of preferential treatment might not end up serving the personal interest of any given member of A (or the personal interests of members of any given subset, say, $a_6 \ldots a_{10} \in A$). If, for example, $a_1 \ldots a_5 \in A$ are treated preferentially, then a_6 (or $a_6 \ldots a_{10} \in A$) will not benefit personally from the preferential treatment.

To be sure, individual members of A do have the right (and in some cases perhaps even duty) to insist that the obligation of preferential treatment towards A be discharged. It may also be allowed (although this point is far from obvious) that members of A differ in this respect from ordinary members of the moral community meaning that they have better or stronger standing than other members of the moral community to advocate for A, and specifically, to call out the injustice of A's underrepresentation and insist that the obligation to redress it be discharged. However, even if it is granted that they have this right or special standing, A's members having this special right or standing only shows that the obligation of preferential treatment is owed to minority group A as a whole. It does not show that the obligation of preferential treatment is owed to members of A individually. At best, that right or standing only entitles members of A to speak up on A's behalf, and to do specifically by insisting that the obligation towards some unspecified subset of the minority group A be discharged.[8] As we have seen, no individual member has the right or standing to insist that the obligation be discharged towards to herself/himself nor to any specific subset of A. By the same token, no member or subset of A is wronged, if the obligation of preferential treatment is properly discharged towards another (suitably large) subset of A.

However, would not the concession I have just made, namely that members of A differ from ordinary members of the moral community because they have better or stronger standing than other members of the moral community to call out the injustice of A's underrepresentation, mean that the obligation of preferential treatment is owed individually to *all* members of A (even if only a given subset of them can be actual beneficiaries of preferential treatment)? I do not think so. The mere fact that one has a special standing or right to call out an injustice or insist on the discharging of an obligation to redress that injustice does not establish that that obligation is owed to one.

This is because even if the demand condition (see condition (ii) above) is held to be crucial for the directedness of obligations, surely, the demand must be rather

[8] There are stronger reasons to talk of such entitlement to advocate for A if A starts to organize itself and, in the process, empowers certain individuals to speak on *A*'s behalf (the final section briefly discusses such scenarios).

specific and concrete.[9] In the above example about your promise to me to visit a friend of ours, what establishes the directedness of your obligation to me (if indeed one wants to rely on the demand theory of directedness) is that I can demand that you do what you promised, namely visit that friend because and in the way I asked you. However, it seems that individual members of *A* cannot make such specific demands with regard to the obligation to redress the injustice of under-representation. They can advocate for *A* and demand the discharging of the obligation to redress the injustice of underrepresentation in general terms and with regard to *A* as a whole, but (as we have seen) they cannot make demands as to which individual members of *A* should be targeted. In short, even if we accept that members of *A* have special standing or right to insist on compliance with the obligation (and I am not sure we should accept this in the first place), this does not establish that that obligation is owed to all (and each) of them, only that it is owed to *A* collectively.

In sum, what this means is that the obligation to the minority group—specifically, the obligation to redress the injustice of its underrepresentation—cannot be cashed out as a concatenation of statements about individual obligations owed to specific members of the relevant group. The obligation in question is not distributive because it is not directed towards specific individuals. Rather, it merely specifies that a given *number* of unspecified individuals within the minority group are to be treated in a certain way.

One final worry is that this argument hinges on the moral acceptability of the use of quotas in affirmative action. This is not the case. I have chosen this—admittedly somewhat controversial—example for detailed analysis to show that the problem of obligations towards collectives does bear on issues of even considerable normative importance and contemporary interest. However, the argument can be illustrated using simpler, less contentious cases as well. For example, imagine that a promise has been made to a group that it will be proportionally represented in some larger body. If so, then there will be an obligation to the group to which this promise was made that is not reducible to obligations to any individual members of this group, and this obligation to the collective cannot be distributed as obligations to individual members.[10]

[9] This is perhaps also the right place to confess that I have serious doubts regarding the plausibility of the demand condition as an explanation of the directedness of obligations in general. *C*'s standing to demand compliance with an obligation seems to be neither sufficient nor necessary for the obligation to be directed towards *C*. May (2015) raises weighty objections against this condition along these lines.

[10] One might also consider cases in which what is owed to the group is not divisible and distributable in some obvious piecemeal fashion (as in the hiring case where the jobs ultimately go to specific individuals). One interesting real-life example would be the obligation to return plundered works of art to the country of their origin. I thank Michael Cholbi for suggesting this example.

Obligations of Collectives and Obligations Towards Collectives

An interesting parallel can be drawn here with the literature on obligations held by (rather than owed to) collectives. Many, perhaps most, who have written about obligations of collectives take the view that at least in some cases the obligations of collectives can be irreducible to the obligations of their members (Copp 2007; Isaacs 2011; Lawford-Smith 2012; Wringe 2016; Collins 2017; Tamminga & Hindriks 2019; Blomberg & Petersson 2022). Moreover, it is also commonly accepted that such obligations are not distributive, whereby non-distributivity is used in the same sense as above, i.e., it is argued that some obligations incurred by collectives cannot be cashed out as concatenations of obligations held by individual members.[11]

The parallels between discussions of obligations of collectives and my suggested analysis of obligations towards collectives is mentioned here not just because of the encouraging similarity in the structure of arguments by which collectivist conclusions are reached with regard to obligations held by/towards certain groups. The reference to that parallel debate is also important because it raises the question whether we have any good reasons to accept a disparity in the status of collectives, so that they can be holders of obligations, but not addressees of obligations.

In general, the obligation reciprocity thesis mentioned in the introduction seems to have at least some initial appeal: if *X* can have obligations, then we can also have obligations towards *X*.[12] It seems unfair that we can impose binding demands on some entity (agent or social object) without that entity having the right to expect any binding commitments from us—not just in a concrete situation, but categorically due to the kind of entity it is. Whatever the case may be, I cannot argue for the obligation reciprocity thesis here. What I want to point out, however, is that denying the main claim of this paper would entail rejecting either the obligation reciprocity thesis or the view that collectives can have obligations. Given the initial plausibility of both the obligation reciprocity thesis and the arguments for the existence of obligations held by collectives, the burden of proof clearly shifts to those who would deny that we can have obligations towards collectives.

This is a specific challenge to be faced by those who accept that collectives can have obligations, but not that they can be obligees. However, there is also a broader burden-of-proof issue here about the moral status of collectives. The question again

[11] Collins (2017) writes that "group agents' duties are distinct from a collection of individual duties", and Tamminga & Hindricks (2019) that "fulfilling an individual obligation is neither necessary nor sufficient for fulfilling a member obligation". By the same token, Lawford-Smith (2012) argues that each member of a collective can satisfy her individual and membership obligations and the group can nevertheless fail to satisfy its collective obligation, and Isaacs (2011) seems to agree.

[12] The converse thesis is more debatable: that we can have obligations towards *X* does not seem to entail that *X* can have obligations. For example, *X* may be dead, or *X* could be an animal.

concerns parity: in this case not specifically in terms of obligations, but rather in terms of other properties relevant to a collective's moral status. There is a growing body of literature on various aspects of this problem. Philosophers have been asking: Can some collectives be morally responsible? Can they have rights? Can they perhaps even feel emotions, suffer pain and experience pleasure? Of course, few would deny that the answer to at least some of these questions is yes, provided the ascriptions of responsibility, rights, sensations, etc. to collectives is understood in a distributive sense—merely picking out properties of some or all individual members of the collective. However, just as in our case of obligations, there is a further question whether such ascriptions can be used *non-distributively* as well.

It is of course not for us to answer these questions here. However, once again it seems fair to diagnose that the burden of proof will be shifted to those who are prepared to ascribe some determinants of moral status to collectives, but not others. Why should we say, for example, that both individuals and groups can be morally responsible in roughly the same sense, but not that they can be obligees in roughly the same sense? In general, why should we assume parity between individuals and collectives with regard to one aspect of moral status (e.g., that of moral responsibility), but not with regard to obligations? These questions will be all the more pressing given that these aspects of moral status are in many cases not independent from one another: the possession of one relevant property could entail the possession of another: for example, on Strawsonian accounts of moral responsibility, the ability to feel reactive emotions is necessary for being a morally responsible agent.[13]

What Kind of Collectives Can Be Obligees?

One question I did not discuss in this paper is whether groups to which obligations can be owed need to have a certain kind of structure and/or possess distinctive organizational features. It has been argued, for example, that only formal organizations with a well-defined division of labor, a stable collective identity, and a formal decision procedure (e.g., corporations), but not random collectives can be morally responsible in a non-distributive sense (List & Pettit 2011). The arguments above do not stipulate such conditions for collectives to qualify as obligees. The target groups of affirmative action measures need not be formally organized groups.[14]

[13] Emphatically, the considerations adduced in this section are merely supposed to highlight the costs of rejecting the view defended and are not meant to serve as knock-down arguments. Indeed, there have been serious attempts to justify the differential treatment of individuals and collectives as regards various determinants of moral status. For example, disparities have been traced back to the fact that individual human beings possess phenomenal consciousness, while collectives do not (List 2018). The arguments above do not aim to discredit such attempts, they are merely meant to show that such arguments are required.

[14] Incidentally, it does not seem to be the case either that such structural requirements would apply to collectives that can have obligations (beyond perhaps some ability of members to communicate with

I do not wish to take a stance here as regards the question which characteristics may be indispensable for collectives to qualify as obligees. However, since the examples I have been focusing on all involve unstructured groups, it is necessary to consider the objection that such groups do not qualify for the specific reason that, given their lack of structure, these groups do not have the capacities required for duties to be directed at them. That is, since they are unorganized, they might not be able to waive an obligation or insist that the obligation be discharged (or complain if it has not been discharged).

In response to this objection, it is worth pointing out four things. First, the interest theory of directed obligations could still work for unorganized groups even if the other two theories mentioned above do not. Second, there are other cases in which it seems quite clear that we have obligations towards various individuals who lack said capacities, e.g., patients in a coma or my yet unborn child. Third, the objection is helpful because it brings out why it may be especially important for certain minority groups to organize themselves and to find a voice, and the account proposed here can in part explain why many of them have done so (Lackey 2018; Townsend 2020). Even if it is not true that such capacities are strictly speaking required to qualify as an obligee, it is clear that only those in possession of such capacities can advocate effectively for the discharging of obligations owed to them. And fourth, even if one nevertheless believes that such capacities are strictly required to qualify as an obligee and one also believes that unorganized collectives do not have these capacities, there can still be good reasons to treat such unorganized collectives as "embryonic" obligees who have the potential to develop into fully enfranchised obligees by adopting the requisite structure or organizational features (among others, by appointing spokespeople on their behalf or by adopting formalized decision-making procedures).[15]

After having surveyed some of the extra theoretical "costs" to be faced by opponents, I close by calling attention to an additional attractive feature of the position defended here, namely that it comes, metaphysically speaking, "cheap". Specifically, the conclusion that we can have obligations towards collectives is not made on the basis of claims about what kind of entities can be persons or agents. Other things being equal, this should make these arguments more appealing than arguments for obligations towards collectives which are based on contentious ideas about group personhood or group agency. By the same token, it should also make these arguments somewhat more difficult to resist by individualists who are opposed to the idea of obligations to collectives *because* they are opposed to collectivism about personhood or agency.

---

one another). Random people at a beach may incur the collective duty of rescuing a drowning child without being members of a formally organized group (see Blomberg & Petersson 2022).

[15] This last argument about "embryonic" obligees parallels (but does not presuppose) Pettit's (2007) arguments about the "developmental rationale" for treating certain unstructured groups as if they were (already) corporate agents.

References

Barabás, György & András Szigeti 2022 "Using quotas as a remedy for structural injustice". *Erkenntnis,* online first: 1–19.

Blomberg, Olle & Björn Petersson 2023 "Team reasoning and Collective Moral Obligation". *Social Theory and Practice,* online first.

Collins, Stephanie 2017 "Duties of group agents and group members". *Journal of Social Philosophy, 48*(1): 38–57.

Copp, David 2007 "The collective moral autonomy thesis". *Journal of Social Philosophy, 38*(3): 369–388.

Held, Virginia 1970 "Can a random collection of individuals be morally responsible?" *Journal of Philosophy, 67*(14): 471–481.

Goldman, Alan H. 1975 "Reparations to individuals or groups?" *Analysis, 35*(5): 168–170.

Goldman, Alan H. 2015 *Justice and reverse discrimination.* Princeton, NJ: Princeton University Press.

Isaacs, Tracy 2011 *Moral responsibility in collective contexts.* New York: Oxford University Press.

Lackey, Jennifer 2018 "Group assertion". *Erkenntnis, 83*(1): 21–42.

Lawford-Smith, Holly 2012 "The feasibility of collectives' actions". *Australasian Journal of Philosophy, 90*(3): 453–467.

Lippert-Rasmussen, Kasper 2013 *Born free and equal?: A philosophical inquiry into the nature of discrimination*. Oxford: Oxford University Press.

Lippert-Rasmussen, Kasper 2018 "The ethics of anti-discrimination policies" in A. Lever, & A. Poama (Eds.) *Routledge handbook of ethics and public policy* (267–280). London & New York: Routledge.

List, Christian 2018 "What is it like to be a group agent?" *Noûs*, *52*(2): 295–319.

List, Christian & Philip Pettit 2011 *Group agency: The possibility, design, and status of corporate agents.* Oxford: Oxford University Press.

May, Simon C. 2015 "Directed duties". *Philosophy Compass, 10*(8): 523–532.

Nunn, William A. 1974 "Reverse discrimination". *Analysis, 34*(5): 151–154.

Pettit, Philip 2007 "Responsibility incorporated". *Ethics*, *117*(2): 171-201.

Tamminga, Allard & Frank Hindriks 2019 "The irreducibility of collective obligations". *Philosophical Studies, 177*(4): 1–25.

Townsend, Leo 2020 "Group assertion and group silencing". *Language & Communication, 70*(1): 28–37.

Wringe, Bill 2016 "Collective obligations: their existence, their explanatory power, and their supervenience on the obligations of individuals". *European Journal of Philosophy, 24*(2): 472–497.

Causal Involvement, Collectives, and Blame

Replies to Petersson

Matthew Talbert

Abstract. This paper argues that there is reason to distinguish between moral responsibility and moral blameworthiness and, in particular, that we can acknowledge that a person is responsible for the negative outcomes of their behavior without this informing our judgments about the person's blameworthiness. This theme is elaborated in the context of a discussion of some of Björn Petersson's work on collective responsibility.

1. Introduction

For some time, I have been interested in how to think about responsibility and blameworthiness for the causal outcomes of behavior. In this paper, I consider how my own thinking on this topic interacts with the views that Björn Petersson has defended.[1] I begin with a sketch of my own (rather inchoate) thoughts on the topic, then I turn to consider aspects of Petersson's work on collective responsibility and their connection with the issues in which I am interested.

[1] I am delighted to contribute to this collection honoring the careers of Dan Egonsson, Björn Petersson, and Toni Rønnow-Rasmussen. Though my paper deals only with Björn's work, I would like to express my gratitude to Toni and Dan, along with Björn, for the hospitality and kindness that they showed me during my years at Lund University.

2. Outcomes, Responsibility, and Moral Blameworthiness

It is generally assumed by those who write about moral responsibility that we can be morally responsible not only for our actions but also for the consequences of our actions. Moral responsibility theorists also tend to assume a close fit—on the negative side of the moral ledger—between moral responsibility and moral blameworthiness. So, if you are morally responsible for some harmful consequence of a wrongful action, then you are also blameworthy on account of the harm that you caused. And, for those who are comfortable with talk of degrees of moral responsibility, if a person bears comparatively greater responsibility for some outcome, then she is, to that extent, more blameworthy on account of that outcome.

Contrary to the above perspective, I am going to argue that moral responsibility and moral blameworthiness come apart in certain ways and specifically with respect to the consequences of actions. I'll sketch the general lines of my thinking in this section and fill in some additional details below in my responses to Petersson.

In my view, judgments about blameworthiness are properly independent of causal judgments. This is because: (i) judgments about blameworthiness are fundamentally judgments about the fittingness of the morally offended responses involved in moral blame; and (ii) the obtaining of a causal relation, or the occurrence of an unwelcome causal outcome, is never in itself morally offensive in a way that makes the responses involved in blame appropriate. It is only when an outcome is caused in a certain way—e.g., by independently morally objectionable intentions and motives—that moral offensiveness arises. Indeed, I am inclined to think that whatever genuine moral offensiveness arises in a given context is entirely accounted for by the moral quality of such things as an agent's intentions and motives, and not at all by the fact that these things happen to bear a causal relation to an unwelcome result.[2]

However, judgments about moral responsibility may well be dependent on causal judgments. Plausibly, a person is morally responsible for something, such as a causal outcome of her behavior, if she fulfills certain control conditions and epistemic conditions with respect to that outcome. And, particularly when it comes to outcomes, control is at least partly a causal or explanatory notion: an agent's having the relevant control over an outcome is partly a function of that agent being causally or explanatorily connected to the outcome. Of course, causing an outcome is not enough to be morally responsible for it, so the control condition is supplemented by, or it includes, an epistemic condition: one must have known or suspected—or one should have known or suspected—that a certain consequence was a likely result of one's action.

[2] There's an obvious connection between what I say in this paper and the usual things that people say when they argue against resultant moral luck. For arguments related to those in this paper, but with more discussion of moral luck, see Enoch & Marmor (2007), Graham (2017), Khoury (2018), and Zimmerman (2002). For my own views on moral luck, see Talbert (2019).

I conclude from the above that, to the degree that judgments about moral responsibility are dependent on causal judgments, and judgments about blameworthiness are not so dependent, judgments about responsibility and blameworthiness can come apart in ways that are not standardly recognized. It seems to me, in particular, that the fact that a person is morally responsible for an outcome may not tell us anything about the degree to which she is open to moral blame.

Instead of separating blameworthiness from moral responsibility, why not simply say that we are not morally responsible for the consequences of our actions, as Andrew Khoury (2018) suggests? I think this is too revisionary a use of "moral responsibility." Thus, I grant that we can be morally responsible for the consequences of our actions insofar as we can intentionally or knowingly (or negligently or recklessly) bear the right causal or explanatory relation to these consequences. The responsibility here goes beyond mere causal responsibility: in virtue of satisfying relevant causal and epistemic conditions, an outcome can be attributed to an agent as an exercise of her powers of agency, and the outcome may be said to be the agent's doing in a way that seems conceptually distinct from an attribution of causal responsibility. But as I've said, I think that moral responsibility for consequences need not affect blameworthiness—it need not affect the aptness of directing towards an agent the negative emotional responses involved in moral blame.

I admit, however, that it also seems an undue revision to normal speech to deny that people can be "to blame" for outcomes. I suggest that to say that someone is to blame for an outcome means no more than that that person is morally responsible for an unwelcome outcome, and that she is responsible for this outcome in virtue of some morally objectionable feature of her self that bears the right causal/explanatory relation to the outcome. These morally objectionable features might be such things as the agent's morally bad motives and intentions, her bad desires and patterns of concern, and so on. Such things are not required for moral responsibility per se—one can be morally responsible for an outcome in virtue of praiseworthy motivations—but if her motives, patterns of concern, etc. hadn't been morally deficient, then she wouldn't be *to blame* for that outcome.

An agent who is to blame for an outcome is blameworthy in virtue of the factors (e.g., her bad intentions) that make it correct to say that she is *to blame* for the outcome. But these morally objectionable features of the agent *on their own* (and even if they hadn't led to a bad outcome) are enough for her blameworthiness. So, an agent is open to blame, and so blameworthy, regardless of whether her objectionable intentions and motives are causally connected to an unwelcome outcome for which the agent is morally responsible and to blame.

Moreover, the fact that a person is morally responsible (and to blame) for an unwelcome outcome does not make her *more* blameworthy than she already is just in virtue of possessing the objectionable features that, together with the obtaining of certain causal connections, make her to blame for the outcome. This is because,

as I suggested above, the obtaining of the causal connections necessary for establishing the *morally responsible for* and *to blame for* relations is not in itself morally offensive. The moral offense is already there in the objectionable features of the agent that may or may not give rise to the unwelcome outcome. Since the actual occurrence of the outcome, and the obtaining of the *to blame for* relation between agent and outcome, does not affect the moral quality of the objectionable features of the agent, it does not amplify blameworthiness (though see Lang 2021, Chapter 2, for the contrary view). Of course, as others have noted (e.g., Enoch & Marmor, 2007; Lawson, 2013), the occurrence of a bad outcome may make an actor's objectionable motives (etc.) more conspicuous than they would otherwise be, but this explains only our readiness to blame in the face of such outcomes, and not the blameworthiness of those whom we blame.

3. Replies to Petersson

I turn now to consider some of Björn Petersson's reflections on collective agency and responsibility for outcomes and how the view outlined in the previous section might interact with Petersson's. I begin with Petersson's 2008 paper, "Collective Omissions and Responsibility" (which builds on Petersson, 2007). The title makes Petersson's focus on omissions clear, but I take my comments to apply to both actions and omissions.

Petersson is a realist about collective responsibility: "groups as such can be morally responsible for effects of their acts, in a sense that cannot be reduced to judgments about individual members' acts" (2008, 244). Collective responsibility, then, is not merely an aggregation of, nor does it collapse into, instances of individual responsibility. Moreover, for Petersson, while "[y]ou are individually responsible … for your intentional marginal contribution to some harm" your "individual responsibility need not coincide with your co-responsibility for the overall harm produced by the collectively responsible group" (2008, 251). This last point is easiest to see in cases in which the collectively caused "harm is over-determined, so that no individual member's act makes any difference to the occurrence of the event in question" (2008, 251).

As far as moral responsibility goes, I see no reason to disagree with what Petersson says. I may be part of a group, and it may be true that the group has caused some harm and that, as an appropriately situated member of this group, I share in its responsibility for doing so. As a member of the group, the harm it caused is partly attributable to me, but my responsibility need not be proportional to my difference-making contribution to the harm. For example, in overdetermination cases, my individual action may make no difference to the occurrence of a harm: had I not acted, the overdetermined harm would have been the same. So, we have at least one

way in which collective responsibility does not collapse into individual responsibility.

But I have trouble seeing how related conclusions about blameworthiness might follow. Blameworthiness seems to me individualized in a way that responsibility may not be. Unfortunately, it is difficult to make this case in the context of the current debate about responsibility since, as I noted above, most theorists assume a tight fit between moral responsibility and blameworthiness. For example, Petersson says that he is working with a "thick" sense of moral responsibility: "[t]o hold a collective morally responsible (in the thick sense) for some harm is to imply that moral sanctions are in place" (2008, 251). It is clear from context that the sanctions Petersson has in mind include moral blame. Indeed, once it is clear that we are considering a case of moral wrongdoing, it tends to make little difference—for Petersson and most other responsibility theorists—whether we speak of "moral responsibility" or of "blameworthiness." This easy transition from *responsibility* to *blameworthiness* is one of the things I am arguing against.

Now, it might be thought that Petersson and I are simply using "moral responsibility" differently, but I don't think this recognizes the disagreement between us. I mean to argue that the conditions that Petersson—rightly, in my view—claims are sufficient for group responsibility do not yield straightforward conclusions about blameworthiness in the way that I take Petersson to suppose.

3.1 Loosely Structured Groups and Individualized Blame

The first point I want to make is that sometimes "blaming a collective" can be suitably analyzed in terms of blaming individuals even if a corresponding analysis of moral responsibility would not be appropriate. This observation is most applicable in cases of "loosely structured" collectives (May, 1990), which are Petersson's central focus in "Collective Omissions and Responsibility." Such loosely structured entities "need not have a common decision procedure, let alone be formally constituted as a group" (Petersson, 2008, 246).

Here is an example (from Petersson, but based on one in Held, 1970) of a loosely structured group and a case in which their collective omission seems blameworthy.

> Suppose an injured person was trapped under a girder, and that the joint effort of two people would be needed to lift it. You and I were the only ones who knew about this accident, and we could have helped if we had acted together…. [but] no collective effort took place. (Petersson, 2008, 257)

Suppose the person trapped under the girder dies. In this case, our loosely structured group is plausibly responsible for a death, and we are plausibly open to blame (depending on what explains our failure to cooperate). But what does blame come to in the case of such a loosely structured collective? A more concretely structured group might be clearly targeted *as a collective entity*: a sports team can be disbanded

or made to forfeit a game, a manufacturing company can be fined, and so on. But with loosely structured entities, it is harder to say what blaming a group comes to—or at least it is harder to see how blaming the group comes to anything more than blaming individuals. In addition, the moral significance of blame (its point and aim) may be exhausted once *individuals* have felt blame's sting, have repented, and so on. After individuals have been reached in this way, what is left over, in the loosely structured case, for "blaming the group" to do?

As Petersson notes, "[w]hen we blame someone, we want the wrongdoer to care about our negative attitudes towards him, as well as about our reasons for having this attitude…" (2008, 249). Again, a concretely structured group—the sports team or the commercial enterprise—might issue a corporate apology that registers the group's acceptance of blame.[3] But how does our desire for such a response apply to a loosely structured group? I suspect that our concern will have been met when the individuals composing such a group—or perhaps a specific subset of these individuals—have taken up an appropriately repentant stance. This stance *might* be expressed in a thought that has, so to speak, collective content. Each individual in Petersson's example might think: "*we* acted wrongly, *we* should have done something."[4] But would anything be missing, would our blame not have achieved its aim, if each individual, or at least those who do have something to apologize for, merely thought: "*I* should have done something different"?[5]

Petersson considers a few variations of the case of the person trapped under the girder. First, it might be that you and I are faultlessly unable to communicate and to coordinate our efforts. Second, one or both of us might be a committed non-cooperator, which explains why our efforts were not effectively coordinated.[6] Finally, it might be the case "that we both considered which options the group had as one unit of causal agency, and that we both agreed" on what was needed to save

[3] Of course, we are liable to think that something is left undone by such an apology if we are uncertain whether members—perhaps specific members—of the group feel personal moral guilt.

[4] Olle Blomberg observes in written comments that Petersson "might prefer to say that each has a thought from the 'we-perspective,' where this is distinct from having a thought with 'merely' collective *content*. … [the collective perspective would be] part of the thought's mode rather than its content." Blomberg directs readers to Petersson (2015; 2017)

[5] Perhaps this thought won't be apt in every case. As Mattias Gunnemyr points out in written comments, the case might be one in which I am stuck with a non-cooperator (and that is why we do not act) or the case might be such that if I act alone (and others fail to join in), then I will simply make things worse. In such cases, Gunnemyr suggests that the appropriate thought might be, "*we* should have done something different." I'm inclined to think, though, that the thought might be, "*he* (or *they*) should have done something different."

[6] In comments, Blomberg notes that if both of us are non-cooperators, then something would be missing if each of us thought only, "I should have done something different" since individual action would have changed nothing; so, the relevant thought must be, "*we* should have done something different." See Blomberg and Petersson (2023).

the trapped person, but we could not agree on how to time our effort, and "[w]hile we argued … the victim died" (Petersson, 2008, 258).

Here is how Petersson assesses responsibility and blameworthiness in these cases (note the easy shift between talk of "blame" and of "moral responsibility"):

> Firstly, we would not blame people for not acting together if they were unable to form joint attitudes to begin with. Secondly, if the attitudes of individual members of a group hinder joint efforts, these individuals are the ones to be blamed for the effects of not acting jointly. Finally, in the third type of case, where a harm is caused by a failure to act that can be explained by members' attitudes concerning a group's options regarding that harm, we find it less inappropriate [to] hold the group as such morally responsible. (2008, 259)

In the third case, what does holding the group responsible—blaming the group—come to? I suspect that the blame applied in the third case will be quite similar to the individual blame applied in the second case. That is, we will blame individuals in both cases, and the blame might take a similar form in both cases. And this will be so even if there is a genuine difference in moral responsibility between the two cases: individual responsibility in the second case, and group responsibility in the third.[7]

Petersson is aware, of course, that "collective sanctions inevitably strike individual members" of a group (2008, 251); I am suggesting that, in some cases, there may not be anything to blaming a group over and above individually felt sanctions. And this indicates a sort of disanalogy with collective moral responsibility. One can have the thought about blame that I have described while still supposing that groups cause things and are morally responsible for outcomes in a way that it is not reducible to individual causal contributions. It may be that collective responsibility does not collapse into individual responsibility, but I suspect that collective blame—or the point of blaming a collective—can collapse in this way, at least in cases of loosely structured groups.

3.2 Groups and "Moral Taint"

I turn now to individual blameworthiness and its relationship to group responsibility.

[7] In comments, Gunnemyr suggests an alternative that I admit has some appeal for me, though I don't have space to develop the idea here. The basic thought is that, in the third case, the members of the group are plausibly engaged in (objectionable) "we-mode thinking" as they fruitlessly (and callously) debate what to do. So, perhaps I could say that it is solely these thoughts—undertaken from a collective perspective—and the objectionable (and, in some sense, collective) quality of will that the thoughts manifest, which grounds blame. This preserves my resistance to allowing outcomes to affect blameworthiness but would allow for blameworthiness to be more inherently collective, in the relevant cases, than I have suggested.

Petersson says that "our method for delimiting the collective agent that is morally responsible for a specific harm should be such that it picks individuals that justifiably can be blamed for what the group has done" (2008, 251). Thus, we should establish "a link between the individual and the group's act" to avoid holding responsible "innocent bystanders who may have been causally involved [in the production of harm] through no fault of their own" (2008, 251).

This is correct, as far as it goes, but in turning to collectively caused, overdetermined harm Petersson says:

> In such cases, we may not be able to assign ordinary individual moral responsibility to any member, while we still find the group's behavior reprehensible. To hold you co-responsible, then, is to hold you to account for the group's act in virtue of the features that [make] you a member of the collective agent, regardless of your individual intentional marginal contribution. The direction of the relation between collective responsibility and co-responsibility is supposed to be top-down—members will be morally tainted by the worth of the collective action. (2008, 252)

Again, I find this a plausible thing to say about responsibility. Depending on our views about causation, we may have difficulty assigning individual moral responsibility in overdetermination cases because it may be unclear whether certain causal links hold between an individual's choices and a harmful outcome. Yet it may still make sense to say that the group of which the individual is a member caused the harm and is responsible for it, and so it may make sense to assign moral responsibility to the individual in virtue of his membership in the group (where this membership is partly established by locating the relevant links "between the individual and the group's act" that Petersson mentions).

But what should we say about blameworthiness? Here I am interested in Petersson's reference to moral taint and to the reprehensible nature of the group's behavior. Behavior, I take it, may be unwelcome, hurtful, and even wrong (though not, I think, wrong in a sense relevant for blameworthiness), independently of the motives and aims that explain the behavior. But I do not see that behavior, or the consequences of behavior, taken independently of facts about motives and aims, can count as morally reprehensible in a way that is relevant to concerns about moral taint and moral blameworthiness. So, I do not see how a group's actions can be reprehensible in this way, and in a way that could plausibly reflect on the group's individual members, except that these individuals' motives, aims, and intentions are morally reprehensible. But in this case, moral taint, which I take to be the grounds for a judgment of blameworthiness, is not top-down; it is, rather, bottom-up.[8]

[8] In comments, Blomberg suggests that a set of micro-aggressions that are individually not particularly morally serious might aggregate in such a way that the target of these slights has grounds for significant moral offense directed at the group that has collectively slighted him. I must admit that this is an appealing proposal, and I am not sure that what I say in the text can accommodate it.

Perhaps moral responsibility can reasonably be conceived of as "top-down," if this means that a person's responsibility is not necessarily explained by their individual causal contribution to an outcome but rather may be explained by their membership in a group and by what the group has done. But blameworthiness, as far as I can see, goes the other way: an individual's openness to blame is explained by facts about that individual, facts beyond their membership in a group. And each individual member of a group may show that they are not open to blame by showing that there is nothing reprehensible in their motives and intentions—importantly, this may not be the same thing as the individual showing that they are not morally responsible for a collectively caused outcome. In other words, it's possible that an individual fulfills relevant causal, epistemic, and group membership conditions such that they count as co-responsible for a harmful outcome and yet their individual motives and intentions may show that they are not blameworthy. (Similarly, you can be morally responsible for a harm that you knowingly brought about, and yet you will tend to avoid blame if your intentions and motives are unobjectionable.)

Of course, it may be that, for certain groups with certain aims, there is no way that an individual can be a willing and informed member of that group without this indicating something reprehensible and blame-grounding about that person. But, again, the individual moral taint here will be a function of the fact that the individual willingly, and with relevant knowledge, joined a group of that sort. There is, I think, no top-level description of a group and its aims, no matter how distasteful, that suggests that individual members of the group are morally tainted without this taint being explicable in terms of reprehensible features of the individual agents.

3.3 Causal Influence and Blameworthiness

In "Co-responsibility and Causal Involvement" (2013), Petersson responds to Christopher Kutz's (2000) argument for rejecting causal involvement as a necessary condition on co-responsibility.[9] Kutz's argument against the causal involvement condition focuses on cases of overdetermination, particularly the Allied bombing of Dresden at the end of WWII.

Given the number of aircraft and bomber crews involved in the raid on Dresden, the destruction of the city was overdetermined: "Each of the 8000 crewmen's causal contribution was … 'marginal to the point of insignificance'" (Petersson, 2013, 848; quoting Kutz, 2000, 118). We might wonder, then, how such minor causal contributions can ground ascriptions of co-responsibility for the fact that Dresden was destroyed. How could such small contributions ground the relatively high degree of moral blame that we might think apt in the case of the intentional destruction of a city populated largely by civilians? It's better, Kutz argues, to think of responsibility and blame in such cases as depending not on individual causal

[9] Again, Petersson is "talking about responsibility in a thick sense, in which moral responsibility is essentially connected to the justifiability of blame and other moral sanctions" (2013, 850).

contributions, but rather on the presence of morally objectionable intentions, such as the intention to destroy civilian-populated Dresden.

In response, Petersson notes that even if an action "made a small, negligible or imperceptible difference to the occurrence of a great harm, it made a difference to the occurrence of that harm" (2013, 858). As such, the case at hand does not threaten "the idea that being co-responsible for something requires making a difference to its occurrence" (2013, 858). "At most," such cases "show that *degree* of co-responsibility for a specific event need not correspond to the size of the causal contribution" (2013, 858, emphasis in original).

The suggestion is that one's degree of responsibility—which will, for Petersson, entail conclusions about blameworthiness—might outstrip one's degree of causal contribution. This is a noteworthy feature of Petersson's account for my purposes. For, as Petersson observes, "it might seem odd to insist that causal involvement is an essential … condition for co-responsibility, while admitting that there is not always any straightforward … relation between causal contribution to an event and degree of co-responsibility for that event" (2013, 858). In fact, I think there is something odd here, and I'll try to draw it out.

Petersson's first response to the oddness he mentions is to note that, on the causal-involvement account, even a small contribution to a horrible outcome may warrant significant blame. Even if a person is blamed in proportion to their causal contribution, "[w]hat she should be blamed for would still vary not only with her share of the total event but also with the value of that event. 1/8000th of an atrocity could be an atrocity" (Petersson, 2013, 858).

It is true that a harm that is small according to some scale of measurement may still be a morally significant harm. Still, I think the above suggests an implausible representation of our blaming practices. Blame and the feelings that express it are not generally divided up to neatly correspond to fine-grained judgements about causal contributions. These responses do not, for example, automatically realign to reflect revised judgments about relative causal contributions when we learn that there were a few more contributors to a harm than we had originally supposed.

More helpful is Petersson's elaboration of the thought that "the causal involvement condition does not imply proportionality between blame and causal contribution" (2013, 858). In this context, Petersson notes that a "justification of an assignment of blame and responsibility will have to appeal to a variety of factors *in addition to* the claim that something bad has happened and that the recipient [of these assignments] was involved in it" (2013, 859; emphasis added).

In addition to attending to a participant's causal involvement, we will also be sensitive when assigning blame "to the agent's type of involvement, the agent's mental capacities, beliefs and intentions," and perhaps also to features of "the social context" in which the agent acted (Petersson 2013, 859). If we are thinking of a collective action, like the bombing of Dresden, it will be relevant to assessments of individual blameworthiness that "participation in a collective project signals a certain kind of commitment" to that project: "[s]uch considerations may explain

why we might consider you highly blameworthy even for a very small contribution to a collectively produced effect" (Petersson 2013, 859). As Petersson observes, "the relative weight of … non-causal considerations typically becomes greater in cases of participation [in a collectively caused outcome] than in cases of single individual actions" (2013, 859).

My suspicion is that, at least in collective cases, the weight of these non-causal considerations can be so great as to swamp our interest in individual causal contributions, at least when it comes to blameworthiness. Petersson disagrees. He says that the above "admission [about the importance of non-causal factors] does not undermine the causal involvement condition for co-responsibility" since "what makes the [individual Dresden] bomber co-responsible for the event in the first place is that his act contributed to its occurrence" (2013, 859).

I agree that a causal contribution is required to establish an individual bomber's partial moral responsibility for the destruction of Dresden. But I suggest that it is much less clear why establishing such a causal relation is necessary for a bomber to be *blameworthy*, particularly when we note how large a role noncausal factors play in our judgments about the appropriateness of blame. If, as Petersson says, we can "explain why we might consider you highly blameworthy even for a very small contribution to a collectively produced effect" by referring to your commitment to an objectionable group project (2013, 859), why should we not regard you as worthy of blame on account of that commitment even if you made *no* contribution to the collectively produced effect?

Suppose that a bomber pilot is deeply and objectionably committed to the goal of bringing about the fiery deaths of Dresden civilians, yet he fails to make even a small contribution to this outcome because the bombs that he drops fail to detonate or because the bomb bay doors of his aircraft malfunction. I can see why this might make a difference to the bomber's causal contribution to an outcome and so also to his moral responsibility for that outcome, but I fail to see how this makes the morally offended attitudes involved in blame any less appropriate than they would be had the bomber been successful in achieving his aim.[10] And this is because, had the bomber been successful, his openness to morally offended responses seems to me to already have a secure footing just in virtue of the morally offensive commitments and attitudes that explain the bomber's choices and actions.

[10] One could opt for a sufficiently fine-grained account of "the bombing of Dresden" such that if the ineffective bomber had been absent, then the fine-grained version of the event would not have occurred. On this account, our bomber's presence (and his failed efforts) would play a role in bringing about (the fine-grained version of) "the bombing of Dresden." One can establish a causal link in this way if one is sufficiently motivated and suitably flexible in their account of causation. But I don't see how establishing a causal link of this sort makes it any clearer that our bomber is a candidate for blame—this seems no more clear than it already was in virtue of the bomber's objectionable moral orientation, which guided his objectionable efforts. For Petersson's take on this fine-grained approach, see (2013, 852-856).

Relatedly, and as I suggested in the previous section, if we had a bomber pilot who made a clear contribution to the destruction of Dresden, and yet we somehow became convinced that there was *nothing* independently objectionable in his motives and intentions, then, while we might assign him causal responsibility, and even moral responsibility if he satisfied certain epistemic conditions, we would have no grounds for finding him worthy of blame. In the absence of such things as independently criticizable motives and intentions, even the most significant causal contribution will not be enough for blameworthiness.[11]

Certainly, Petersson and I disagree about much of the foregoing. But occasionally, the disagreement seems to me less stark than it might initially appear. Consider the following passage, which seems to express straightforward disagreement with one of my central claims above:

> If we have evidence of an agent being committed to contributing to an outcome along with others, but it is clear to us that the agent completely fails to contribute to that outcome, I would regard it as absurd to blame the agent *for* that outcome. (Petersson, 2013, 864; italics in original)

Again, this may seem straightforwardly at odds with the account I gave of the unsuccessful Dresden bomber. But note that Petersson speaks here of blame *for an outcome*. I agree that if a Dresden pilot made no causal contribution to the fact that Dresden was bombed, then it would be absurd to blame him, even partially, *for that outcome*. But I claim that the unsuccessful bomber is blameworthy—that is, he is open to the responses involved in blame—on account of his bad motives and intentions, and that he is blameworthy to the same extent that he would be had he been partly responsible for (and so, partly *to blame for*) the fact that Dresden was bombed.

Petersson seems willing to meet me at least part of the way here. Consider the following example of his, inspired by a television comedy:

> the thoughtful mother of a blind young neo-Nazi regularly swaps the son's swastika-badges … for completely innocent symbols, without his knowledge. Suppose this son joins a neo-Nazi-demonstration and that the demonstration to some minor extent is successful in creating conflict and violence. At the same time, unintentionally the blind son radically diminishes this harmful effect of the collective behaviour, just by being visible in the crowd with cute symbols on his clothes. (2013, 864)

Petersson says of the son:

[11] As I noted in the previous section, it may be difficult to see how one could fulfill relevant epistemic conditions, and engage in certain courses of conduct, without this evincing an independently criticizable moral orientation, but still, the attribution of such an orientation seems to be necessary for blameworthiness.

> He fully shares the participatory intentions of his fellows, but he obstructs the fulfillment of those intentions. We should blame him for trying but surely we cannot blame him for the small raise in conflict and violence that is an effect of the demonstration. It turned out that his mother's well-intentioned deception successfully prevented him from making himself responsible for that sort of harm. (2013, 864)

Again, I agree. The son is not morally responsible for the outcome in question because he does not contribute to it, and so we cannot say that he is to blame for the outcome. But, as Petersson allows, the son is open to blame on account of *what he tried to do*, and presumably this has to do with the fact that his trying was explained by his sharing the objectionable motives and intentions of his (more causally effective) fellow neo-Nazis.

And suppose that the son had succeeded in helping to bring about the sort of unwelcome outcome at which he aimed. This would not make him more worthy of (or worthy of more) blame because it would not make his bad motives and bad intentions more morally offensive—more blame grounding—than they already were. Of course, if the son were to have caused some bad outcome, this might draw our attention to his objectionable aims and intentions and give us additional reason to regret their presence in him. In this way, the occurrence of the outcome might serve to explain why we blame without providing further moral grounds for blame.

Conclusion

As I just observed, the occurrence of an unwelcome outcome can draw attention to blame-grounding motives and intentions. However, Petersson says that "[t]he idea that causal links are relevant merely as indicators of intentions gets things the wrong way around" (2013, 864; Petersson directs this comment toward Lawson 2013).

For Petersson, reference to causal involvement is essential to our responsibility practices. When we blame people, we aim, he says,

> to make them react with corresponding feelings of guilt or remorse, not over past states of mind as such, but over what these states of mind have led to…. We want to make them realize that their choices had an impact—that it was no coincidence that bad things happened when they made those choices. Their choices *mattered*, not by themselves, but because of their connections to events that made the world worse. (Petersson 2013, 864; emphasis in original)

There's some tension, I think, between these claims and Petersson's observations about the blind neo-Nazi, which are immediately prior in the text. As Petersson suggested, we blame the neo-Nazi for what he tried to do, and we presumably do so on account of the intentions that moved him to so try, and independently of the fact that his attempt failed.

Still, there's something right in what Petersson says here. It is important that it is "no coincidence" that intentions lead to choices, that choices regularly lead to actions, and that actions regularly have effects in the world. If bad intentions and objectionable strivings never led to consequences that concern us, then I assume that we would not have our habitual concern about people's bad intentions. But because there is a fairly reliable connection between bad intentions and bad outcomes, we do care about people's intentions, and certainly about their attempts, independently of whether their bad intentions lead to bad consequences on a particular occasion. Indeed, and more generally, we expend a great deal of energy thinking about how we stand in other people's estimations, about what they *really* think about us, even if these inner orientations are not revealed in their actions.

The centrality of internal factors, such as intentions, for our responsibility practices is, I think, most prominent when we consider excuses. As I've suggested, no matter how unwelcome the consequences of someone's actions, if we are convinced that the action is explained by morally faultless motives and intentions, then we have no adequate grounds for blame since, by our own lights, there is no moral affront on which our morally offended blaming responses might reasonably be founded. And if an agent's action turns out to be harmless, and yet we become convinced that he had the most objectionable motives and intentions, then it is easy to understand why moral offense is aroused on the part of those who have luckily not been exposed to harm. I concede, again, that we would not have this interest in others' bare intentions and motives if these things did not regularly lead to happy and unhappy outcomes. But given this regular connection, our moral interest in mere intentions (and other internal states of agents) is perfectly intelligible even when these things do not give rise to external outcomes.[12]

References

Blomberg, Olle & Björn Petersson (2023) "Team reasoning and collective moral obligation". *Social theory and practice*. Advance online publication. https://doi.org/10.5840/soctheorpract2023120177

Enoch, David & Andrei Marmor (2007) "The case against moral luck". *Law and philosophy*, 26(4): 405-36.

Graham, Peter (2017) "The epistemic condition on moral blameworthiness: A theoretical epiphenomenon" in P. Robichaud & J. W. Wieland (Eds.) *Responsibility: The epistemic dimension* (163-79). Oxford, Oxford University Press.

[12] I presented some of the ideas in the second section of this paper at seminars at Lund University and Gothenburg University. I'd like to thank all the participants at these seminars for their thoughts. Particular gratitude is owed to Olle Blomberg, Gunnar Björnsson, Mattias Gunnemyr, Björn Petersson, Toni Rønnow-Rasmussen, András Szigeti, and Caroline Torpe Touborg. Blomberg and Gunnemyr are owed additional thanks for their generous and helpful written comments.

Held, Virginia (1970) "Can a random collection of individuals be morally responsible?". *Journal of philosophy*, 67(14): 471-81.

Khoury, Andrew (2018) "The objects of moral responsibility". *Philosophical studies*, 175(6): 1357-81.

Kutz, Christopher (2000) *Complicity: Ethics and law for a collective age*. Cambridge: Cambridge University Press.

Lang, Gerald (2021) *Strokes of luck: A study in moral and political philosophy*. Oxford: Oxford University Press.

Lawson, Brian (2013) "Individual complicity in collective wrongdoing". *Ethical theory and moral practice*, 16(2): 227-43.

May, Larry (1990) "Collective inaction and shared responsibility". *Nous*, 24(2): 269-77.

Petersson, Björn (2007) "Collectivity and circularity". *Journal of philosophy*, 104(3): 138-56.

Petersson, Björn (2008) "Collective omissions and responsibility". *Philosophical papers*, 37(2): 243-61.

Petersson, Björn (2013) "Co-responsibility and causal involvement". Philosophia, 41(3): 847-66.

Petersson, Björn (2015) "Bratman, Searle, and simplicity. A comment on Bratman, *Shared agency: A planning theory of acting together*". *Journal of social ontology*, 1(1): 27-37.

Petersson, Björn (2017) "Team reasoning and collective intentionality". *Review of philosophy and psychology*, 8(2): 199-218.

Talbert, Matthew (2019) "The Attributionist Approach to Moral Luck". *Midwest studies in philosophy*, 43(1): 24-41.

Zimmerman, Michael (2002) "Taking luck seriously". *Journal of philosophy*, 99(11): 553-76.

Emotions as Value Enablers

Fabrice Teroni

For a philosopher interested like I am in issues surrounding our access to values and the role played by emotions therein, Toni's continuing exploration of the FA analysis is a constant source of inspiration. I take the opportunity to express my merited and right-kind-of-reasons-responsive gratitude, Toni.

In this paper, I wish to focus on an intriguing claim that Toni recently put forward in joint work with Wlodek Rabinowicz in reply to a worry raised by the FA analysis of value. Let me start with a probably unnecessary reminder. The FA analysis claims that values are nothing over and above what makes attitudes merited or fitting.[1] The funny, for instance, would come down to what merits amusement; the offensive to what merits anger. There are many distinct families of values, and I shall concentrate on "emotional values", i.e. values for which the FA analysis looks most attractive, as their relation to specific psychological attitudes – emotions – is obvious (the funny, the offensive, the shameful, the sad, the hopeful, the regrettable, etc.).

The worry to which Toni replies is that the FA analysis fosters a revisionary understanding of these values. Pre-theoretically, values are "located in" the objects to which we attribute them: funniness is the property of a joke, shamefulness the property of a deed. By analysing emotional values in terms of merited emotions, the FA analysis would force us to shift from this pre-theoretical conception to a relational understanding of these values. Toni's reply consists in claiming that the worry misconstrues the role of emotions in the FA analysis. According to him, the worry presupposes that the FA analysis conceives of emotions as *sources of value*, while it actually conceives of them as *enablers* for the relevant properties of the object (the joke, the deed, etc.) to constitute values.

I am aware that there is a lot to unpack here, and I shall try my best to do so in what follows. I wish to focus on the distinction between source of value and value

[1] In what follows, I shall mostly deal with the metaphysical aspect of the FA analysis. For a discussion of the role of emotions in value concepts, see Deonna and Teroni (2021).

enabler because I agree with Toni that it is key to the development of an attractive FA analysis. However, its application to the relation between emotions and emotional values raises complex issues. In particular, we should wonder as to how exactly emotions function as value enablers. And how does one's approach to that issue relate to the project at the core of the FA analysis, i.e. that of demystifying value through psychological attitudes?

My aim in what follows is to assess whether advocates of the FA analysis are well-advised to respond to the aforementioned worry in terms of the contrast between sources of value and value enablers. It is not to assess the FA analysis in light of other criticisms, such as the traditional Wrong Kind of Reasons objection (Rabinowicz and Rønnow-Rasmussen 2004) or the more recent explanatory objection put forwards by Francesco Orsi and Andrés Garcia (2021, 2022). The discussion is structured as follows. §1 lays out the worry that the FA analysis fosters a revisionary understanding of emotional values. §2 introduces the distinction between enablers and favourers and how it is pressed into service by Toni to reply to this worry. While I agree that the reply is attractive, since casting emotions in the role of enablers chimes well with how we pre-theoretically understand the relations between emotions and values, I observe that doing so requires that we tackle two connected issues. First, how do emotions function as value enablers? Second, is the resulting picture compatible with the FA analysis? The rest of the discussion is structured around these issues. §3 looks at the role of emotions within the FA analysis so as to specify the kind of enabling role they can play. On this backdrop, I explore in §4 a contrast between how belief relates to truth and how emotions relate to values, a contrast which helps uncover what we are after. A first reaction to this contrast, according to which emotions are value enablers by allowing us to access values, which differ from truth, is examined in §5. I argue that this idea cannot do justice to the key insight of the FA analysis. §6 defends an alternative idea, according to which emotions are enablers in virtue of their attitudinal shapes.

§1 Overcounted Adicity

Pre-theoretically, we conceive of ascriptions of emotional values to objects as being made true or false by the intrinsic properties of these objects. Whether or not a joke is funny or a painting fascinating, say, has to do with their respective intrinsic properties. This is manifest in the fact that, when challenged, we are not tempted to consider anything but the joke or the painting in order to justify our claim that it is funny or fascinating. Of course, we acknowledge the existence of "relational" or "personal" values (Rønnow-Rasmussen 2011): we admit that something may be shameful, threatening or hopeful only for some people (in virtue of their specific physical or psychological constitution, etc.). We conceive of ascriptions of personal

values as being made true by properties of the object and its relation to the relevant people – and nothing else.[2]

Whether we focus on intrinsic or on relational value, the worry is that the FA analysis fixes the adicity of value one unit too high. How so? The FA analysis claims that an object has a value if and only if it merits an attitude. This suggests that it understands what we pre-theoretically conceive as a monadic value in terms of a relation between the object and the attitude that it merits. Here is the worry expressed by Jonathan Dancy, who targets an FA analysis in terms of reasons: "The reason for supposing that goodness is less polyadic than reasons is that reasons belong to, are for individuals." (Dancy 2000: 170) Whether it appeals to reasons, merit or fittingness, the worry is that the FA analysis makes the presence of goodness and of specific thick values depend on certain entities in a way that does not correspond to our pre-theoretical understanding of them (Stratton-Lake 2013: 91). What we conceive as a monadic property (adicity 1) comes out as a two-term relation (adicity 2) between that object and an attitude. According to the FA analysis, funniness would be a relation between the properties of a remark and the amusement that it merits. Similarly, what we conceive as a personal value – and so, as a relation between an object and a subject (adicity 2) – comes out as a three-term relation (adicity 3) between an object, a subject and an attitude. A threat would be a relation between a nearby wolf, say, Toni and the fear that the situation merits.

The worry is substantial. While a philosophical account of an entity may somewhat deviate from the folk understanding of it, we are ready to tolerate more significant deviations for some entities (natural kinds, say) than for others (mental states, perhaps). Values seem to belong to the second category.[3] If the "overcounted adicity" objection is along the right track, the FA analysis indeed fosters a revisionary understanding of value. In addition, a commonly advertised selling point of the FA analysis is its neutrality vis-à-vis many substantial issues in the philosophy of value (Orsi 2015, Rønnow-Rasmussen 2022). This alleged neutrality is undermined if the FA analysis rules out the very intelligibility of monadic value.

A natural reaction to the foregoing proceeds as follows. Perhaps the objection under discussion threatens the FA analysis of the thin values of the good and the bad, as well as of some thick values (justice and lewdness, say), but why on earth would it threaten the FA analysis of the emotional values on which we have decided to concentrate? One may after all insist that an FA analysis of these values cannot

[2] Some (emotional) values are neither intrinsic nor relational. It would for instance be regrettable if an employee of the Massaschusetts Historical Society were to inadvertently throw Abraham Lincoln's pen into the dustbin. The disvalue of this act is not grounded in the intrinsic properties of the pen, yet may not be a relational value either as it is regrettable for everyone. Such values raise important questions, but they do not – as far as I can see – affect the points I am going to make. Thanks to Robert Pál-Wallin for having drawn my attention to this issue.

[3] This goes against some forms of naturalism, which understand value on the model of natural kinds (Copp 2007). I shall rest content here with observing that a natural kind inspired approach to value would probably dissolve the overcounted adicity objection rather than take it seriously.

qualify as revisionary as "there is hardly any lexical room for anyone to disagree with a version of FA restricted to such value properties." (Orsi and Garcia 2021: 1220) This reaction is correct as far as the existence of a constitutive relation between emotional values and emotions is concerned: reference to the emotions in the analysis of emotional values should not come as a surprise. Still, as we shall realize in §2, the FA analysis would also run against the folk understanding of emotional values if it were to cast emotions and properties of the object on an equal footing as sources of value. One gets a whiff of the different roles played by these factors by contrasting these two claims: "the joke is funny because of its timing and incongruity" and "the joke is funny because it merits amusement".

So, the FA analysis faces the overcounted adicity objection. According to Toni, it can steer clear of this objection by appealing to a distinction between sources of value and value enablers, a distinction to which I now turn.

§2 Enablers as Rescue Team

In a recent paper co-written with Wlodek Rabinowicz, Toni replies to the overcounted adicity objection (Rabinowicz and Rønnow-Rasmussen 2021).[4] Here is how I understand the reply. The objection would be premised on a mistaken understanding of the role psychological attitudes play within the FA analysis. The mistake consists in thinking that, according to the FA analysis, attitudes are *sources of value*, i.e. that from which an object's value (partly) originates. The FA analysis does not (or at least should not) cast attitudes in the role of sources of value, the reply continues, but rather in the role of *value enablers*.[5] Since the key notion of an enabler is borrowed from Dancy's seminal discussion (Dancy 2004: chap.3), let us first try to get a grip on how Dancy understands it.[6]

It helps to follow Dancy's steps and focus first on the contrast between enabler and favourer in the realm of reasons. A favourer is a consideration that speaks in favour of something. To take Dancy's own example, the fact that one promised to help a friend is a consideration that (defeasibly) speaks in favour of one's helping her. Similarly, the fact that it rained ten minutes ago is a consideration that

[4] One caveat. Toni's response is actually directed at Orsi and Garcia's (2021, 2022) explanatory objection to the FA analysis. I shall excise the appeal to the distinction between sources of value and value enablers from its role in addressing this objection. What interests me is the very idea that emotions are value enablers.

[5] As Rabinowicz and Rønnow-Rasmussen acknowledge, the distinction between source of value and value enabler is introduced by Orsi and Garcia (2021), who think that it is of no help to reply to their explanatory objection.

[6] There is a dash of irony here: with the idea of an enabler, Dancy would have provided a way for the advocates of the FA analysis to respond to the overcounted adicity objection that he puts forward.

(defeasibly) speaks in favour of believing that the streets are wet. Enablers play a different role, as they do not provide considerations in favour of the relevant action or attitude. In the case of the promise to help, one enabler mentioned by Dancy is the fact that the promise was not made under duress. According to him, this negative fact is not a consideration in favour of doing what one promised. It plays another role, that of enabling the promise to count in favour of doing what one promised. This enabling role is made manifest by the following counterfactual: if the promise had been made under duress, it would not speak in favour of doing what one promised. As regards the belief that the streets are wet, one may plausibly think of one's understanding of the concepts of a street and of wetness as enablers. One's understanding of these concepts does not play the same role as the fact that it rained ten minutes ago; it is not an additional consideration in favour of believing that the streets are wet. Understanding plays a different role, that of enabling the fact that it rained ten minutes ago to speak in favour of one's believing that the streets are wet. The enabling role of concept understanding is revealed by a parallel counterfactual: if one did not understand these concepts, the fact that it rained ten minutes ago would not speak in favour of one's believing that the streets are wet. In a nutshell, a variety of factors allow favourers to favour what they favour without themselves favouring. (Some of) these are enablers.

Armed with the distinction between favourers and enablers in the realm of reasons, let us now try to deploy it in the context of the FA analysis of value. This demands that we move from issues surrounding reasons to issues surrounding the ontology of value, as well as from relations between facts and attitudes to relations between attitudes and values. While we should in general be cautious in transposing claims from one of these domains to another, the transposition seems in the case at hand to proceed smoothly. Toni's strategy is to appeal to the distinction between favourers and enablers to avoid the worry that the FA analysis fixes the adicity of value one unit too high. This strategy has it that the worry is premised on an understanding of the FA analysis according to which attitudes play the role of favourers or sources of value. This is wrong-headed, Toni claims, since the FA analysis casts attitudes in the role of value enablers. True, the FA analysis has it that an object would not have a given value (a joke would not be funny) save for the existence of the relevant attitude (amusement). However, we realized that the truth of this counterfactual may reveal that the attitude enables the relevant properties of the object to be the sole source of its value. As regards the joke, the idea is that amusement enables the joke's timing and incongruity to constitute its funniness. The source of fun thus remains squarely within the boundaries of the joke; funniness comes out as a monadic value.

Recall the contrast that we met in §1 between two claims about amusement: "the joke is funny because of its timing and incongruity" and "the joke is funny because it merits amusement". The distinction between source of value and value enabler allows us to diagnose what governs this intuitively grasped contrast: the first claim refers to sources of funniness, the second to what enables these properties to

function as such sources. So, casting attitudes in the role of enablers not only holds the promise of preserving the adicity of values, it also corresponds to how we pre-theoretically think about the emotions and their relation to values. Suppose that someone challenges your claim that a joke is funny or that a painting is fascinating. In response, it is awkward if not downright misguided to refer to your emotions (Deonna and Teroni 2012). "Look, the joke amused me" and "I was fascinated by the painting" do nothing to vindicate your initial verdicts. The idea that emotions are enablers offers a convincing explanation of why this is so. Responses to such normative challenges that refer to emotions are awkward because the challenger is after the source of an object's value. The response he gets does not comply with that request, since it refers to an enabler. The situation is structurally similar to the one in which you ask a friend why she thinks that the streets are wet and she replies that she understands the relevant concepts. Hardly what you are after, even if it turns out that this understanding enables the fact that it rained ten minutes ago to favour your friend's belief.

Let me emphasize two essential aspects of the relation between emotions and values that the "emotions as enablers" claim nicely takes into account. The first aspect is normative. Your initial attribution of value (the joke is funny) has already informed the challenger that you think of the object as meriting a given emotion (amusement). So, responding to the challenge by pointing out that the object elicited that emotion is unsatisfactory. Casting emotions in the role of value enablers nicely explains why this is so. The second aspect is psychological. The fact that an object has elicited an emotion is rarely part of our perspective when we attribute values to objects (compare Dancy 2004: 46 on belief). True, we sometimes start reflecting on the value that an object may have by realizing that it engaged us emotionally. This hardly qualifies as the basic case, however: emotions rather tend to directly elicit value judgements about objects without a detour through such a reflexive process. And, even when reflexivity takes place, the way we think about the emotions supports the idea that they function as enablers. We may start reflecting on the object's value by realizing that we reacted emotionally to it, but we then try to locate which features of the object could support this emotional verdict. This again chimes well with the idea that emotions are value enablers.

§3 Emotions in the FA Analysis

We have realized how attractive the claim that emotions are value enablers is: it holds the promise of solving the overcounted adicity objection and corresponds to the way we pre-theoretically think about the emotions. Yet, to deliver on this promise, we need to know more about the role of emotions in the FA analysis. What exactly does the enabling role of emotions consist in? Is this role compatible with the FA analysis of values in terms of merited emotions? The difficulty is

compounded by the fact that, as Joseph Raz emphasizes, "the category of being an enabler is very diverse", so diverse in fact that there may well be "no theoretically interesting common feature among the roles which enablers perform" (Raz 2006: 106). This is not surprising, as our grip on enablers is mainly negative: enablers are not sources of value but factors that have to be in place so that sources of value function as such. Fortunately, we can further specify the enabling role of emotions in light of the account of value characteristic of the FA analysis.

The first thing to appreciate is that, while there are contingent enablers (the fact that I did not exercise may contingently enable a situation to be threatening, something which may also have been enabled by the fact that I broke my ankle, etc.), emotions cannot be cast in such a contingent role. According to the FA analysis, emotions are key to what emotional values are. The relation between a given emotion and a value is thus necessary. Nothing would be funny if there was no amusement, nothing would be fascinating if there was no fascination.[7] Moreover, the FA analysis aims at explaining values in terms of more basic constituents of reality: emotions and the non-evaluative properties of their objects that merit these emotions account for values, not the other way around.[8]

None of this threatens the claim that emotions are enablers: there are necessary enablers (one's capacity to discharge them may be a necessary enabler of duties) and an enabler may be a more primitive component of reality than what it enables (same example). Neither is the claim that emotions are enablers in tension with the fact that emotional value concepts "wear their emotional origin on their sleeves", so to say. This may well be what we should expect for necessary enablers. For instance, the concept of duty is partly structured around the "ought implies can" principle, and this despite the fact that a subject's capacity to act does not favour but enables. Similarly, the fact that values such as the funny, the fascinating, etc. parade their relations to emotions is not in tension with the claim that these emotions are (necessary) enablers.

The obvious relation between emotions and emotional values should lead us to insist on two points. The first is that this fact reveals that emotions are manifest in consciousness. This is perhaps best shown by appreciating how the FA analysis of values differs from a dispositionalist account of colours (e.g., Cohen 2009). The latter should come (and is typically sold) as a surprise – nothing in our pre-theoretical apprehension of colours makes such a connection with visual experience obvious. This contrasts with the FA analysis of emotional values. One explanation of this difference that I pursued elsewhere with Julien Deonna starts with the observation that visual experience is transparent, whereas emotional experience is opaque. Contrary to what happens in visual experience, we can focus our attention

[7] Amusement need not exist "in all worlds" where there is funniness (Rabinowicz and Rønnow-Rasmussen 2021) – the FA analysis typically proceeds in counterfactual terms.

[8] Toni expresses at many places his understanding of the FA analysis as an analysis *stricto sensu*, as opposed to the "no-priority view" championed for instance by Wiggins (1987).

on emotional experience (the experience of amusement or fear, for instance) rather than on what this experience is about (a joke, a nearby wolf). The fact that emotions are opaque means that they are available as building blocks for the relevant value concepts (Deonna and Teroni 2021). The second point bears upon the idea that the FA analysis appeals to constituents of reality that are claimed to be more basic than values. Emotions are our "entry point" into the realm of emotional values, since something has a value only insofar as it merits an emotion. This is well-trodden territory, but the point deserves emphasis as it is crucial for our discussion to build an understanding of emotions as value enablers that is in tune with the intimate relation between emotions and values, as the FA analysis portrays them.

Toni's own way of acknowledging these points while casting emotions in the role of value enablers leads him to bring up two possibilities:

> One possibility would be that attitudes come with inherent standards—say, admiration comes with criteria that specify what is to be admired, or perhaps with paradigmatic examples—and that because of these standards some properties of objects (in virtue of which the objects satisfy or approximate the standards) become value-makers. Because of the inherent standards of admiration certain properties of an object make it admirable. But a more obvious and less contentious option is that the properties of an attitude enable the properties of the object to be value-makers simply because they determine the nature of the value in question. Clearly, admirability is a value whose nature in part is determined by what the attitude of admiration consists in. Likewise, the nature of desirability in part is determined by the constitutive properties of desire, and so on. (Rabinowicz and Rønnow-Rasmussen 2021: 2476-2477)

In his discussion, Toni's aim is not to flesh out these possibilities – the ideas of inherent standards and of value determination – in any detail, and I am not confident I fully understand them and how they relate to one another. What follows is an attempt to clarify these issues to get a grip on the role of emotions in the FA analysis, and on whether this role is compatible with the claim that they function as value enablers.

§4 Truth vs. Emotional Values

In this section, I shall focus on a natural strategy for developing the claim that emotions are value enablers thanks to their inherent standards and/or thanks to their contribution to value determination. The strategy consists in using the relations between belief and truth as one's model. While I shall argue that this model does not pay sufficient attention to the role of emotions in the FA analysis, its exploration will help illuminate what we are after.

The strategy is to flesh out the claim that emotions are value enablers by understanding the relation between emotions and values on the model of the relation between belief and truth. Here is how the strategy will unfold. There are facts, truths and beliefs. Belief is an attitude that is correct if and only if the proposition believed is true.[9] Absent belief (or representation more generally), there would be facts, but there would be no truths. There are monadic facts, but there are no monadic truths: truths wear their relational nature on their sleeves. Talk of truth is talk of correspondence between beliefs and what is (the facts). There need be no actual belief for a fact to be a truth. Still, belief enables facts to constitute truths – a fact is a truth thanks to its corresponding to a belief. Misunderstanding of what truth is and of the dependence of truth on belief would lead one to systemically overcount the adicity of truths. A truth is always a fact, although the nature of truth is not to be found in the intrinsic properties of any fact that is a truth: it is rather to be found in the function of that fact – the function of being that against which the correctness of a belief is assessed. Belief is not a source of truth, although a detour via belief is needed for talk of truth to get traction. This is the sense in which belief enables facts to constitute truths.[10]

The two possibilities brought up by Toni to flesh out the claim that attitudes are enablers materialize in the relation between belief and truth. First, belief comes with inherent truth-related standards and it is in light of the foregoing observations attractive to claim that "because of the inherent standards of [belief] certain properties of [a fact] make it [a truth]." Second, these same observations also make a convincing case for the claim that "[truth] is a value whose nature is in part determined by what the attitude of [belief] consists in". The temptation is thus strong to apply the model of the relations between facts, truths and beliefs to the relations between natural properties of objects, values and emotions.

Here is how the application proceeds. The relevant triad is now: intrinsic properties of objects, values and emotions. Emotions are attitudes that are correct if and only if their objects have specific values. Amusement is correct if and only if its object is funny, shame is correct if and only if its object is shameful, etc.[11] According to the FA analysis, absent the emotions, objects would have natural properties, but they would have no values. Talk of value is talk of attitudes that objects merit in virtue of their natural properties.[12] In this sense, (possible) emotions

[9] I shall concentrate on this widespread idea and leave aside debated issues in the normativity of belief. On these issues, see Fassio (2015).

[10] Dummett's (1996: chap.2) claims regarding the origin of the concept of truth in the correctness of assertions have interesting similarities with these remarks on truth. I am indebted to Roberto Keller and Hemdat Lerman for this reference.

[11] On this idea, which is common territory amongst different approaches to the emotions, see Deonna and Teroni (2022a).

[12] This is the "buck-passing" account of value, according to which value does not give reasons for attitudes – only the properties on which value supervenes do (Scanlon 1998).

enable talk of values. Misunderstanding of what value is and of its dependence on emotions would lead one to overcount the adicity of value. A (monadic) value is always a property of the relevant object, although the nature of value is not to be found in the intrinsic properties of any valuable object: it is rather to be found in the function of these properties – the function of being that against which the merit of emotions is assessed. Emotions are not sources of value, although a detour via emotions is needed for value talk to get traction.

As attractive as this application of the model of the relations between facts, truths and beliefs to the relations between natural properties, values and emotions may look at first sight, it would be a mistake to adopt it as part of an FA analysis. Put in the context of the FA analysis, this model suffers indeed from a fatal flaw, since it overlooks a basic difference in our apprehension of truths and of emotional values. The difference is this: truth is not the property of meriting belief, whereas there is, according to the FA analysis, no ontological gap between emotional values and what merits the relevant emotions (no gap between the funny and what merits amusement, for instance). The fact that truth is not what merits belief transpires from the discussions surrounding trivial truths, such as the proverbial number of sand grains on a beach. A trivial truth is a fact that does not merit belief (see e.g., David 2001).

This should lead us to distinguish two claims. The first claim is that talk of a given entity is talk of that entity in relation to a psychological attitude. If the foregoing is along the right track, truth is such an entity: truth talk is talk of facts in relation to beliefs. The second, stronger claim is that talk of this entity is structured around the idea of what merits this psychological attitude. This is not the case for truth: trivial truths are testimony to the fact that truth talk is not structured around the idea of what merits belief. Truth is a property that tallies with the functional approach sketched above: a truth is a fact insofar as that fact corresponds to a belief, but the psychological relatum of this relation does not feature centrally in our understanding of truths. The domain of truths is the domain of facts as these facts correspond to beliefs, but we often speak of truths independently of their relation to merited beliefs.

This is why we should not apply this model to the relation between natural properties, values and emotions. According to the FA analysis, there is in the value domain no distinction that parallels that between truth and what merits belief. There are just the natural properties and the value: funniness and the fascinating are not properties that need to be refined, so to say, into what merits amusement and fascination – they are the very properties of meriting these emotions.[13] Something

[13] This is often supported by the so-called "shapelessness" of values vis-à-vis non-evaluative properties – roughly, the idea that there is no unity to the various constellations of natural properties that are instances of funniness or shamefulness. No unity, that is, except the one that they receive thanks to all meriting amusement or shame. On this issue, see Roberts (2011).

that is too bland to be amused by is not fun, something that is too superficial to merit shame is not shameful, etc.[14]

There is thus an important contrast between truth and values: only in the latter case does the idea of merited attitudes structure our basic understanding of the relevant entities. So, if emotions are not sources of value but value enablers, this cannot be assimilated to the claim that beliefs enable facts to be truths. This contrast is not to be taken lightly, since it reveals something crucial about values and how they differ from truth, as I shall argue in the next sections.

§5 Enabling as Accessing and its Limits

If our aim is to develop the idea that emotions are enablers in a way congenial to the FA analysis, how should we react to the contrast between truth and emotional values? In this section, I shall consider a first such reaction, which diagnoses the contrast in representational terms.

The reaction consists in claiming that the differing relations between belief and truth and between emotions and values are explained by the properties that belief and emotions respectively represent. According to this line of thought, the belief that p represents the truth of p, whereas emotions represent values – amusement represents the funniness of the joke, etc. The contrast at issue would be due to the fact that beliefs and emotions represent distinct properties: beliefs represent a property – truth – that differs from what merits belief (as is evidenced by trivial truths), whereas emotions represent values that are (as per the FA analysis) identical to what merits these emotions. Call this the "pure access" account.

This account allows us to implement the claim that emotions are enablers in the following way. Talk of value is talk of normativity. Now, the fact that a normative claim bears on a subject presupposes that he accesses the relevant entities. Consider a rather uncontroversial example: if you ought to mow the lawn today, you are in a position to access that fact (or at least have been in a position to access it). Of course, we cannot transpose what holds for obligations directly to the relations between emotions and values, as the FA analysis under discussion does not equate emotional values with what *ought* to elicit a given emotion, but rather with what *merits* to elicit it.[15] Still, one may insist that something merits a response from a subject only if this

[14] The existence of this contrast between truth and emotional values does not mean that the FA analysis implies that, all things considered, we should be amused by all funny jokes. The contrast only suggests that, whereas a proposition can be true without meriting belief, an object always has a value in virtue of meriting a response (even if, all things considered, we should not respond in this way). To put the point in another terminology, the FA analysis under discussion targets *pro tanto* value. I am indebted to Jakob Werkmäster for having insisted on this issue.

[15] The FA analysis is rarely laid out in terms of obligations, because emotions are not directly subject to the will. On this issue, see Gert (2003) and Svavarsdóttir (2014).

subject can access it. Emotions would allow properties of their objects – the sources of value – to function as such by providing the sort of access that is a precondition of value talk. A remark may be incongruous and have a specific timing without anyone accessing it, a canvas may exemplify forms and colours without anyone accessing them. However, the funny and the fascinating, being normative properties constituted by a relation of merit between these properties and an emotion, presuppose that the relevant subjects access these properties. Emotions are value enablers in virtue of providing such an access. And they are necessary value enablers because we cannot access the relevant properties in any other way.[16]

This understanding of the contrast between truth and emotional values in terms of representation and the associated pure access account chime well with a widespread approach according to which emotions are representations of values (e.g., Milona 2016, Tappolet 2016). Whatever its merits[17], it will not do in the present context, which consists in trying to combine the idea that emotions are enablers with the FA analysis.

Casting emotions in the role of pure access providers indeed goes against the spirit of the FA analysis. Advocates of the FA analysis – and John McDowell (1985) in particular – have emphasized that it differs from dispositionalism since it does not refer to the responses that an entity tends to generate, but to the responses that it merits. This is difficult to square with the idea that emotions are enablers in virtue of providing access to value. To see why, observe that when we talk of merit in the mental realm, we do so only in relation to emotions, desires and, perhaps, beliefs.[18] What do these mental states have in common? Their primary business seems to be that of *responding to what we represent*, rather than that of representing: emotions, desires and beliefs are reactions to what we apprehend perceptually or otherwise.[19] They contrast with mental states which are in the business of representing – which are pure access providers – and to which the notion of merit does not apply. Consider two central cases: perception as a way of accessing our surroundings and (episodic) memory as a way of accessing our personal past. In both cases, the notion of merit fails to get a grip: colours, shapes, forms and past events are not entities that we understand by reference to what merits to be perceived or remembered.

So, there seems to be an essential relation between the notion of merit and psychological reactions. Why is that so? Well, to merit is to deserve or to be worthy of a response, on account of some properties. And the properties to which the

[16] This last claim raises a number of issues in the epistemology of emotion that I cannot enter into here. For a discussion, see Deonna and Teroni (2022b).

[17] I assess them e.g., in Deonna and Teroni (2012: chap.6).

[18] Merit seems to qualify belief only if we allow some stretching – "merits belief" seems to refer to trust rather than to propositional belief.

[19] The fact that emotions are reactions is at the centre of the approaches defended in Deonna and Teroni (2012, 2022), Müller (2019) and Mulligan (2007).

response responds must somehow be represented – otherwise we would not face a (merited) response, but a mere coincidental cooccurrence. It is thus no surprise that mental states can be relata of the merit relation only when they react to the representation of such properties. Desire qualifies because it is a reaction of pursuing what one apprehends perceptually or otherwise. The same is true of the emotions, which also react to our apprehension of their objects – one is sad about an event one remembers, amused by a remark one hears or fascinated by a painting one sees.

Another way to drive the point home is to emphasize that the pure access account gives with one hand what it takes back with the other. It claims that emotions represent a property that merits to be responded to in a given way. However, in casting the emotions in the role of representing what merits to be responded to in such a way, the account prevents the emotions to constitute these responses. This is the case because, when we say that a joke is funny, we certainly don't (only or primarily) say that it merits to be represented. This would hardly make sense and would completely neglect the specific contribution of emotions: the fact that they are ways of (dis)favouring, which advocates of the FA analysis (and Toni in particular)[20] rightly put at the centre of their account of value.

All in all, we should not cast emotions in the enabling role of pure access providers – this fails to do justice to the idea of merit at the centre of the FA analysis. Is it possible to preserve the idea that emotions are enablers while insisting on the fact that they are responses to what we represent? The last section is devoted to this issue.

§6 Enabling via Attitudinal Shape

Let us trace back our steps a little. There is an important contrast between truth and emotional values. A first reaction is to claim that this contrast derives from what beliefs and emotions respectively represent. We have concluded that this reaction and its companion pure access account of the emotions does not do justice to the key insight of the FA analysis. I now wish to put forward an alternative reaction according to which the contrast at issue is rather due to the fact that emotions have attitudinal shapes that importantly differ from that of belief.

In a nutshell, the pure access account is unsatisfactory because it misunderstands what we mean when we say that an object merits an emotion: we do not thereby mean that it merits to be represented. We rather mean that the emotion somehow "does justice" to the object. Can we develop this idea in less metaphorical terms? I

[20] Toni acknowledges (2022: 128-129) that his previous attempt to connect the notion of merit to representational content is not satisfactory. His suggestion to connect it instead to psychological modes corresponds to the account in terms of attitudinal shapes that I shall present in the next section.

think that we can if we take into consideration the rich attitudinal shapes of emotions.

How best to characterize emotional shapes? We should do so by emphasizing the essential relations between emotions and attention. When we undergo an emotion, our cognitive resources are channelled towards its object so as to prepare us to deal with it in a specific way – emotions are partly constituted by distinct action tendencies (Deonna and Teroni 2012, Frijda 2007, Scarantino 2014). In fear, our body is mobilized to neutralize something; in anger, it is mobilized for a form of active hostility; in shame, for moving away from the gaze of others; in fascination, for further exploring an object that we may have difficulties in fully grasping. These ways of dealing with the object can take place at the level of behaviour and/or thought. Now, these emotional responses essentially take time, and the time that they take is a function of the aims of the emotions, which consist in modifying or maintaining a given relation with an object. Emotions are individuated by a system that organizes consciousness and channels it towards fulfilling these aims. In fear, this takes place within an organization of the subject's resources to avoid a threat; in sadness, these resources are organized to cope with a loss; in fascination, to further explore an object for its own sake, etc.

The foregoing observations suggest that the notion of merit applies to the emotions because they organize and occupy consciousness in specific processual ways. This sounds terribly abstract, but it actually chimes well with how we pretheoretically conceive of value, and constitutes one key asset of the FA analysis. Our conception of emotional values is essentially the conception of what merits to occupy consciousness and attention in specific ways. A threat is something that merits to be avoided, an offense something that merits to be righted, the funny what merits to be laughed at. And when we deplore a lack of emotion in ourselves or others, we claim that an object merits to occupy consciousness in a given way. We thus assess whether emotions are correct as a function of whether their objects merit to occupy consciousness in these ways (Deonna and Teroni 2021). There can be too little or too much fear directed at a threat, as there can be too little or too much amusement directed at a joke – "too little" and "too much" cover here the intensity as well as the duration of emotional processes.

This contrasts sharply with the way we assess intellectual states like judgements and beliefs, where these ideas fail to get a grip (Na'aman 2021). Failing to realize this is the fundamental shortcoming of the pure access account discussed in §5. As Sigrún Svavarsdottir writes regarding our criticism of unmerited emotions,

> The alleged mistake is not that of misrepresentation but, rather, that of misplacement of emotional and motivational energies. It is some kind of misplacement or waste of emotional and motivational resources to train them on an object of little or no value. That is, I submit, the drift of the criticism. (Svavarsdóttir 2014: 89)

Talk of what merits emotions is talk of what merits to occupy consciousness in specific ways. Still, we have observed above that, in order to qualify as (merited) responses to some properties of objects, emotions should not simply cooccur with the representation of these properties. Do emotions qualify as such responses? They do, insofar as they display the relevant sensitivity to evidence – and it seems plausible to think that amusement is sensitive to fun-related considerations, anger to offense-related considerations, etc. (Deonna and Teroni 2022a). If this is along the right track, we can maintain that emotions are ways of responding to objects – of favouring or disfavouring them – to which the notion of merit applies.

The final question is: how does the claim that emotions are enablers fare in relation to this characterization of their attitudinal shapes? Recall the two possibilities that Toni puts forward to flesh out the claim that attitudes are enablers: the first refers to the idea that attitudes come with inherent standards, the second to the idea that values are partly determined by the relevant attitudes. What we have just said about the attitudinal shapes of emotions and their sensitivity to evidence supports the idea that some properties of objects – the incongruity and timing of remarks, the distribution of forms and colours on canvas, etc. – are sources of values because of the existence of emotions with specific attitudinal shapes and inherent standards. Emotions have some plasticity, as they can be directed towards new objects – the plausible claim that we are endowed with different sensibilities (to the funny, the shameful, the fascinating, etc.) that evolve during our lives requires that much. But this plasticity is restricted – in order to engage our emotional responses, these new objects must relate, by several intermediate steps perhaps, to "paradigmatic scenarios" (de Sousa 1987, D'Arms and Jacobson 2010). The paradigmatic scenario for anger may be situations that constitute unjustified encroachments on the ends that we pursue (jostling someone, intrusion in our private space, etc.); that of amusement may be of the "slipping on a banana peel" or "grossly breaking a social norm" variety; that of fear situations in which we are at the mercy of a predator.

The plasticity of emotions is moderate and anchored in these paradigmatic scenarios precisely because the attitudinal shapes of emotions are determinate: emotions are never thin favourings or disfavourings, they are always determinate ways of doing so that lend themselves to the sorts of descriptions sketched above.[21] This puts substantial constraints on what can merit a given emotional response. Almost any fact can be believed, almost any situation can be desired, but only quite specific objects or situations can be responded to with fear, anger, amusement, fascination or shame. It is in these determinate attitudinal shapes that we should anchor the function of emotions to partly determine emotional values. The restricted plasticity of emotions allows them to trace relatively stable and interpersonally recognizable paths through the space of natural properties, and the properties that make up these paths are disunified except for the fact that they merit these specific

[21] This chimes well with Toni's appeal to thick attitudes (Rønnow-Rasmussen 2022: 82).

forms of attention.[22] As opposed to this, belief displays no restricted plasticity and its shape is much less determinate than those of the emotions. This may well be the source of the contrast between truth and what merits belief. Whatever the final verdict about belief, the two possibilities mentioned by Toni turn out, at least in the case of the emotions and their relation to values, to be two faces of the same coin.

Conclusion

Starting with the worry that the FA analysis fosters a revisionary understanding of values, I explored different ways of combining Toni's response to this worry – emotions do not function as sources of value but as value enablers – with the FA analysis. I have emphasized a contrast between the way belief relates to truth and the way emotions relate to values and examined two reactions to this contrast. I have argued that the first reaction, according to which emotions enable by granting access to values, is unattractive and have endorsed a second reaction, according to which emotions enable thanks to their attitudinal shapes.[23]

Bibliography

Cohen, Jonathan (2009). *The Red and the Real. An Essay on Color Ontology*. New York: Oxford University Press.

Copp, David (2007). *Morality in a Natural World. Selected Essays in Metaethics*. New York: Oxford University Press.

D'Arms, Justin and Jacobson, Daniel (2010). Demystifying Sensibilities: Sentimental Values and the Instability of Affect. In Peter Goldie (ed.), *The Oxford Handbook of Philosophy of Emotion* (pp.585-613). New York: Oxford University Press.

Dancy, Jonathan (2000). Should we Pass the Buck? In A. O'Hear (ed.), *Philosophy, the Good, the True and the Beautiful, Royal Institute of Philosophy Supplement*, 47: 159-173.

Dancy, Jonathan (2004). *Ethics Without Principles*. Oxford: Oxford University Press.

[22] You may remember the aforementioned claim (ftn13) that values are shapeless vis-à-vis non-evaluative properties. In the present context, it should be developed as follows. The source of value of each concrete instance of a monadic emotional value lies in the natural properties of an object. But what unifies the various constellations of relevant natural properties – what makes them qualify as instances of a given emotional value – is their relation of merit to the attitudinal shape of an emotion, which functions as an enabler.

[23] I am grateful to Julien Deonna, Roberto Keller, Julia Langkau, Robert Pál-Wallin and Jakob Werkmäster for their comments on a previous version of this paper.

David, Marian (2001). Truth as the Epistemic Goal. In M. Steup (ed.), *Knowledge, Truth, and Duty* (pp. 151-170). New York: Oxford University Press.

De Sousa, Ronald (1987). *The Rationality of Emotion*. Cambridge, MA: MIT Press.

Deonna, Julien and Teroni, Fabrice (2012). *The Emotions. A Philosophical Introduction*. New York: Routledge.

Deonna, Julien and Teroni, Fabrice (2021). Which Attitudes Fit the Fitting Attitude Analysis of Value? *Theoria* 89: 1099-1122.

Deonna, Julien and Teroni, Fabrice (2022a). Emotions and their Correctness Conditions: A Defense of Attitudinalism. *Erkenntnis.*

Deonna, Julien and Teroni, Fabrice (2022b). Why Are Emotions Epistemically Indispensable? *Inquiry*.

Dummett, Michael (1996). *The Seas of Language*. Oxford: Oxford University Press.

Fassio, Davide (2015). The Aim of Belief. *Internet Encyclopaedia of Philosophy*.

Frijda, Niko (2007). *The Laws of Emotion*. Mahwah, NJ: Lawrence Erlbaum.

Gert, Joshua (2003). Requiring and Justifying: Two Dimensions of Normative Strength. *Erkenntnis* 59: 5-36.

McDowell, John (1985). Values and Secondary Qualities. In T. Honderich (ed.), *Morality and Objectivity* (pp. 110-129). London: Routledge.

Milona, Michael (2016). Taking the Perceptual Analogy Seriously. *Ethical Theory and Moral Practice* 19(4): 897-915.

Müller, Jean Moritz (2019). *The World-Directedness of Emotional Feeling: On Affect and Intentionality*. Cham: Palgrave Macmillan.

Mulligan, Kevin (2007). Intentionality, Knowledge and Formal Objects. *Disputatio* 2(23): 205-228.

Na'aman, Oded (2021). The Rationality of Emotional Change: Towards a Process View. *Noûs* 55(2): 245-269.

Orsi, Francesco (2015). *Value Theory*. London: Bloomsbury.

Orsi, Francesco and Garcia, Andrés (2021). The Explanatory Objection to the Fitting Attitude Analysis of Value. *Philosophical Studies* 178: 1207-1221.

Orsi, Francesco and Garcia, Andrés (2022). The New Explanatory Objection Against the Fitting Attitude Account of Value. *Philosophia* 50(4): 1845-1860.

Rabinowicz, Wlodek and Rønnow-Rasmussen, Toni (2004). The Strike of the Demon: On Fitting Pro-attitudes and Value. *Ethics* 114(3): 391-423.

Rabinowicz, Wlodek and Rønnow-Rasmussen, Toni (2021). Explaining Value: On Orsi and Garcia's Explanatory Objection to the Fitting-attitude Analysis. *Philosophical Studies* 178: 2473-2482.

Raz, Joseph (2006). The Trouble with Particularism (Dancy's Version). *Mind* 115.457: 99-120.

Roberts, Debbie (2011). Shapelessness and the Thick. *Ethics* 121(3): 489-520.

Rønnow-Rasmussen, Toni (2011). *Personal Value*. Oxford: Oxford University Press.

Rønnow-Rasmussen, Toni (2022). *The Value Gap*. Oxford: Oxford University Press.

Scanlon, Thomas (1998). *What we Owe to Each Other*. Cambridge, Mass.: Harvard University Press.

Scarantino, Andrea (2014). The Motivational Theory of Emotions. In J. D'Arms and D. Jacobson (eds.), *Moral Psychology and Human Agency* (pp.156-185). Oxford: Oxford University Press.

Stratton-Lake, P. (2013). Dancy on Buck-passing. In D. Backhurst (ed.), *Thinking About Reasons: Themes from the Philosophy of Jonathan Dancy* (pp.76-96). New York: Oxford University Press.

Svavarsdóttir, Sigrún (2014). Having Value and Being Worth Valuing. *The Journal of Philosophy* 111(2): 84-109.

Tappolet, Christine (2016). *Emotions, Value, and Agency*. Oxford: Oxford University Press.

Wiggins, David (1987). A Sensible Subjectivism? In D. Wiggins (ed.), *Needs, Values, Truth: Essays in the Philosophy of Value*. Oxford: Blackwell.

Causation, Responsibility, and Norms

Re-evaluating Our Norms in the Face of Climate Change

Caroline Torpe Touborg

Abstract. In this paper, I combine two ideas. First: causation is relative to a possibility horizon, i.e., a class of possible worlds containing just those worlds that represent serious possibilities. Second: you are responsible for an outcome only if you have performed an action or omission that caused the outcome. Combining these two ideas raises a question: what is the relevant possibility horizon for evaluating whether the causal condition for responsibility is satisfied? I suggest that norms play a role in selecting the relevant possibility horizon. This yields a picture with norms as input. These norms select the relevant possibility horizon for assessing whether the causal condition for responsibility is satisfied. And together with other conditions for responsibility, this yields the output – namely, attributions of responsibility. This picture suggests a way to evaluate norms: when a bad outcome (such as climate change) occurs, taking one set of norms as input may yield the output that nobody is responsible – it just happened. By contrast, a different set of norms may allow us to attribute responsibility. In that case, I suggest that we should, *ceteris paribus*, prefer the norms that allow us to attribute responsibility for the outcome.

Until recently, it was widely assumed that causation is a binary relation: *c* is a cause of *e*. However, a growing number of authors now question this assumption.[1] An

[1] See footnote 3.

alternative suggestion is that causation is a three-place relation: *c* is a cause of *e* within a possibility horizon H, where a possibility horizon is simply a class of possible worlds containing just those possible worlds that represent serious possibilities. My aim in this paper is to explore what happens when we combine this idea with the widely held idea that causation is a necessary condition for responsibility: you are responsible for an outcome only if you have performed an action or omission that caused the outcome.

I begin by setting out the motivation for thinking that causation is a three-place relation with a possibility horizon as its third relatum (section 1). Once we combine this idea with the idea that causation is necessary for responsibility, we face a question: how do we select a possibility horizon for evaluating whether the causal condition for responsibility is satisfied? (section 2). In answering this question, my aim is to offer an idealised description of how we, as a matter of fact, make this selection. Based, among other things, on experimental work, we find that norms play a crucial role in how we select a possibility horizon. For moral responsibility, moral norms play this role; for legal responsibility, legal norms play this role, etc. (section 3). This yields a picture with norms as input; these norms select the possibility horizon we use to assess whether the causal condition for responsibility is satisfied; and together with other conditions for responsibility, this yields the output, which is attributions of responsibility (section 4).

In the second part of the paper, I argue that this picture of the relation between norms, causation, and responsibility suggests a way to evaluate norms: when a bad outcome (such as climate change) occurs, taking one set of norms as input may yield the output that nobody is responsible; taking a different set of norms as input may yield the output that some agents *are* responsible. In that case, I suggest that a system of norms that attributes responsibility for the bad outcome is, *ceteris paribus*, superior to a system of norms that does not: a system of norms that attributes responsibility for the bad outcome allows us to give a particular kind of *explanation* of why the bad outcome happened – namely, an explanation in terms of someone's having violated the norms (section 5). I end by showing how this way of evaluating norms may be used to re-evaluate our norms in the face of climate change (section 6).

There is an important shift between the first and the second part of the paper: in the first part of the paper, my aim is simply to offer a schematic *description* of how causation, responsibility, and norms relate to each other in our practice of attributing responsibility for outcomes. Such a description of our practice of course leaves it open whether this is how we *should* go about attributing responsibility for outcomes: is our practice of attributing responsibility for outcomes sound, or should it be revised? I do not attempt to answer this question. However, in the second part of the paper, I consider what follows if we assume that our practice *is* sound: I suggest that *if* our practice of attributing responsibility for outcomes is sound, then we may evaluate different systems of norms in terms of the attributions of responsibility they can support.

1. Causation Within a Possibility Horizon

In the following, I intend to talk about the ordinary, everyday notion of causation – the one we appeal to when we say that "the gust of wind caused the curtains to flutter" or "the lack of grass caused the wildebeest to migrate." This ordinary notion of causation has two important features. First, omissions and absences may play the role of cause and effect, as in the example above where the lack of grass (an absence) is cited as a cause of the wildebeest's migration. Second, causation is selective. As Hart and Honoré write:

> In most cases where a fire has broken out the lawyer, the historian, and the plain man would refuse to say that the cause of the fire was the presence of oxygen, though no fire would have occurred without it: they would reserve the title of cause for something of the order of a short-circuit, the dropping of a lighted cigarette, or lightning. (Hart and Honoré, 1985, p. 11)

This second feature of causation has prompted a growing number of authors to question the orthodoxy that causation is a binary relation between a cause *c* and an effect *e*. To see why, consider the following example:

Suppose that there is a lightning strike, and a forest fire starts thereafter. In ordinary contexts, such as a conversation among the forest rangers, it seems inappropriate to assert (1):

(1) The presence of oxygen caused the forest fire.

Indeed, if one of the forest rangers were to assert (1), it would be perfectly appropriate if a second forest ranger replied by denying this, saying

(2) The presence of oxygen did not cause the forest fire.[2]

There are, however, more unusual contexts where these judgements are reversed. Putnam gives the following charming example:

> Imagine that Venusians land on earth and observe a forest fire. One of them says, "*I* know what caused that – the atmosphere of the darned planet is saturated with oxygen." (Putnam, 1982, p. 150)

[2] In saying that it would be perfectly appropriate for a second forest ranger to respond by uttering (2), I am following the verdict of Kaiserman (2017), McGrath (2005), and Schaffer (2012). What matters here is simply that the second forest ranger *could* choose to respond by uttering (2), and this response would be appropriate. But of course, the second forest ranger *need* not respond in this way. She might instead respond in a more conciliatory way – first acknowledging that there is indeed a way to look at the situation where the presence of oxygen counts as a cause, and then directing attention towards the present causal inquiry: "why was there a fire now in this forest, rather than at some other time or in some other forest?" In *this* causal inquiry, the presence of oxygen is merely a background condition and does not count as a cause.

In this example, it is perfectly appropriate for the visiting Venusian to assert (1), and it would be inappropriate for the others to reply by asserting (2).

The orthodox view that causation is a binary relation has a hard time accommodating these data. Instead, several authors have suggested that causation has a third relatum, which we may call a possibility horizon. Let me explain this in more detail:

It is natural to understand the difference between the context of the forest rangers and the context of the visiting Venusians in terms of which possibilities are taken seriously in the two contexts. In the context of the forest rangers, the presence of oxygen is simply taken for granted – in this context, it is not treated as a serious possibility that there might have been no oxygen. In the context of the visiting Venusians, by contrast, it *is* treated as a serious possibility that there might have been no oxygen. We may use the notion of a *possibility horizon* to capture this idea, where a possibility horizon is simply a class of possible worlds. A possibility horizon H represents which possibilities we are taking seriously and which we are ignoring: if a world is included in H, it represents a possibility we are taking seriously; if a world is not included in H, it is being ignored. The suggestion then is that causation is relative to a possibility horizon:[3]

> *Ternary causation*: the causal relation has three relata and takes the form: *c* is a cause of *e* within possibility horizon H.

According to *Ternary causation*, the causal relation has three relata: the cause *c*, the effect *e*, and a possibility horizon H. Correspondingly, a complete causal claim takes the form "*c* is a cause of *e* within possibility horizon H." But of course, we do not usually mention possibility horizons when we are making causal claims. Rather, a typical causal claim takes the form "*c* is a cause of *e*," and the relevant possibility horizon is supplied by context. In a context where the relevant possibility horizon is H_1, for example, an utterance of "*c* is a cause of *e*" expresses the complete causal claim "*c* is a cause of *e* within possibility horizon H_1"; in a context where the relevant possibility horizon is H_2, an utterance of "*c* is a cause of *e*" expresses the different complete causal claim "*c* is a cause of *e* within possibility horizon H_2."

When is such a complete causal claim true? I will not here attempt to set out necessary and sufficient conditions for causation within a possibility horizon.[4] For

[3] This suggestion is implicit in the causal modelling approach to causation exemplified in the work of e.g. Halpern, Hitchcock, Pearl, and Woodward. According to the causal modelling approach, causation is relative to a causal model (see e.g. Halpern and Pearl, 2005, p. 845), and any given causal model represents certain possibilities and leaves out others. In this way, the claim that causation is relative to a model subsumes the claim that causation is relative to a possibility horizon. Woodward, for example, is clear that his causal modelling approach involves "relativizing causal judgments to a set of serious possibilities (or, what I take to be the same thing, to the choice of some system of representation that reflects those possibilities)" (Woodward, 2003, p. 90). Outside the causal modelling framework, the idea of relativizing causation to a possibility horizon is developed by Kaiserman (2017) and Touborg (2018).

[4] For an attempt to do so, see Touborg (2018).

now, all we need is the following *necessary* condition, which captures the idea that background conditions do not count as causes:

> *Candidate cause*: *c* is a cause of *e* within H only if there is a world in H where *c* does not occur.

In the case of the forest fire, this allows us to capture the data as follows: in the context of a conversation among the forest rangers, it is not treated as a serious possibility that there might have been no oxygen – the forest rangers are taking the presence of oxygen for granted. Correspondingly, the relevant possibility horizon $H_{Rangers}$ does not include any worlds where there is no oxygen. In the context of the forest rangers, (1) expresses the complete causal claim (1*):

> (1*) Within possibility horizon $H_{Rangers}$, the presence of oxygen caused the forest fire.

Since *Candidate cause* fails to be satisfied, (1*) is false. Correspondingly, its negation – which is expressed by (2) – comes out true. This fits with the observation that it is inappropriate to assert (1) in the context of the forest rangers and appropriate to assert (2).

In the context of the visiting Venusians, by contrast, it *is* treated as a serious possibility that there might have been no oxygen – after all, there is no oxygen where the Venusians come from. Correspondingly, the relevant possibility horizon $H_{Venusians}$ *does* include worlds where there is no oxygen. In the context of the Venusians, an utterance of (1) expresses the complete causal claim (1**):

> (1**) Within possibility horizon $H_{Venusians}$, the presence of oxygen caused the forest fire.

Here, *Candidate cause* is satisfied. It is also clear that the remaining conditions for causation (whatever they are) are satisfied. Thus, (1**) comes out true. Correspondingly, its negation – which is expressed by (2) – is false. This fits with the observation that it is appropriate to assert (1) in the context of the Venusians and inappropriate to assert (2).

2. Causation and Responsibility

How is causation related to responsibility?

The kind of responsibility I am interested in here is responsibility in the sense of *accountability*, as McKenna characterises it below:

> Treating another as accountable is treating her as one who is a candidate for moral demands and thus as one who is held to expectations that when complied with (or exceeded) merits praise and sometimes reward, and when violated merits blame and sometimes punishment. (McKenna, 2012, pp. 7-8)

Whenever I talk about "responsibility" in the remainder of this paper, it should be understood in this sense of accountability. The objects of responsibility – that is, the things one may be held responsible *for* – include actions, omissions, outcomes, and maybe more. In the following, I will focus on responsibility for outcomes – specifically, responsibility for *bad* outcomes.

It is widely held in the literature on moral responsibility that causation is a necessary condition for responsibility for outcomes.[5] This causal condition for responsibility is typically stated as follows:

> *The causal condition for responsibility*: you are responsible for a bad outcome *e* only if you have performed some action or omission *c*, such that *c* is a cause of *e*.

To see the intuitive support for *The causal condition for responsibility*, suppose that someone holds you responsible for some bad outcome *e* – say, the breaking of a vase. One way in which you might defend yourself is precisely by pointing out that you had nothing to do with the breaking of the vase: there is no causal connection between your behaviour and the breaking of the vase. If you succeed in showing that your behaviour did not cause the bad outcome, then it does indeed seem clear that you are not responsible for it.

The causal condition for responsibility simply requires that *c* is a cause of *e*. However, when we combine this with the idea that causation is a three-place relation with a possibility horizon as its third relatum, we face a crucial question: how do we (as a matter of describing our actual practice) select a possibility horizon for evaluating whether *The causal condition for responsibility* is satisfied?

3. How Do We Select a Possibility Horizon?

In this section, I am going to suggest that moral norms play a crucial role in our selection of a possibility horizon for evaluating whether *The causal condition for responsibility* is satisfied.

What are moral norms? As I shall use the term in the following, the content of a moral norm is simply a claim about what is morally permitted or required – for example, "you are morally required to keep your promises." Saying that a person is committed to a particular moral norm (or system of moral norms) amounts roughly to saying that she believes that these moral norms are true.

It has frequently been observed that our moral norms play a role in shaping our ordinary causal judgements, especially in contexts where we are concerned with attributing moral responsibility. Consider the following case:

[5] According to Sartorio (2007), this is the most widely held view about the relationship between causation and responsibility for outcomes. Against the causal condition for responsibility, see Sartorio (2004).

> *Flowers*: Billy has promised to water Suzy's flowers while she is away. However, he fails to do so, and the flowers die. If Billy had watered the flowers, they would have continued to bloom. It is also true that if the Queen had watered the flowers, they would have continued to bloom.[6]

In this case, it seems perfectly appropriate to assert (3) but inappropriate to assert (4):

(3) Billy's failure to water the flowers caused their death.

(4) The Queen's failure to water the flowers caused their death.

Furthermore, it seems clear that the reason we are treating Billy and the Queen differently in this case has to do with our moral norms: Billy's failure to water the flowers violated a moral norm that we are committed to (namely, the norm that "you are morally required to keep your promises"). By contrast, the Queen's failure to water the flowers did not violate any of our moral norms – according to our moral norms, she was under no obligation to look after Suzy's flowers.

The literature on causation as well as the experimental philosophy literature abounds with cases illustrating this phenomenon: when something bad happens (e.g. the flowers die),[7] we tend to treat behaviour that violates a moral norm we are committed to (e.g. Billy's failure to water the flowers as promised) as a candidate cause, while treating behaviour that is in accordance with our norms (e.g. the Queen's failure to water the flowers) as a mere background condition.[8] This suggests that norms play a crucial role in our selection of a possibility horizon: when someone violates a norm we are committed to, we take seriously the possibility that they might instead have acted as the norm requires. By contrast, when someone acts in accordance with the norms we are committed to, we do not take seriously the possibility that they might have acted differently.[9] In *Flowers*, for example, our

[6] This type of case has been widely discussed. See e.g. Hart and Honoré (1985), p. 38; Sartorio (2004); Beebee (2004); McGrath (2005); and Blanchard and Schaffer (2017). For experimental support for the verdicts I use, see Willemsen (2018) and Henne et al. (2017).

[7] The principle sketched here only seems to apply when we are asking about the causes of a bad outcome. For the sake of brevity, I omit discussion of neutral and good outcomes.

[8] For example, Henne et al. (2017) summarise their findings as follows (p. 274): "when an omission does not violate a norm (and there is counterfactual dependence) it will not be identified as a cause, and when it does violate a norm it will be identified as a cause." For an overview of the empirical literature, see Willemsen and Kirfel (2019); cf. Hitchcock and Knobe (2009).

[9] The question I am concerned with here is simply the question of which possibilities we take into consideration when we evaluate who caused a bad outcome, such as the flowers' death. It is a separate question whether an agent *could* have acted differently, in the sense that is relevant for free will. For example, it is perfectly consistent to say that the Queen *could* have acted differently (in the sense relevant for free will) – she could have made different choices about what to say, what to wear, etc.; perhaps, she could even have watered Suzy's flowers (if she had tried, she would have succeeded) – but we simply ignore these possibilities when we evaluate who is responsible for the flowers' death.

moral norms select a possibility horizon $H_{Flowers}$ that contains worlds in which Billy waters the flowers but does not contain worlds where the Queen waters the flowers. Relative to $H_{Flowers}$, it is then true that Billy's failure to water the flowers caused their death: *Candidate cause* is satisfied, and it seems clear that the remaining conditions for causation (whatever they are) are satisfied too. By contrast, it is false that the Queen's failure to water the flowers caused their death: since our possibility horizon does not include any world where the Queen waters the flowers, *Candidate cause* fails to be satisfied. This accommodates the intuitive judgement that it is appropriate to assert (3) but inappropriate to assert (4).

The way in which we evaluate causation for the purpose of attributing legal responsibility offers a clear parallel. While the relevant norms in the case of moral responsibility are moral norms, the relevant norms in the case of legal responsibility are specified by the laws and legal practice. Schaffer (2010) offers the following illustration: suppose that a lifeguard naps while he is on duty, and a swimmer drowns on his watch. Did the lifeguard's negligence cause the swimmer's death? When a judge is evaluating this question, she does not consider the closest possible world where the lifeguard does not nap – as Schaffer notes, that might be a world in which the lifeguard sneaks off to have a cigarette, and what happens in such a world seems clearly irrelevant to the question we are interested in. Rather, the judge considers the closest world in which the lifeguard lives up to what is legally required of him. Based on this and other cases, Schaffer suggests the following description of our current legal practice:[10]

> Generalizing, it seems that causal judgments in the law are based on a comparison between the actual course of events and an alternative scenario in which the defendant acts lawfully. (Schaffer, 2010, p. 272)

Similarly, Halpern and Hitchcock make the following observations about the role of legal norms in determining legal causation:

> The law suggests a variety of principles for determining the norms that are used in the evaluation of actual causation. In criminal law, norms are determined by direct legislation. For example, if there are legal standards for the strength of seat belts in an automobile, a seat belt that did not meet this standard could be judged a cause of a traffic fatality. By contrast, if a seat belt complied with the legal standard, but nonetheless broke because of the extreme forces it was subjected to during a

[10] Schaffer uses his contrastive account of causation (Schaffer, 2005) to accommodate this observation: according to Schaffer, causation is a four-place relation between a cause, a contrast to the cause, an effect, and a contrast to the effect. When we are evaluating causation in the law, Schaffer's suggestion then is that the relevant contrast to the cause is lawful conduct. However, Schaffer's more general observation – namely, that when we are evaluating causation in the law, we take seriously the possibility that the defendant might have acted in accordance with the law, but do not take seriously e.g. the possibility that the defendant might have broken the law in some other way (say, by sneaking off to get a cigarette) – may just as easily be understood as an observation about how we select a possibility horizon.

> particular accident, the fatality would be blamed on the circumstances of the accident, rather than the seat belt. In such a case, the manufacturers of the seat belt would not be guilty of criminal negligence. In contract law, compliance with the terms of a contract has the force of a norm. In tort law, actions are often judged against the standard of "the reasonable person." (Halpern and Hitchcock, 2010)

We may think of the law as guiding our selection of which possibilities we do and do not take seriously when we are evaluating legal causation: the possibility horizon we select for the purpose of evaluating causation in the law is determined, to a large extent, by the content of the law. We select a possibility horizon that includes possible worlds where the defendant acts in accordance with the law (more carefully, where the defendant does what the law minimally requires), while it leaves out possible worlds where the defendant breaks the law in some other way than he actually did, or where people who in fact acted in accordance with the law act differently.

Let us take stock. The picture that has emerged so far suggests that there are different *flavours* of responsibility: when we are concerned with moral responsibility, we assess whether *The causal condition for responsibility* is satisfied by considering causation within the possibility horizon that is selected by our moral norms (that is, the moral norms we believe to be true). When we are concerned with legal responsibility (within a particular jurisdiction), we assess whether *The causal condition for responsibility* is satisfied by considering causation within the possibility horizon that is selected by the relevant legal norms (that is, the legal norms that are contained in the laws and legal practice of the relevant jurisdiction). And there may be additional flavours of responsibility following the same pattern. For example, there might be a particular flavour of responsibility that applies to sports. Here, the relevant norms might be norms of skill – for example, the level of skill that can be expected of players in the Champions League – and a coach or sports journalist might hold a player responsible for a failure to score a goal when their actions or omissions count as causes relative to the possibility horizon selected by the relevant norms of skill.[11] In the following, I will leave these other flavours aside and focus on moral responsibility.

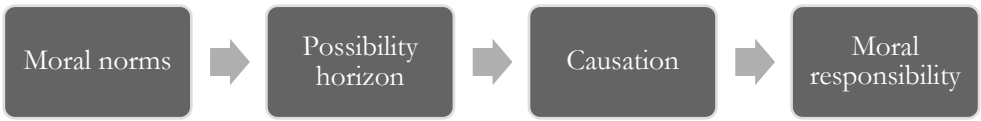
4. The Emerging Picture

We have now arrived at a picture where our moral norms play a role in how we select a possibility horizon when the purpose of our causal inquiry is to attribute moral responsibility for some bad outcome.[12] This possibility horizon in turn shapes

[11] See e.g. the general account of blame and credit in Björnsson (2017).

[12] In a context where we have a different purpose – for example, understanding the physics of how forest fires occur – it may well be the case that our moral or legal norms play no role in our selection of a possibility horizon.

our causal judgements; and these causal judgements in turn inform our attributions of responsibility. These relations between moral norms, possibility horizons, causation, and moral responsibility are summarized in the diagram below:



One way in which we may put this picture to use is in understanding disagreements: the picture suggests that disagreements about moral responsibility may be the result of underlying disagreements about causation, which are themselves the result of underlying disagreements about moral norms. More carefully, suppose that A takes S_A to be the correct system of moral norms and B takes a different system S_B to be the correct system of moral norms. In this situation, A is going to select a possibility horizon H_A on the basis of S_A and use H_A as the basis for her causal judgements and eventually her attributions of responsibility. Similarly, B is going to select a possibility horizon H_B on the basis of S_B and use H_B as the basis for her causal judgements and eventually her attributions of responsibility. Thus, A and B may reach different verdicts about causation and moral responsibility because of an underlying disagreement about whether S_A or S_B is the correct system of moral norms.[13]

This picture resembles a suggestion made by Mackie in his paper "Responsibility and language" (1955). Mackie here presents the following case:

> In Sydney some time ago a motor cyclist was exceeding the speed limit; a traffic policeman, also on a motor cycle, chased him, and soon they were both travelling, according to the reports, at 70 m.p.h. Then an unobservant citizen stepped off a bus into the policeman's path; in the crash that resulted the other man was killed at once; the policeman died the next day.
>
> There was some disagreement as to who was responsible for this accident. The police announced that when they caught the original speedster they would charge him with causing the two deaths. The general public was inclined at first to hold the policeman responsible for the other man's death, but tended to change its mind a little when he died himself. (Mackie, 1955, p. 143)

This disagreement about responsibility, Mackie suggests, is the result of an underlying disagreement about what is "the normal, proper, or expected course of events":

[13] For a nice example of how disagreements about moral norms may lead to disagreements about causation, see Kaiserman (2017), pp. 57-58.

> The answer we choose will depend on what we take to be the normal, proper, or expected course of events; the person that we hold responsible is the one who steps outside this expected pattern. Thus if we assume, as apparently the police did, that it is normal and proper for traffic police to pursue, relentlessly and with all the means in their power, those who break the speed limit, and that it is normal and proper for people to step off buses without taking precautions against motor cycles passing at 70 m.p.h., but that it is not normal and proper for cyclists (other than police in pursuit of a criminal) to break the speed limit, then we shall hold the cyclist responsible. The behaviour of the policeman and the bus traveller belonged, on this view, to the normal pattern, but that of the cyclist was an intrusion into it.
>
> On the other hand, the general public is inclined to take a less legalistic view, and not to identify what is normal and proper with strict conformity to the law and the police regulations. It might hold, therefore, that the cyclist's conduct, though illegal, was yet normal and expected, including his increase in speed when chased, whereas it was not normal, not "reasonable", for the policeman to go to such lengths to catch a speedster. Making these assumptions, the general public would conclude that the policeman was responsible for the accident. (Mackie, 1955, p. 144)

The structure of Mackie's suggestion is parallel to what I have suggested above: in both cases, the idea is that we may trace disagreements about moral responsibility back to disagreements about causation, which may in turn be traced back to an underlying disagreement about what should be treated as a mere background condition (as belonging to "the normal pattern" in Mackie's terms) and what should be treated as a candidate cause (an "intrusion" into the normal pattern). The main difference between my suggestion and Mackie's consists in the fact that Mackie traces disagreements about what should be treated as a mere background condition and what should be treated as a candidate cause to disagreements about what is "the normal, proper, or expected course of events," where what is "expected" is understood broadly, including both what is expected in a normative sense and what is statistically expected. By contrast, I trace disagreements about moral responsibility more specifically to disagreements about moral norms.

From here, there are several ways one might go. So far, I have focused on *describing* the relation between causation, responsibility, and moral norms in our practice of attributing moral responsibility for outcomes. One question one might consider at this point is whether this practice is sound. For now, I will put this question aside. In the following, I will instead consider what happens if we accept the picture I have set out above, not only as a descriptive picture of how we in fact go about attributing responsibility for outcomes, but also as a prescriptive picture of how we *should* go about attributing responsibility for outcomes. I will argue that *if* we accept that this is how we should go about attributing responsibility for outcomes, the picture I have sketched suggests a way to *evaluate* different systems of norms in terms of the attributions of responsibility they can and cannot support.

5. Evaluating Systems of Norms

The picture presented in sections 1-4 sets out how taking a particular system of moral norms as input yields particular attributions of moral responsibility as output. This imposes a requirement of consistency between our moral norms and our attributions of responsibility: the verdicts we get when we use our moral norms as input have to *support* the attributions of responsibility we make. In our everyday lives, we typically achieve consistency by taking our system of moral norms as given and then simply accepting the verdicts about moral responsibility that it delivers. However, we may also look at things the other way around: if we have some prior understanding of the verdicts about responsibility we *should* be getting – and in this section, I will suggest that there are independent reasons to prefer some verdicts over others – then we may *reverse-engineer* our system of moral norms so that it can support those verdicts. To the extent that we know what the output (attributions of responsibility) should be, the picture therefore offers a way to compare and evaluate different systems of norms in terms of their ability to support those verdicts.

Let me illustrate this with a simple example: suppose that Ben starts a camp fire in a dry forest, a gust of wind blows a spark into some nearby dry grass, the grass catches fire, and a large and destructive forest fire develops. Is Ben responsible for the forest fire? The answer depends on which norms we take as input. Consider the following two moral norms:

N_1: It is not permissible to start a camp fire in a dry forest.

N_2: It is permissible to start a camp fire in a dry forest.

Using N_1 as input, we get a possibility horizon H_1 that includes worlds where Ben does not start a camp fire. Relative to H_1, we find that Ben's action is a cause of the forest fire: *Candidate cause* is satisfied and the remaining conditions for causation (whatever they are) are clearly satisfied too. Thus, Ben satisfies *The causal condition for responsibility*. Of course, *The causal condition for responsibility* is not a sufficient condition for moral responsibility. However, provided that Ben also satisfies the remaining conditions, we get the verdict that he is responsible for the forest fire.[14]

Using N_2 as input, on the other hand, we get a possibility horizon H_2 where Ben starts a camp fire in *every* world in H_2. That is, Ben's starting a camp fire is simply

[14] What are these other conditions for being responsible for an outcome? I will not attempt a complete answer to this question here. However, at least the following requirement seems plausible: an agent is only responsible for an outcome if he is responsible for the action or omission that in turn caused the outcome. For example, Ben is only responsible for the forest fire if he is responsible for starting the camp fire: we should not hold him responsible for the forest fire if someone held a gun to his head and forced him to start the camp fire. For an overview, see e.g. Sartorio (2007).

treated as a background condition in H_2. Relative to H_2, we find that Ben's action is *not* a cause of the forest fire: *Candidate cause* fails to be satisfied. Thus, Ben does not satisfy *The causal condition for responsibility*, and it immediately follows that he is not responsible for the forest fire. If we also treat the other circumstances – the gust of wind, the proximity of the dry grass, and the dryness of the forest – as background conditions, we do not find *any* causes of the forest fire. Instead, the forest fire appears inevitable: it occurs in every world in H_2. If we instead treat it as a serious possibility that there might not have been a gust of wind, that there might not have been dry grass where the spark landed, or that the forest might not have been so dry, we do find causes of the forest fire – namely, the gust of wind, the proximity of the dry grass, and the dryness of the forest. But either way, Ben's action does not count as a cause of the forest fire, and so we get the verdict that he is not responsible.[15]

This reveals an important difference between the two norms, N_1 and N_2: N_1 allows us to hold an agent responsible for the forest fire – namely Ben, who should have refrained from starting his camp fire. By contrast, N_2 does not allow us to hold anyone responsible.

Do we have a principled reason to prefer one of these verdicts over the other? I believe we do: as a tentative suggestion, one might think of a system of moral norms as a theory for explaining bad outcomes in terms of the actions and omissions of moral agents. One such theory is superior to another when it can explain more. For example, considered as a theory for explaining bad outcomes in terms of the actions and omissions of moral agents, a system of norms that includes N_1 is, *ceteris paribus*, superior to a system of norms that includes N_2 instead: a system of norms that includes N_1 can *explain* the occurrence of the forest fire (a bad outcome) in terms of Ben's lighting a camp fire; by contrast, N_2 cannot offer any explanation of this outcome in terms of the actions and omissions of moral agents – according to N_2, the forest fire is simply an accident (either seen as inevitable or as being caused by non-agential features of the situation, such as the gust of wind, the proximity of dry grass, or the dryness of the forest).

[15] A person who is committed to N_2 may still recognize that there is a perspective from which Ben's starting the camp fire *is* a cause of the forest fire: for example, when the purpose of the causal inquiry is to understand the *physics* of how the forest fire started, moral norms play no role in the selection of a possibility horizon – and thus, even someone who is committed to N_2 may select a possibility horizon $H_{Physics}$ that treats it as a serious possibility that Ben might not have started a camp fire. The important point, however, is that, to a person who is committed to N_2, the fact that Ben's starting a camp fire is a cause of the forest fire within $H_{Physics}$ has no bearing on the question whether *The causal condition for responsibility* is satisfied: what matters for this question is whether Ben's starting the camp fire is a cause relative to the relevant *normative* possibility horizon. A different example might bring out the idea more clearly: suppose you decide to walk home one evening instead of taking a cab. On the way home, you get robbed. In this case, there obviously is a perspective from which your decision to walk home is a cause of your getting robbed – if you had taken a cab instead, you would not have been robbed. However, in a discussion about who is responsible for the robbery, it would seem entirely misplaced to point out that your decision to walk home was a cause of your getting robbed: the norm-free perspective from which your decision to walk home is a cause of the robbery is simply irrelevant when we want to determine whether *The causal condition for responsibility* is satisfied.

The suggestion that we may think of a system of moral norms as a theory for explaining bad outcomes in terms of the actions and omissions of moral agents draws on the principle of moral harmony. In *Utilitarianism and Co-operation*, Regan expresses this principle as follows:

> [T]here is the intuition that whatever the correct moral theory is, […] [i]t ought to be the case that if all agents satisfy the theory, then the class of all agents produce the best consequences they can produce collectively by any pattern of behaviour. (Regan, 1980, p. 3)

By contraposition, the principle states that if a suboptimal outcome is produced, there must be at least one agent who has failed to satisfy the moral theory, i.e. who has failed to behave as the theory requires (Pinkert, 2015, pp. 975-77). It is natural to understand this precisely as a principle about explanation: when a suboptimal outcome occurs, the correct moral theory will explain *why* it occurred in terms of one or more agents failing to behave as the theory requires.

This gives us a way to evaluate different systems of norms: a system of norms that allows us to explain a bad outcome in terms of the actions and omissions of moral agents is, *ceteris paribus*, superior to an alternative system of norms that treats the outcome as unexplainable – i.e., as something that "just happened." This is, of course, not the only relevant consideration when evaluating systems of norms: sometimes a system of norms that allows us to explain a bad outcome in terms of the actions and omissions of moral agents should still be rejected because it is deficient in other ways. However, when two systems of norms are equally good in all other respects, I believe we should prefer the one with more explanatory power.

6. Evaluating Our Norms in the Face of Climate Change

In a newspaper article about the deadly 2021 heatwave in Canada, the climate journalist Eric Holthaus wrote:

> Climate change isn't just a thing that's happening, it's a series of choices made by actual people who are sharing this planet with us. (Holthaus, 2021)

The picture I have sketched here yields an intriguing interpretation of this statement:[16]

First, there exist some systems of norms according to which climate change, and more specifically climate-change-related disasters, "just happen" – that is, these systems of norms do not allow us to *explain* these disasters in terms of the actions and omissions of moral agents. Consider, for example, the following simplified system of norms:

[16] This interpretation may, of course, go beyond what Holthaus intended; that is not the point.

Unrestricted permission to emit greenhouse gases

The norms for individuals, businesses, and governments are as follows:

a) *Individuals:* it is permissible to drive, fly, shop, eat an ordinary Western diet, and do all the other things that are part of an affluent Western lifestyle.
b) *Businesses:* within legal limits, it is permissible to produce any goods that are demanded in the market, in whichever way is most profitable.
c) *Governments:* it is permissible for members of the public to vote and prioritise political issues as they see fit, without regard for the interests of future generations; and it is permissible for governments to act in accordance with the priorities of the general public.

If we use this system of norms as input, we find that climate-change-related disasters "just happen" – they are not caused by the actions and omissions of moral agents, and thus, we cannot appropriately hold anyone responsible for them.

Second, there also exist systems of norms which would not commit us to the conclusion that climate change is "just a thing that's happening." Consider, for example, the simplified system of norms below:

Restricted permission to emit greenhouse gases

The norms for individuals, businesses, and governments are as follows:

a*) *Individuals:* it is not permissible to emit greenhouse gases for the sake of trivial benefits.
b*) *Businesses:* it is not permissible to emit greenhouse gases during the production of goods and services, when these emissions could easily be avoided.
c*) *Governments:* it is not permissible for voters and governments to neglect issues that could be devastating for future generations.

The precise content of these norms is not important. The important point is that there exist systems of norms – exemplified by *Restricted permission* – that would allow us to hold individuals, businesses, and governments responsible for climate-change-related disasters. If we were to adopt such a system of norms, we would indeed see climate change as the result of "a series of choices made by actual people who are sharing this planet with us."

By stating that "[c]limate change isn't just a thing that's happening, it's a series of choices made by actual people who are sharing this planet with us," Holthaus implicitly rejects norms such as *Unrestricted permission*, according to which climate change "just happens," and endorses norms such as *Restricted permission*, according to which climate change is indeed the result of "a series of choices made by actual people who are sharing this planet with us." The arguments I have presented support this: considered as a theory for explaining bad outcomes in terms of the actions and omissions of moral agents, a system of norms that includes *Restricted permission* is, *ceteris paribus*, superior to a system of norms that includes *Unrestricted permission*.

7. Conclusion

In this paper I have done two things. First, I have suggested a picture of the relation between moral norms, causation, and moral responsibility: our moral norms play a role in our selection of a possibility horizon for assessing whether the causal condition for moral responsibility is satisfied. Second, I have suggested that, if we assume that this picture captures how we *should* attribute responsibility for outcomes, we get a way to evaluate different systems of moral norms: when a bad outcome happens, a system of norms that allows us to *explain* this bad outcome in terms of the actions and omissions of moral agents is, *ceteris paribus*, superior to a system of norms that treats the outcome as something that "just happens." Applied to climate change, this suggests that norms such as *Restricted permission* are, *ceteris paribus*, preferable to norms such as *Unrestricted permission*.[17]

References

Beebee, Helen (2004) "Causing and nothingness" in J. Collins, N. Hall, & L. A. Paul (Eds.) *Causation and counterfactuals* (291-308). Cambridge, Mass.: MIT Press.

Björnsson, Gunnar (2017) "Explaining (away) the epistemic condition on moral responsibility" in P. Robichaud, & J. W. Wieland (Eds.) *Responsibility: the epistemic condition* (146-62). Oxford: Oxford University Press.

Blanchard, Thomas, & Jonathan Schaffer (2017) "Cause without default" in H. Beebee, C. Hitchcock and H. Price (Eds.) *Making a difference: essays on the philosophy of causation* (175-214). Oxford: Oxford University Press.

Halpern, Joseph Y., & Judea Pearl (2005) "Causes and explanations: a structural-model approach. Part I: causes". *British Journal for the Philosophy of Science*, *56*(4): 843-887.

Halpern, Joseph Y., & Christopher Hitchcock (2010) "Actual causation and the art of modelling" in R. Dechter, H. Geffner, & J. Halpern (Eds.) *Causality, probability, and heuristics: a tribute to Judea Pearl* (383-406). London: College Publications.

Hart, H. L. A., & Tony Honoré (1985) *Causation in the law*. Oxford: Clarendon Press.

Henne, Paul, Ángel Pinillos, & Felipe De Brigard (2017) "Cause by omission and norm: not watering plants". *Australasian Journal of Philosophy*, *95*(2): 270-283.

Hitchcock, Christopher, & Joshua Knobe (2009) "Cause and norm". *Journal of Philosophy*, *106*(11): 587-612.

[17] I am grateful to audiences at the Lund Higher Seminar in Philosophy, the Umeå Higher Seminar in Philosophy, the "Causation and Responsibility" workshop at Bern University, and the Stockholm June Workshop in Philosophy. I would especially like to thank Per Algander, Erik Carlson, Anton Emilsson, Mattias Gunnemyr, Madeleine Hayenhjelm, Sofia Jeppsson, Alex Kaiserman, Christian Löw, Luise Mirow, Matt Talbert, Bram Vaassen, and Martín Abreu Zavaleta for their detailed comments.

Holthaus, Eric (30th June 2021) "How did a small town in Canada become one of the hottest places on Earth?" *The Guardian*. https://www.theguardian.com/commentisfree/2021/jun/30/lytton-hottest-places-world-climate-emergency

Kaiserman, Alex (2017) "Necessary connections in context". *Erkenntnis*, *82*(1): 45-64.

Mackie, John L. (1955) "Responsibility and language". *Australasian Journal of Philosophy*, *33*(3): 143-159.

McGrath, Sarah (2005) "Causation by omission: a dilemma". *Philosophical Studies*, *123*(1-2): 125-148.

McKenna, Michael (2012) *Conversation and responsibility*. Oxford: Oxford University Press.

Pinkert, Felix (2015) "What if I cannot make a difference (and know it)". *Ethics*, *125*(4): 971-998.

Putnam, Hilary (1982) "Why there isn't a ready-made world". *Synthese*, *51*(2): 141-167.

Regan, Donald (1980) *Utilitarianism and co-operation*. Oxford: Clarendon Press.

Sartorio, Carolina (2004) "How to be responsible for something without causing it". *Philosophical Perspectives*, *18*(1): 315-36.

Sartorio, Carolina (2007) "Causation and responsibility". *Philosophy Compass*, *2*(5): 749-765.

Schaffer, Jonathan (2005) "Contrastive causation". *The Philosophical Review*, *114*(3): 327-58.

Schaffer, Jonathan (2010) "Contrastive causation in the law". *Legal Theory*, *16*(4): 259-97.

Schaffer, Jonathan (2012) "Causal contextualism" in M. Blaauw (Ed.) *Contrastivism in philosophy* (35-63). New York: Routledge.

Touborg, Caroline (2018) *The dual nature of causation: two necessary and jointly sufficient conditions* (Doctoral thesis). University of St Andrews. https://research-repository.st-andrews.ac.uk/handle/10023/16561

Willemsen, Pascale (2018) "Omissions and expectations: a new approach to the things we failed to do". *Synthese*, *195*(4): 1587-1614.

Willemsen, Pascale, & Lara Kirfel (2019) "Recent empirical work on the relationship between causal judgements and norms". *Philosophy Compass*, *14*(1): e12562.

Woodward, James (2003) *Making things happen: a theory of causal explanation*. Oxford: Oxford University Press.

The Truth about Social Entities

Tobias Hansson Wahlberg

1. Introduction

There is much ado these days about a sub-field of metaphysics called 'social ontology'. According to its SEP-entry, '[social ontology] is concerned with analysing the various entities in the world that arise from social interaction' (Epstein 2021). Examples of the putative entities under study are social properties such as *being a dollar bill* and *being Prime Minister*, and social objects such as corporations and social groups. In earlier writings I have argued in some detail on a case-by-case basis that, although there certainly are *truths* about such objects and features, there is little reason to suppose that there are such entities 'out there in the world', in a substantive ontic sense. In this paper, I will bring together the main ideas and claims of these papers, to provide an overview of the position I defend.[1] If my reasoning is on the right track, there is in fact no domain of social entities for social ontologists to quantify over, using objectual or referential (first and second order) singular quantifiers. Nevertheless, existential claims about such entities can very well be true, provided that they are understood in terms of substitutional quantifiers.

The structure of the paper is as follows. I begin by briefly characterising truthmaker theory (Sect. 2) which my approach to social entities relies on. Following this, I address, first, the ontic status of social properties (Sect. 3), and then the ontic status of social objects (Sect. 4), deploying the truthmaker framework canvassed in Sect. 2. I then develop and clarify my view by introducing and discussing the old (but nowadays little-attended to) distinction between objectual and substitutional

[1] These ideas have materialised while I have been a member of the *Metaphysics and Collectivity* research group, founded by associate professor Björn Petersson and others. I am very grateful to Björn for many fruitful discussions over the years. Björn's stimulating 2007-paper in *The Journal of Philosophy* was in fact one of my entry points to social ontology, and specifically to the issue of the causal standing of social objects.

quantifiers (Sect. 5). Subsequently (Sect. 6), I address a potential problem for my approach: the fact that social entities are, in some sense, causal. I end with some concluding remarks in Sect. 7.

2. Truthmaker Theory

In my theorising about 'social reality', I set off from the idea that contingent truths in general have *truthmakers*: entities (objects, properties, states of affairs, events) in the world that make true sentences/statements/propositions true – i.e., entities in virtue of which the truths are true (for general discussion of truthmaker theory, see e.g. Heil 2003; Armstrong 2004; Cameron 2008; Mellor 2009/2012). Importantly, truthmakers need not exactly mirror or correspond to the content of the truths in question. Truthmakers can very well be, in David Armstrong's terminology (2004: 33), 'deflationary'. Such truthmakers typically do not, at first sight, look 'fully dressed up' for the occasion, but nevertheless they suffice to make the relevant statement true. Here are some examples from the metaphysical literature that illustrate the notion of deflationary truthmaking:

- Tensed statements made true by B-facts (e.g. Mellor 1998).
- Dispositional statements made true by categorical properties plus laws of nature (e.g. Armstrong 1997).
- Statements about rainbows made true by sunlight-reflecting raindrops (e.g. Mellor 2009/2012).
- Statements about macroscopical objects and properties made true by fundamental particles arranged X-wise (e.g. Heil 2003; Cameron 2008).[2]

Note that it does not follow from the fact that a statement has deflationary truthmakers that the putative state of affairs expressed by that statement has been 'reduced' to the relevant deflationary truthmakers. Reduction is typically understood in terms of identification (see my 2019a for detailed discussion). But deflationary truthmakers cannot, in general, be taken to be *identical* with the putative states of affairs expressed by the relevant statements: for example, states of

[2] To fill in some more detail: Truthmaker B-theorists will say that a true utterance of the tensed 'I ran yesterday' is made true by the B-fact – or 'tenseless' state of affairs – that the utterance is located one day after the day on which the utterer (or a temporal counterpart of her) runs. Truthmaker categoricalists will say that a true utterance of the dispositional 'This substance is corrosive' is made true by the categorical properties of the substance plus the obtaining laws of nature. Deflationists about rainbows (as we may call them) will say that a true utterance of 'There is a rainbow east of us' is made true by sun-light reflecting raindrops east of the persons in question. And truthmaker deflationists about macro-scopic objects and properties will say that a true utterance of 'This brick is rectangular' is made true by fundamental particles arranged in certain complex ways ('rectangular-brick-wise').

affairs *not* containing any A-properties (i.e., properties such as being past, being present and being future), so-called B-facts, cannot be identified with putative states of affairs that contain such properties (so-called A-facts). B-theorists in the philosophy of time instead typically claim that, in an ontic, worldly sense, *there are no* A-facts – although there certainly are *true tensed statements* (which are made true by B-facts).

Although the adoption of a (deflationary) truthmaker theory is quite popular in general metaphysics, surprisingly, very few social ontologists (if any) have invoked such a theory in relation to social ontology. A deflationary truthmaker approach is, I think, particularly suitable in relation to putative social entities – as I will try to explain and illustrate in the following sections.

3. Social Properties

I begin by applying a deflationary truthmaker approach to (putative) social properties. An important sub-category of the social properties, often focused on by social ontologists, is the category of institutional properties (e.g. Searle 1995; 2010). These are properties or 'statuses' which depend for their 'existence' on our acceptance of constitutive rules ('institutions'), which have the illocutionary force of *declarations*. The relevant constitutive rules are standardly taken to be of the form 'X counts as Y in context C', where the X term picks out an object (or a kind of object) and the Y term expresses a property – such as *being a dollar bill*, *being President*, *having grade G in subject S* – which is simply *assigned* to X, and which consequently is not reducible to any brute, physical properties of X. Below, I first briefly discuss institutional properties, and then I go on to address non-institutional social properties. I propose a deflationary truthmaker account of both kinds of property.

3.1 Institutional properties

Pace Searle (e.g. 2010: 11-12), I do not think we should conceive of the assignment of an institutional property to an object (or to a collection of objects of a certain kind) as involving *creation* of something worldly: a token of an ontic property, which was not instantiated by the relevant object (or objects) prior to the assignment. The notion that something is literally created by such an assignment suggests peculiar action at a distance or even magic: it is hard to see that there could be a naturalistic mechanism at work (Effingham 2010; my 2019b; 2021).[3] Such a

[3] Admittedly, Searle maintains that such properties are 'ontologically subjective' (e.g. Searle 2010: 18). However, it is unclear what this alleged mode of existence amounts to and how claims of ontological subjectivity cohere with statements such as these: '[Declarations] change the world by declaring that a state of affairs exists and thus bringing that state of affairs into existence' (Searle 2010:

mechanism would in any case violate the special theory of relativity, since the mechanism would involve, in standard cases, *instantaneous* generation of the institutional property (in the reference frame of the assigner) (see my 2021 for detailed discussion). Even more troublesome, sometimes we assign institutional properties, as it were, *backwards in time*. For example, at universities, grades and appointments are regularly assigned retroactively. A standard scenario: a student completes an assignment on a certain date *t*, but a busy professor grades the assignment at a later date *t′* and in so doing gives the student an official grade which is valid from the *earlier* time *t* onwards. If something is literally created by such retroactive assignments, they will involve backwards generation.[4] In fact, in some reference frames (moving at high velocity relative to the reference frame in which the assignment takes place), ordinary *synchronic* assignments will, from their points of view, involve backwards generation, if such assignments involve creation of ontic properties (this is illustrated in detail in my 2021). All of this suggests, I think, that assignments of institutional properties do *not* involve creation of worldly, ontic properties.

Fortunately, a deflationary truthmaker theory enables us to explain what goes on in these cases, without us having to postulate institutional properties as ontic entities. In a nutshell, the account is this: When we accept a constitutive rule or declaration of the form 'X counts as Y (at time *t*)', that acceptance makes it *true* that X is Y (at time *t*). Thus, the truthmakers for statements about X's being Y (at time *t*) are simply these acceptances or states of mind (which may be 'located' at times differing from *t*) – the truthmakers do not involve an institutional property, Y, instantiated by X (at time *t*). The relevant institutional *predicate* applies to the object in question, but this is not because the object has started to instantiate an ontic property; rather, the predicate applies in virtue of the collective *acceptance* of the constitutive rule or declaration in question. Thus, the application of the institutional predicate is simply an instance of 'mere Cambridge change' (cf. Geach 1969: 71-72): a predicate begins to apply to an object at a certain location because of a physical or mental change that happens elsewhere, even at another time (for detailed discussion, see my 2021).[5]

The resulting view of putative institutional properties can thus be regarded as a form of *predicate nominalism* (cf., e.g., Armstrong 1978: 12-14). On this view, an

12); 'the whole point of having the notion of "fact" [or state of affairs] is to have a notion for that which stands outside the statement but which makes it true, or in virtue of which it is true, if it is true' (Searle 1995: 211). For extensive critical discussion, see my (2021).

[4] Some social ontologists apparently happily embrace this consequence; see e.g. Silver (2022).

[5] Dan Sperber has informed me that he proposed a similar 'mere-Cambridge-change' account of institutional properties already in 2011. See his *Seventh European Congress of Analytic Philosophy* (ECAP7) lecture in Milano entitled 'The Deconstruction of Social Unreality' (unpublished in written form), available online at: https://vimeo.com/28924148. I was completely unaware of this talk when I wrote my (2021), but it is certainly exciting and encouraging that our analyses converge in this way: as Sperber put it (personal communication), this convergence may be taken as 'indirect evidence that we may well be on the right track'.

object can be said to have a 'property' simply in virtue of a suitable predicate applying to the object in question. Another way of putting the idea is to say that institutional properties are merely so-called *abundant properties*, not sparse properties, in the terminology introduced by David Lewis (see his 1986: 59-60; for discussions of the sparse/abundant distinction, see my 2021 and 2022). That is, institutional properties are not immanent universals or tropes, but should be understood merely in terms of true predications.

3.2 Non-institutional Social Properties

What about social properties which are not *assigned* to objects but which, allegedly, somehow *emerge* due to social interactions (see e.g. Bunge 1996; Elder-Vass 2010; Lawson 2013; 2016). Putative examples of such properties include *being able to arrest suspects* (a property held by police officers), *being able to dismiss employees* (a property held by corporations), and *being able to influence the normative beliefs and behaviour of individual persons* (a property held by social groups) (ibid.).[6] Do such properties exist in an ontic sense? In order to address this issue, we need first to distinguish between (supposedly) non-institutional social properties had by *individuals* (or, possibly, by physical objects) and (purportedly) non-institutional social properties had by *social objects*, such as corporations and social groups. Let me begin by addressing the former properties.

Social properties had by individuals are clearly *extrinsic*, even if they are not institutional: they are 'properties' that individuals have because they stand in various 'relations' to other individuals (or to 'social objects' of which they are 'part' or 'related' to). A physical duplicate of an individual with such 'properties', existing in social isolation on a remote planet, would not have these properties – which is why it makes sense to call them *social* (Lawson 2013; my 2020). Should we, then, think of non-institutional social properties as ontic (or, in Lewis's terminology, as sparse)? Elder-Vass, Lawson and other critical realists say *yes*, because they hold that such properties are so-called *powers*, i.e. ontic causal properties which have their causal profile essentially. The fatal problem with this proposal is that powers are – in the general metaphysical literature – supposed to be *intrinsic* features of things (see, e.g, Harré 1970; Molnar 2003; Bird 2007). Non-institutional social properties had by individuals can thus only be conceptualised as powers on pain of contradiction (see my 2020 for extensive discussion). Being contradictory, we can

[6] I think many of the examples referred to in relevant literature are, in fact, institutional. For example, a person typically satisfies 'is a police officer' by way of a collectively accepted declaration (such as a signed diploma), and thereby the person also typically satisfies predicates such as 'has the right to arrest suspects' (which is a mere deontic-power predicate – it is not, as such, a causal predicate); see my 2020 for detailed discussion. Here, however, I proceed on the assumption that the relevant social properties are not institutional, for the sake of the argument.

conclude that non-institutional social properties *qua* powers (had by individuals) do not exist in an ontic sense.

But suppose believers in non-institutional social properties denied that such properties are powers and merely maintained that they are ontic, extrinsic properties that individuals have because they stand in various ontic relations to other individuals (or social objects). Is such a view tenable? One objection to such a position is that it seems to be in conflict with Ockham's Razor: the extrinsic properties in question seem to be ontologically superfluous.

Consider a purely spatial example: if we postulate two individuals *a* and *b*, and an ontic dyadic relation of *being spatially separated by 10 m* which is jointly instantiated by *a* and *b*, is it not then redundant to also postulate an ontic, extrinsic monadic property, *being separated from a by 10 m*, which is instantiated by *b*, and a corresponding ontic, extrinsic monadic property, *being separated from b by 10 m*, which is instantiated by *a*? If we have the ontic dyadic relation (jointly instantiated by *a* and *b*), it seems we already have all we need to explain why the monadic predicate 'is separated from *a* by 10 m' is true of *b,* and why the monadic predicate 'is separated from *b* by 10 m' is true of *a*. These predicates apply to *b* and *a*, respectively, because of the ontic *dyadic relation* that is jointly instantiated by *a* and *b*. Likewise for the full sentences '*b* is separated from *a* by 10 m' and '*a* is separated from *b* by 10 m': a deflationary truthmaker theorist will maintain that both sentences are made true by the 'deflationary' state of affairs that *a* and *b* are separated by 10 m; there is no reason to postulate two distinct 'inflationary' states of affairs here: *b's* having the ontic, extrinsic property of being separated from *a* by 10 m (which makes the first sentence true), and *a's* having the ontic, extrinsic property of being separated from *b* by 10 m (which makes the second sentence true). To postulate such inflationary states of affairs would be to violate Ockham's razor.

I suggest that a deflationary truthmaker theorist should respond similarly with respect to putative non-institutional social properties had by individuals. It is not immediately obvious what exactly the relevant underlying ontic relations are supposed to be in these cases,[7] but the general strategy for deflationary truthmaker theorists is clear (given that suitable ontic relations can be identified): maintain that it is true to *say* that individual *i* is *F* (where '*F*' is a non-institutional social predicate), but hold that the truthmakers for the relevant truth do not involve any ontic, extrinsic non-institutional social properties but merely the relevant ontic relations that hold among the relevant individuals. (Alternatively – if it is hard to find plausible ontic relations here – maintain that the truths are made true by the mental attitudes of the individuals involved. However, if the latter route is taken,

[7] The authors in question speak of 'interactions', 'collective practices', 'relational organisation', 'organising structure', etc. Insofar as these notions are supposed to refer to deontic relations (e.g. Lawson 2016: 364-365) – or presuppose the ontic existence of social objects or wholes (of which the individuals in question are parts, components or members) – I would deny that these terms succeed in picking out genuine, ontic relations (see my 2020, and below, Sect. 4).

the deflationary truthmaker theorist is coming very close to adopting the proposal discussed in Sect. 3.1 concerning institutional properties. In the end, this may very well be the most advisable approach; cf. note 6 above; see my 2020 for further discussion.[8])

Next, consider alleged non-institutional social properties had by *social objects*, such as corporations and social groups. Such properties need not be extrinsic, but can be intrinsic to the objects in question. (Extrinsic social properties had by social objects face the same issues as those just described in relation to individuals.) Such intrinsic social properties can consistently be taken to be powers. However, the notion that there are ontic powers at the level of social objects as wholes faces a causal exclusion problem analogous to the causal exclusion problem discussed in the philosophy of mind (e.g. Kim 2005): given the causal abilities and performances of the individuals who make up the social object (in the case of a social group), or who manage and administer the social object on behalf of the social object (in the case of a corporation), the causal powers of the social object itself, as a whole, seem redundant. The postulation of such 'holistic' powers seems to entail systematic causal overdetermination, at least if they are taken to be manifested (see my 2014a, 2014b, 2020 for extensive discussion). Ockham's Razor rules that we should not postulate such redundant ontic properties.

A more fundamental problem, however, is that, arguably, social objects do not even exist, in an ontic sense – a thesis I will support in the next section. If they do not exist in an ontic sense, they cannot instantiate social properties: for to instantiate such properties, they must exist in an ontic sense.

4. Social Objects

As with social properties, a distinction can be made between *institutional* and *non-institutional* instances. I will begin by addressing institutional objects, and then I will discuss non-institutional social objects. I will propose a deflationary truthmaker account of both kinds of object.

4.1 Institutional Objects

Institutional objects are non-identical with physical/brute objects and are, allegedly, *declared* into being.[9] Searle exemplifies with corporations and non-cash money.

[8] To me, it seems quite plausible to maintain that the reason a police officer can arrest someone by uttering 'You're under arrest!' (perhaps while physically grabbing the person in question), is that such an utterance is a *declaration* conforming with the collectively accepted rights and duties of police officers (cf. note 6 and my 2020).

[9] Elsewhere (Hansson Wahlberg 2014c), I have argued that, strictly speaking, a physical object or person X who comes to satisfy an institutional *sortal* predicate Y can also be said to be an 'institutional

Additional possible examples include, I take it, universities, borders, States and laws.

Searle writes about the creation of a corporation:

> In this case we seem to have created a remarkably potent object, a limited liability corporation, so to speak out of thin air. No pre-existing object was operated on to turn it into a corporation. Rather, we simply made it the case by fiat, by Declaration, that the corporation exists. (Searle 2010: 98)[10]

As I argue in my (2021), the idea that an object is literally created in this way is misguided. It is more sensible, I suggest, to adapt the deflationary truthmaker account of institutional properties and apply it to objects. On this account, a declaration to the effect that a corporation exists is made (e.g., a signing of a certain document), and because of this declaration it becomes true to *say*, 'A corporation, founded in such and such a way, exists'. The truthmaker for such an existential assertion should not be taken to be a new, ontic, institutional object that is somehow brought into existence in the world (perhaps, at its 'institutional location', cf. Hindriks 2013: 418) simultaneously with (or perhaps even before) the declaration. That would lead to difficulties of the kind discussed above, in Sect 3.1. Rather, the deflationary truthmakers should be assumed to consist simply of the declaration itself, together with representations of the relevant legal regulation (for discussion, see my 2021).

4.2. Non-institutional Social Objects

Some (putative) objects can be called social simply because they have individuals (two or more) as *members*. Such objects need not be institutional, i.e. they need not be declared into being. Examples of non-institutional social objects, spoken of in the social sciences, are *collectives* of various sorts (e.g., crowds, audiences and mobs), *categories* (e.g., people over fifty, redheads) and, possibly, (at least some instances of) *social groups* (e.g. street bands, football teams and book clubs).[11] As characterised in standard social science textbooks on the topic (e.g. Forsyth 2019), collectives and categories are (roughly) mere collections of individuals who happen

object'. Hence, such an object/person X is *both* a physical object and an institutional object. Here, however, I reserve the term 'institutional object' for the so-called free-standing-Y-term cases (see e.g. Searle 2010: 98).

[10] Searle quotes the California Code in support of his view; similar formulations can be found in Swedish law.

[11] Some social groups do seem to be introduced via declarations. Faculty committees are arguably cases in point (cf. Epstein 2019; my ms.). Thus, perhaps we should allow that at least some social groups are institutional(-ish) objects. In any case, a deflationary truthmaker account can handle them – either along the way characterized above (4.1), regarding corporations, or along the way described below, in this section.

to be located at the same place (collectives) or who happen to share some characteristic (categories). I have not, in earlier work, written specifically about collectives and categories, but I conjecture that it would be quite straightforward to offer a deflationary truthmaker theory of such entities (or rather, of *truths* about them) simply in terms of pluralities of individuals who happen to have appropriate locations/properties.[12]

Social groups, by contrast, do seem, at least prima facie, to be less easy to account for in terms of deflationary truthmakers. As has been repeatedly pointed out by social ontologists, social groups are conceptualised as *non-extensional* entities, both in colloquial speech and in the social-scientific literature: we maintain – truly, we would like to think – that distinct social groups can have the same members. For example, a chess club and an orchestra can consist of the same members. Because of the non-extensional character of social groups, many philosophers (e.g. Uzquiano 2004; Ritchie 2013; Epstein 2015, 2019) think that statements about social groups are made true by *sui generis* entities which are irreducible to sets/sums/pluralities of individuals, but which are *constituted* or *grounded* by such entities (where constitution/grounding relations are taken to be asymmetric dependence relations distinct from *n*-adic identity relations). Such philosophers thus tend to accept a bifurcated, levelled ontology in the social realm: over and above the relevant individuals (the members of the social groups in question) there are (co-membered) ontic social groups (see the diagrams in, e.g., Sawyer 2005: 70; Elder-Vass 2010: 50; Forsyth 2019: 36). I will now argue, drawing on my (ms.), that the non-extensionality of social groups can in fact be accounted for on a deflationary truthmaker account. Thus, this feature does not force us to postulate inflationary – i.e., constituted or grounded – ontic social groups as truthmakers for truths about them.[13]

To start with, on a deflationary truthmaker account it can be true to *say* that a certain social group (a street band, a book club) has been formed/created, at a certain time *t*, simply because some individuals have started to behave in a certain way (playing music in a coordinated way on a street) or entered certain states of mind (started to think of each other as members of a common book club) at *t*. The statement in question need not be made true by an ontic social group which popped into being in the external world at *t*. Now, if it is true to say of a further 'social group' that 'it' was created at time t' ($t \neq t'$), then the first group and the second

[12] To allow for membership change – if collectives and categories are conceptualised as being able to change members, which is a bit unclear – the truthmakers may more specifically be taken to involve *distinct* individuals (or temporal parts or stages of distinct individuals) at distinct times. The relevant, distinct pluralities may in effect be said to be successive 'temporal parts' or 'stages' of the 'collective' / 'category' in question. See my (2014c) and (2019a) for discussion of temporal parts and stages of institutional objects when the latter are understood as ontic entities.

[13] In my (ms.), I handle additional allegedly problematic features of social groups in terms of deflationary truthmakers.

group cannot consistently be held to be identical even if they 'consist' of the same members (i.e., even if the truthmakers for the claims in question involve the same individuals). The 'groups' will have distinct 'properties' – i.e., there will be distinct truths about them, e.g., about when they were 'created' – and hence they cannot be identified, on pain of violating Leibniz's Law. Moreover, social groups that were created simultaneously, can still – indeed, must – be distinguished if it is true to say that 'they' are governed by distinct rules or norms. The relevant deflationary truthmakers here may simply be external documents, or people's attitudes or dispositions (for further discussion, see my ms.). Thus, the fact that social groups are conceptualised, and truly described, as non-extensional does not force us to recognise them as worldly, ontic entities, over and above individuals acting under distinct rules or norms.

5. 'Existence' in Ontic and Non-ontic Senses: Objectual and Substitutional Quantification

When I say that social objects such as corporations and social groups do not exist in an ontic sense, what exactly do I mean by that? What I mean is that the existential, singular quantifier, ∃, when understood in the standard *objectual* or *referential* sense (e.g. Quine 1948/1953), does not succeed in ranging over any such objects.[14] Thus, if a true ordinary language statement, such as 'A book club has now been formed', made at time t_1, is regimented as '$(\exists x)\ (Fx \wedge Lxt_1 \wedge \neg Lxt_0)$' (where F = _is a book club with such and such features, and L = _is located at time_, and t_0 is an arbitrary time before t_1), and the existential quantifier is read as an objectual quantifier, then, on my view, the regimented version expresses a *falsehood*. However, if a *substitutional* interpretation of the existential quantifier is adopted (e.g. Marcus 1972/1993; Kripke 1976; Haack 1978, Ch. 4; in which case the symbol 'Σ' is often used), the formalised version *does* express a truth – assuming that there is a *true substitution instance* of the form '$Fa \wedge Lat_1$ & $\neg Lat_0$', as the existential quantifier, on this reading, says that there is.[15] The *truthmakers* for such a substitution instance are, I suggest, simply of the kind described above (4.2) – they are individuals

[14] The *plural* existential quantifier, in sentences such as '$\exists xx\ (Sxx, a)$', may indeed succeed in ranging over pluralities of individuals, even if it is read referentially (e.g., when formalising ordinary language sentences such as 'Some individuals surround object *a*'). However, such 'social objects' are *plural*, not singular, entities. Some people may hold it is a misuse of the term to speak of mere pluralities (i.e., several entities) as 'objects' or 'entities'. For relevant discussion, see Oliver and Smiley (2016), especially (Ch. 15).

[15] The formal language in question is assumed to have a suitable stock of names – but this assumption can be relaxed by merely requiring that we *could* have introduced a suitable name which would have allowed us to state or form a true substitutions instance of the kind just described.

thinking of each other as members of a book club (see my ms. for further discussion).

Similarly, when I say that there are no social properties in an ontic (or sparse) sense, I mean the following: if '∃' is understood as a *referential* second-order quantifier it will not succeed in ranging over any social properties; formal sentences beginning '(∃*F*) …' (where '*F'* is supposed to be a predicate variable ranging over social properties) will consequently be *false*. But since there can be *true substitution instances* for sentences beginning '(∃*F*) …', that involve social *predicates*, '∃' can figure as a *substitutional* second-order quantifier in true second-order existential social sentences (e.g., of the form '(∃*F*) *Fa*', saying, in effect, that there is a social *truth* concerning *a* – which is the case if, e.g., *'a* is president' is true, which it is if people accept the relevant constitutive rules, as outlined above, in Sect. 3.1).

Thus, I suggest that when we, in ordinary language, say that there are social objects such as corporations and social groups, and social properties such as the property (or 'status') of being money and the property (or 'status') of being married, we should be taken to be implicitly using substitutional quantifiers.

I should perhaps highlight that my invocation of substitutional quantifiers sets me apart from standard truthmaker theorists. To my knowledge, truthmaker theorists do not make use of the distinction between objectual and substitutional quantifiers. I think, however, that this distinction helps to clarify how there can be existential truths about entities that do not 'really' exist (as, e.g., Cameron 2008 puts it).

6. The Causal Impact of Social Entities

As canvassed above, my view is that there are no social entities (objects and properties) in an ontic sense, although there are truths concerning such entities. This position seems to face an immediate problem though: in colloquial speech and in the social sciences, we speak of social entities as causes and effects; but to be causes and effects, social entities must, apparently, be real. How else could they be related by causal relations?

My answer: yes indeed, there are causal *truths* involving institutional entities as relata, but such truths need not be made true by ontic social entities standing in ontic causal relations. Compare: there are causal truths involving *absences* as putative relata, and such truths are evidently not made true (partly) by absences standing in ontic causal relations. Absences are *nothings* and simply cannot serve as ontic relata. Nevertheless, it can be true to *say* 'the gardener's failure to water the flowers caused them to wither'. For example, if such a statement is analysed, in line with Lewis (2004), in terms of a pair of true statements ('the gardener fails to water the flowers at t_1', and 'the flowers wither at t_2'), which are such that had the first statement been false, the second statement would have been false, we can see how the original causal statement can be *true* even though there are no absences in an ontic sense.

(Other accounts can be adopted here, such as Mellor's (1995) or Woodward's (2003).)

Likewise, I suggest, for causal truths about social entities. If we analyse 'Joe's bad grades caused him to be unemployed' in terms of the two statements 'Joe has bad grades at t_1' and 'Joe is unemployed at t_2', which are such that they are both true, but had the first statement been false, the second statement would have been false, we can begin to see how the initial, explicitly causal sentence may be true although there are no social properties in an ontic sense. (Again, various accounts of causal statements may be adopted here.) In a slogan, my view is that true causal statements about social entities express 'mere abundant causation' (for detailed discussion of this notion, see my 2022). That is, such statements are not made true by ontic causal relations (generative processes, to be more precise, which would be instances of 'sparse causation') which connect the putative 'relata' in question. Nevertheless, they are true. The development of a detailed account of the relevant deflationary truthmakers for such truths (an account which avoids committing to non-actual possible worlds) is currently work-in-progress. The general idea, however, is that the truthmakers consist, at least partly, in people's *representations* of social entities, and the way these representations (or their physical substrates/realisers) affect people's decision making (for some preliminary discussions, see my 2014a; 2014b; 2020; 2021; 2022).

Lastly, some words about the Eleatic Principle (EP) – roughly, that to be is to be causal. EP is endorsed by many metaphysicians and social ontologists. However, I think it needs to be restricted or specified in order to be acceptable. As we saw above, absences can truly be said to be causal, and arguably, likewise for social entities. But absences and social entities do not exist in an ontic sense (this should be completely uncontroversial for absences). Thus, I suggest that EP should be understood as saying: to be in an ontic sense is to be sparsely causal. In this version of EP, my view of social 'entities' as not being ontic but as partaking in mere abundant causation is fully compatible with the principle (see my 2022 and ms. for further discussion).[16]

7. Conclusion

I have argued that there are no social entities in an ontic sense.[17] If I am correct about this, there is in fact no social *ontology* in the sense of a domain of (singular) social objects such as corporations and social groups, and social properties or

[16] Perhaps EP has to be rejected in any case: this may be so if we have to accept Platonic entities, such as numbers, in our ontology (cf. Colyvan 1998).

[17] With the exception of pluralities of individuals – if pluralities of individuals are properly referred to as 'social objects' or 'social entities', albeit plural ones. See note 14 above.

'statuses' such as being money and being a professor. In this purely extensional sense of 'social ontology', there has, in my view, been 'much ado about nothing' over the last few decades. Of course, this is not to suggest that the *subject* or *discipline* social ontology (understood as a sub-field of metaphysics) is otiose – for example, we still have to figure out what exactly the relevant truthmakers are for the various social truths in question, and these are issues which no doubt are very complex, difficult and important. Thus, I end by citing Donald Davidson, who reportedly said in relation to another philosophical topic: 'It's good to know we shan't run out of work'.[18]

References

Armstrong D. M., 1978, *Universals & Scientific Realism, Vol. 1: Nominalism & Realism*, Cambridge: Cambridge University Press.

Armstrong D. M., 1997, *A World of States of Affairs*, Cambridge: Cambridge University Press.

Armstrong D. M., 2004, *Truth and Truthmakers*, Cambridge: Cambridge University Press.

Bird A., 2007, *Nature's Metaphysics: Laws and Properties*, Oxford: Oxford University Press.

Bunge M., 1996, *Finding Philosophy in Social Science*, New Haven, CT: Yale University Press.

Cameron R, 2008, 'Truthmakers and Ontological Commitment: or how to deal with Complex Objects and Mathematical Ontology without getting into Trouble', *Philosophical Studies*, 140(1): 1-18.

Colyvan M., 1998, 'Can the Eleatic Principle be Justified?', *Canadian Journal of Philosophy*, 28(3): 313-335.

Effingham N., 2010, 'The Metaphysics of Groups', *Philosophical Studies*, 149(2): 251-267.

Elder-Vass D., 2010, *The Causal Power of Social Structures*, Cambridge: Cambridge University Press.

Epstein B., 2015, *The Ant Trap: Rebuilding the Foundations of the Social Sciences*, Oxford: Oxford University Press.

Epstein B., 2019, 'What are Social Groups? Their Metaphysics and how to Classify them', *Synthese*, 196(12): 4899-4932.

Epstein B., 2021, 'Social Ontology', *The Stanford Encyclopedia of Philosophy* (Winter 2021 Edition), Edward N. Zalta (Ed.), https://plato.stanford.edu/entries/social-ontology/

[18] See Haack (1978: 121). I have been unable, however, to track down the formulation in Davidson's own work.

Forsyth D. R., 2019, *Group Dynamics*, 7th edition, Boston: Cengage.

Geach P., 1969, *God and the Soul*, London: Routledge.

Haack S., 1978, *Philosophy of Logics*, Cambridge: Cambridge University Press.

Hansson Wahlberg T., 2014a, 'Elder-Vass on the Causal Power of Social Structures', *Philosophy of the Social Sciences*, 44(6): 774-791.

Hansson Wahlberg T., 2014b, 'Causally Redundant Social Objects: Rejoinder to Elder-Vass', *Philosophy of the Social Sciences*, 44(6): 798-809.

Hansson Wahlberg T., 2014c, 'Institutional Objects, Reductionism and Theories of Persistence', *dialectica*, 68(4): 525-562.

Hansson Wahlberg T., 2019a, 'Why the Social Sciences are Irreducible', *Synthese*, 196(12): 4961-4987.

Hansson Wahlberg T., 2019b, 'Are there any Institutional Facts?', in Hansson Wahlberg T. and Stenwall R. (eds.), *Maurinian Truths*, pp. 83-88, Lund: Media-Tryck.

Hansson Wahlberg T., 2020, 'Causal Powers and Social Ontology', *Synthese*, 197(3): 1357-1377.

Hansson Wahlberg T., 2021, 'The Creation of Institutional Reality, Special Theory of Relativity, and Mere Cambridge Change', *Synthese*, 198(6): 5835-5860.

Hansson Wahlberg T., 2022, 'Sparse Causation and Mere Abundant Causation', *Philosophical Studies*, 179(11): 3259-3280.

Hansson Wahlberg, ms., 'Towards a Deflationary Truthmakers Account of Social Groups', manuscript under review.

Harré R., 1970, 'Powers', *The British Journal for the Philosophy of Science*, 21(1): 81-101.

Heil J., 2003, *From an Ontological Point of View*, Oxford: Oxford University Press.

Hindriks F., 2013, 'The Location Problem in Social Ontology', *Synthese*, 190(3): 413-437.

Kim J., 2005, *Physicalism, or something near enough*, Princeton N.J.: Princeton University Press.

Kripke S., 1976, 'Is There a Problem about Substitutional Quantification?', in Evans G. and McDowell J. (eds.), *Truth and Meaning: Essays in Semantics*, Oxford: Oxford University Press.

Lawson T., 2013, 'Emergence and Social Causation', in Groff R. and Greco G. (eds.), *Powers and Capacities in Philosophy: The New Aristotelianism*, 2013, New York: Routledge.

Lawson T., 2016, 'Comparing Conceptions of Social Ontology: Emergent Social Entities and/or Institutional Facts?', *Journal for the Theory of Social Behaviour*, 64(4): 359-399.

Lewis D., 1986, *On the Plurality of Worlds*, Oxford: Blackwell Publishing.

Lewis D., 2004, 'Causation as Influence', in Collins et al. (eds.), *Causation and Counterfactuals*, pp. 75-106, London: The MIT Press.

Marcus R. B., 1972/1993, 'Quantification and Ontology', in *Modalities: Philosophical Essays*, 1993, pp. 75-87, Oxford: Oxford University Press.

Mellor D. H., 1995, *The Facts of Causation*, London: Routledge.

Mellor D. H., 1998, *Real Time II*, London: Routledge.

Mellor D. H., 2009/2012, 'Truthmakers for What?', in *Mind, Meaning, and Reality: Essays in Philosophy*, pp. 96-112, 2012, Oxford: Oxford University Press.

Molnar G., 2003, *Powers: A Study in Metaphysics*, Oxford: Oxford University Press.

Oliver A. and Smiley T., 2016, *Plural Logic*, 2nd edition, Oxford: Oxford University Press.

Petersson B., 2007, 'Collectivity and Circularity', *The Journal of Philosophy*, 104(3): 138-156.

Quine, 1948/1953, 'On What There Is', in *From a Logical Point of View*, Cambridge, MA: Harvard University Press.

Ritchie K, 2013, 'What are Groups?', *Philosophical Studies*, 166(2): 257-272.

Sawyer R. K., 2005, *Social Emergence: Societies as Complex Systems*, Cambridge: Cambridge University Press.

Searle J. R., 1995, *The Construction of Social Reality*, London: Penguin Books.

Searle J. R., 2010, *Making the Social World: The Structure of Human Civilization*, Oxford: Oxford University Press.

Silver K., 2022, 'Backwards Causation in Social Institutions', *Erkenntnis*, Online First, https://doi.org/10.1007/s10670-022-00613-y

Uzquiano G., 2004, 'The Supreme Court Justices: A Metaphysical Puzzle', *Noûs*, 38(1): 135-153.

Woodward J., 2003, *Making Things Happen – A Theory of Causal Explanation*, Oxford: Oxford University Press

Love, Blame, and What We are Owed

Understanding Relational Values

Jakob Werkmäster

1. Introduction

In several of his works, Toni Rønnow-Rasmussen has made the case for distinguishing between personal value and impersonal value (Rønnow-Rasmussen, 2009; 2011; 2022). In Rønnow-Rasmussian fashion the aim of this paper is to make a further distinction in value, that between relational value and non-relational value. My goal is to argue that not only are these value concepts distinct from the distinction between impersonal value and personal value; they are orthogonal to one another. As I show in the paper, I believe that some of the most important values in our everyday life, such as lovable and blameworthiness, are best understood as relational values.

The structure of the paper is as follows. In this first introductory section, I present the Fitting Attitudes Analysis of Value (henceforth FA) and how Rønnow-Rasmussen makes use of it to distinguish between personal- and impersonal value. In the second section, I show that FA offers conceptual space for another distinction in value, modifying for whom an attitude is fitting. I stipulate that we call this a distinction between relational values and non-relational values and present an initial characterization. In the third and fourth sections, I show how the distinction between relational value and non-relational value resonates with our everyday evaluative thinking and is philosophically illuminating when it comes to discussions about morality and directed moral obligations. In doing so I also conclude that our initial characterization is inadequate. Just focusing on for whom an attitude is fitting is inadequate. It fails to capture vital aspects of the phenomenon we are after, collapses into the controversial distinction between agent-relative and agent-neutral value, and inherits the problems facing said distinction. A second improved

characterization is given that not only focuses on *for whom* an attitude is fitting, but also on which attitude is fitting. This characterization is found more satisfactory albeit not without its own set of challenges.

According to FA, to be valuable is to be the fitting target of a valenced attitude. To be good is to be the fitting target of a pro-attitude and to be bad is to be the fitting target of a con-attitude. This pattern of analysis goes back to Brentano (1889/2009) and Ewing (1948), and has had a renaissance in contemporary philosophy following Scanlon (1998) but also in large part due to the contribution of Rønnow-Rasmussen and Rabinowicz (2004). FA consists of two components, a normative component (i.e., the 'fittingness') and an attitudinal component (i.e., the 'pro/con-attitude'). On FA, to be lovable is to be the fitting object of love, admirable the fitting target of admiration, blameworthy to be the fitting object of blame and so on. There are several ways to understand what is meant by 'fitting'. In line with the wider reasons-first ideology and Rønnow-Rasmussen's writings, I follow suit and understand fittingness in terms of reasons.[1]

Rønnow-Rasmussen (2011) aims to make sense of how certain objects can have value for someone, the two examples he uses are a poem written by his daughter when she was a child and remnants of a bookcase his father made for him. These objects, he argues, are for good for Rønnow-Rasmussen. They have a personal value for Rønnow-Rasmussen. Whether they also have an impersonal (final) value is a different question. Rønnow-Rasmussen shows how we can distinguish between two fundamentally different kinds of values, impersonal value and personal value, by utilizing the attitudinal component of FA to make sense of personal- and impersonal values. For an object to have impersonal value is for it to be fitting to favor it, but for an object to be good for *a* is for it to be fitting to favor the object x *for a's sake* (Rønnow-Rasmussen, 2011). Whether it is personally or impersonally valuable is therefore made evident by the attitudinal part of FA, by the way it is fitting to favor it – with an eye to for whom's sake one should favor it.

Rønnow-Rasmussen argues that the fact that an object has personal value does not entail that it has impersonal value and that something has impersonal value does not entail that it has personal value. The two are distinct and non-reducible. One and the same object can, however, be both personally- and impersonally valuable. This allows Rønnow-Rasmussen to explain how his daugher's poem can be good *for* him, without having to commit himself to the claim that his daugher's poem has impersonal value. It is fitting for Rønnow-Rasmussen to favor the poem for *his sake* – further it is fitting for *anyone* to favor the poem for *his* sake. Personal values are genuine values and not just subjective ascriptions about what Rønnow-Rasmussen

[1] Reasons-first is the claim that normative reasons are the metaphysical rock-bottom of normativity and all other normative properties can be understood in terms of reasons (Rowland, 2019). Space does not allow me to argue for reasons first, or at least that FA should be understood in terms of reasons. For the purposes of the paper, I think the position is reasonable and popular enough that I can simply have it as a presupposition of the paper.

likes. Some objects can be *good for A* even if A does not know about them, or even dislike them.[2]

The vital aspect of Rønnow-Rasmussen approach is that it is solely the attitudinal aspect of FA that is modified to accommodate for personal value. It is from "for whom's sake" it is fitting to have the attitude that determines whether for whom an object is a personal value, and whether it is a personal value. There is one aspect of the normative component of the FA analysis, the fittingness, that seldom gets discussed – for whom does it need to be fitting? The beginning of an answer to this question is the focus of the following section.

2. Fitting for Whom?

As FA is usually formulated the agential component, 'for whom' it needs to be fitting, is omitted all together. It was omitted in the previous section. Most remain satisfied with the claim that it is "fitting to favor an object". It is likely that the agent(s) for whom it needs be fitting is omitted in formulations of FA because *who* is fittingly favoring intuitively does not matter. Intuitively, if *x* is admirable it is fitting for anyone to admire *x*. Likewise, in so far as something is an increase in welfare it seems that it is fitting for anyone, or everyone, to favor it.[3] However, given what Rønnow-Rasmussen calls the "personalizability of reasons", all reasons are reasons *for someone* and if an attitude is fitting, the attitude is fitting *for someone* (2009).[4] Given that formulations of FA usually omit for whom it needs be fitting it requires a bit of speculation what philosophers have in mind. To my knowledge, when it is not omitted, it is expressed in terms of 'everyone' or 'anyone'; see for instance (McHugh & Way, 2016; Orsi, 2015; Rabinowicz, 2013; Rowland, 2019; Schroeder, 2010).

By twisting the gears in the machination of FA we can distinguish personal values from impersonal values by tinkering with the attitudinal gears of FA, whether we

[2] For Rønnow-Rasmussen, it is important that *good-for* is not equivalent with *welfare,* welfare might always be good *for someone* but not everything that is *good for someone* has to do with welfare. For our present purposes, we need not linger on this issue. For other accounts about personal value see (Darwall, 2002; Rosati, 2008).

[3] A small caveat. Some argue that there is an epistemic constraint on fittingly favoring an object (Bykvist, 2009), i.e., it is not fitting for someone who has never contemplated the welfare increase or knows about the welfare increase to favor it – but in principle as long as the salient considerations are available to the agent it is fitting for anyone to favor a just outcome or an increase in welfare.

[4] It is possible that if one rather than understanding fittingness in terms of reasons one takes fittingness as basic, there is no need for an agential component at all. Admiration just fits the object, without any reference to an agential component. In other words, it has been suggested to me that it is possible that while there is a "personalizability of reasons" there might not be a "personalizability of fittingness". Thanks to Thomas Schmidt and Andrés Garcia for suggesting reasoning along these lines.

should favor an object *for someone's sake* or not. Similarly, by tinkering with the attitudinal component of FA we can get at a distinction between final/instrumental value, by looking at valence of the attitude we can determine whether an object is good or bad, by looking at the specific type of attitude we can determine whether something is loveable or admirable and so on.[5]

As a matter of intellectual curiosity what happens if we rather than tinkering with the attitudinal component, we tinker with the agential component of FA? Tinker with for *whom* an attitude is fitting! Some objects are such that it is only fitting for some to favor it, other objects are such that it is fitting for anyone to favor it. Undoubtedly, there is conceptual space for such a maneuver. I stipulate that objects that it is fitting for anyone to favor, pace eventual epistemic constraints, have what we can call a non-relational value and objects where it is *'fitting for some but not all*' have a relational value.

> FA-NR1: *x* is non-relationally valuable if and only if, and because, it is fitting for *anyone* to favor *x*.[6]

> FA-R1: *x* is valuable in relation to A if and only if, and because, it is fitting for A but not necessarily anyone else to favor *x*.

Conceptual possibility, however, in a way comes cheap. In order to be philosophically interesting, not only does one need to prove conceptual possibility but also explain what it means; whether the distinction is actually instantiated in the world, and that the distinction is robust and does not collapse into previously made distinctions. The goal for the next sections is to show that the distinction between relational values and non-relational values is feasible, that there are plausible examples of both present in both our everyday thinking about values and our philosophical theorizing. In doing so, however, I also show that our first characterization is inadequate. Our first characterization collapses into the distinction between agent-relative and agent-neutral values and fails to fully capture the inter-personal relational aspects of the phenomenon we are trying to capture.

---

[5] Final goodness as that which we should favor for its own sake. Instrumental goodness as that we should favor for the sake of its effects. An object is good rather than bad if it is fitting to favor it rather than disfavor it. An object is, say, very admirable if it is fitting to admire it a lot. An object is admirable rather than despicable if it is fitting to admire it rather than despise it.

[6] Some might argue that non-relational values, objects where it is fitting for *anyone* to favor, is just a special case of a relational value but that it has this value in relation to any possible agent. I do not want to argue about terms and labels. There is a philosophical interesting difference between objects in which it is only fitting for some to favor it and objects in which it is fitting for anyone to favor it. The latter object's value is in no way explained by or dependent on the agent(s) or properties of the agent(s) who would be doing the favoring. For more on this see footnote 7.

3. Love, Blame, and What We Owe to Each Other

The purpose of this section is to give substantive intuitive examples of objects where it is fitting for some but not just anyone to have an attitude, and that this affects what kind of value these objects have. Examples where we want to say that the object has value even though it is not fitting for anyone (but to some) to have a pro-attitude towards it.

3.1 Love

To give an intuitive example of the distinction between relational and non-relational value consider the following (true) claim: "My parents are very loveable." This claim can be analyzed in different ways. One way is that it is fitting for anyone to love them for their own sake. Highlighting my parent's final value. Another way to analyze this claim is that it is fitting to love them *for my sake*. Thereby highlighting the personal value my parents have for me. There, however, seems to be a third possible interpretation. It is fitting *for me* to love my parents in a sense in which it is not fitting for others to do. They have, I want to argue, a value in relation to me.[7] This highly personal aspect of an intuitive sense of love is not able to be captured only by appeal to a distinction between personal- and impersonal value.

Given the qualities of my parents, I believe that all three ways of analyzing the claim can be true. Note also that in the third relational sense of loveable, it is possible that it is fitting *for me* to love them *for my sake* or for their own sake. It is therefore, at least conceptually, possible to have relational personal value or relational impersonal value. In other words, the distinction between personal/impersonal value and the distinction between relational/non-relational value are orthogonal. This should perhaps not come as a surprise since we get the distinction between personal/impersonal value by looking at the attitudinal component and relational/non-relational value by looking at the agential component.

Getting into the details of the nature of love and the value of lovability would take us too far astray.[8] Safe to say, what I have in mind is a narrow sense of love. There is a sense of love, in which love is more like liking. Then there is another sense of love including romantic love and familial love that I hope the reader is intimately familiar with. It, however, would be impertinent if I failed to mention Rønnow-Rasmussen's (2008) writing on the subject. Rønnow-Rasmussen (2008) discusses the possibility that the object of love is not the properties of the beloved but the

[7] That some of what makes them valuable in relation to me is the relationship I stand to them entails, I think, that all relational values by necessity will be extrinsic rather than intrinsic values.

[8] For some classical writings about the nature of love, see (Frankfurt, 2001; Howard, 2019; Kolodny, 2003; Plato, transl. 1998). For a new interesting view on the nature of love, see Werkmäster Johansson (MS).

particular person. The beloved is non-fungible and it would not be fitting to love an identical copy with the same properties.[9] The properties of the beloved might be what causes us to love them but, according to Rønnow-Rasmussen, is not the object of love. Rønnow-Rasmussen's insightful writing of love as a value is thought provoking. There is one section in his paper that could be interpreted as endorsing that the beloved, over and above having impersonal and personal value, also has a relational value. One of the insights is that regardless of whether there is a sense in which one can be lovable in a way in which ascribes impersonal value there is at least one additional sense. He writes:

Instead of talking about value period, it might seem more plausible to ascribe *agent-relative value-for to the beloved*. I suspect that many who would be hesitant to say that their beloved carried a final value period would at least be ready to say promptly that *the beloved is good for or has value for them* (people are probably not ready to say to the same extent that their beloved has some final value period that nobody else has)." (Rønnow-Rasmussen, 2008, p. 502 emphasis added).

In interpreting what Rønnow-Rasmussen is saying here, there is a slight tension. On the one hand, he is claiming that the beloved is *good-for* the agent. On the other hand, he seems to claim that the beloved has *agent-relative* value. In *Personal Value* (2011) it is made clear that good for is not an agent-relative value in the traditional sense; it is fitting for anyone to love my parents *for my sake*; Rønnow-Rasmussen's daughter's poem is not good-relative to Rønnow-Rasmussen, but good *for* Rønnow-Rasmussen.[10] Perhaps the most natural way to interpret (Rønnow-Rasmussen, 2008) here is that the paper on love is an earlier work and the terminology of agent-relative is something that disappeared as the work matured.

An alternative interpretation is that there is the possibility that Rønnow-Rasmussen oscillates between personal value and relational value. The leading example in his 2008 paper is Rønnow-Rasmussen's love for his wife Ellie. In his writing, it seems to be implicitly taken for granted that Rønnow-Rasmussen has reason to love his wife that we lack, or reasons to love her in a way in which we lack.[11] Not a love that we could have reason to have towards Ellie for Rønnow-Rasmussen's sake, but a different kind of love. Just as with the example of my parents and familial love, Rønnow-Rasmussen's example of romantic love seems to highlight the same structure. It is without a doubt the case that there is a sense of love in which we have reasons to love Rønnow-Rasmussen's beloved for his sake. Ellie is good for Rønnow-Rasmussen. Over and above, being good for him, what I

[9] For more on objections against what is usually called the "Qualitative view" of love see (Howard, 2019; Kolodny, 2003).

[10] For a discussion about agent-relative values, see section four.

[11] This is not, I think, an example of what Rønnow-Rasmussen (2011) calls "Janus-Values", values where it is not fitting for, say, the admirable to admire herself on pain of undermining her admirability, but fitting for others to admire her (for a's sake).

want to argue, which is something I believe is already implicitly acknowledged in his writing, is that his beloved has a value in relation to him.

This should not be read as a form of subjectivism. It is not a claim about Rønnow-Rasmussen's desire or a report of his preferences or beliefs. Such a reductive claim would not capture the very real way in which my parents are valuable in relation to me, regardless of my desires or motivational states.[12]

It is just that some of the reasons to love my parents are grounded in not just properties about them but also in properties about our relationship, which are such that I am in a unique position to have such a reason to love them. This, however, is not to say that all relational values are such that it is only ever fitting for a single agent to have some attitude. We can easily imagine where several agents stand in some relation to something such that it is fitting for a group or several individual agents to have some attitudes, but not fitting for just anyone.

3.2 Blame and Directed Duties

Moving on from love to blame. Blameworthiness is one of the more central values within our moral practices.

On FA to be blameworthy is to be the fitting target of blame. There is a conceptual connection between blameworthiness and moral wrongdoing. Lastly, and perhaps trivially, there is a conceptual connection between moral wrongdoing and moral obligations.

This has led some philosophers to argue for a buck-passing account of moral wrongdoing and moral obligations (c.f., Darwall, 2006; Skorupski, 2010).[13]

> *Wrongness-BP*: An act F is morally wrong if and only if it is fitting to blame an agent for Fing (I.e. the agent is blameworthy for Fing).[14]

Given the following intuitive and trivial principle

> *OB-W*: An act F is morally obligatory if and only if not-Fing is morally wrong (forbidden).

We get the following buck-passing account of duties.

---

[12] This is not to say that it is incompatible with subjectivism. It is. Subjectivism is not entailed, or presupposed, by anything I say.

[13] Why not pass the buck directly from moral obligations to reasons for actions, such as an act is morally obligatory if and only if it is what one has most reason to do? The answer to this is that such an account is unable to account for supererogation (Werkmäster, 2019).

[14] Wrongness-BP is a controversial thesis. Space does allow me to investigate the advantages or disadvantages of Wrongness-BP as such. While I believe there are many merits to it, for our present purposes, it is not important whether it is correct or not.

> *OB-BP:* An act F is morally obligatory if and only if it is fitting to blame an agent that does not F (i.e., F is morally obligatory if the agent would be blameworthy if she failed to F.).

One thing about moral obligations, is that some are directed, owed to others, while others are not directed, owed to no one in particular. When I make a promise to you, I have an obligation to you. If I fail to fulfill my promise without a proper excuse, not only do I do something wrong, you are *wronged* by me. Say that I promise you to give my friend a gift. If I fail to fulfil the promise, my friend is not wronged, you are. Keeping my promise is owed to you even if someone else is the benefactor. This directed character of certain duties is importantly different from other duties we might have such as a duty to not destroy some piece of art or an untouched forest. Scanlon (2008), Darwall (2006), and Wallace (2019) are a few of the philosophers that have recently been attracted to understanding morality via a focus on our relationships and to the nature of morality as an essentially interpersonal phenomenon. In short, all moral duties are directed duties.

While the first order questions in virtue of what some duties are directed, owed to others, our present purposes are with the structural issue. An undirected duty can be construed as a two-placed relation between an agent and an action. A directed duty on the other hand rather takes the form of a three-placed relation between an agent, an action, and the party who stands to be wronged. If our deontic buck-passing account, our FA analysis of duties and wrongness in terms of blameworthiness is to be correct, it should be able to account for the difference between directed and non-directed duties. This difference should be reflected in the BPA analysis of duties and wrongness in terms of blameworthiness.

A first attempt is that we can do so by distinguishing *for whom* it is fitting to blame agents that flout directed duties and non-directed duties.

> *Wronged:* A wrongs B by Fing if and only if it is fitting for B to blame A in a sense in which it is not fitting for anyone else to blame A for Fing.
>
> *Wrong:* A's Fing is wrong if and only if it is fitting for anyone to blame A for Fing.[15]

Some, such as Darwall (2006), might argue that even if A wrongs B it is possible for a third party to not just blame A, but to blame A on behalf of B. I argue, however, that blaming on behalf of someone does not capture the sense in which the wronged party can blame the wrongdoer. Blaming on behalf of B, rather seems to be to blame

[15] What is the relation between doing something morally wrong and wronging someone? Arguably, to wrong someone implies that one does something morally wrong, but not the other way around. If so, whenever A wrongs B it is fitting for anyone to blame A since A also does something that is just wrong. Not only is this a possible implication, it sounds quite plausible. Morality is not a private matter, albeit often a relational one. It is wrong to wrong someone. This, however, does not diminish that the wronged party can fittingly blame the wrongdoer in a sense not available to others.

A for *B's* sake.[16] This, however, would entail that A is blameworthy *for* B (personal blameworthiness rather than relational blameworthiness). I leave it open whether there is both impersonal- and personal blameworthiness. Nothing that I say speaks against this possibility. The possibility, however, in no way tells against the existence of relational blameworthiness.

There is furthermore, a limit to extent one can hold a wrongdoer responsible on behalf of the wronged. A third-party can for instance not forgive nor accept an apology on behalf of the wronged. By accepting the idea that there is such a thing as relational blameworthiness, we get the tools to provide a straightforward explanation of why it is the wronged who is owed an apology and why there is a sense in which it is only the wronged party who can grant forgiveness.

4. Am I Just Re-inventing Agent-relative Values?

So far, I have argued that relational values are not to be conflated with personal values. I have given substantive examples of values I think are better captured by appealing to the distinction between relational- and non-relational values. So far so good, one question that the reader might have been thinking throughout this paper is, however, the following: "Isn't he just re-inventing the distinction between agent-relative- and agent-neutral values?"

In this section, I argue that relational values are not to be conflated with agent-relative values.[17]

In order to arbitrate whether the distinction between relational values and non-relational values collapses into the distinction between agent-relative and agent-neutral values we first need a clear definition of agent-relative values. What does it mean for something to be good-relative-to? Sadly, there is no uncontroversial way to express the distinction between agent-relative value and agent-neutral value, or if a distinction even in principle could be made.

The motivation for philosophers who want to argue for a distinction between agent-relative- and agent-neutral values has often been to find a way to allow consequentialists to implement side-constraints in their moral theories. In other

[16] Rather than blaming on behalf, the locution of blaming as a representative of the moral community is sometimes used. If the moral community was wronged it makes sense that as far as A is blameworthy in relation to the moral community that any member of the moral community could relationally blame A. The metaphysics of whether groups, such as the moral community, could be owed directed duties is something I leave open for debate. What matters is that there is a difference between blaming on behalf of someone and blaming as a proper representative of someone. A man could fittingly blame sexist hiring practices on behalf of women, but not as a representative of women.

[17] However, note that I am not proposing the distinction between relational value and non-relational value as an alternative to agent-relative and agent-neutral value; it is possible that both exist even if I am skeptical of the distinction between agent-relative and agent-neutral value.

words, an explanation for why it is not the case that I have a moral obligation to murder one in order to prevent someone else murdering two, even though on pure consequentialist grounds murdering one maximizes value.[18] The distinction between agent-relative- and agent-neutral promises to deliver on this. It allows the consequentialist to say that while murdering one might be neutrally better than allowing two others to be murdered, it is better relative to you that you do not murder rather than someone else murdering. If one is supposed to maximize agent-relative value then the consequentialist can explain deontic side-constraints.

Some have tried to capture the agent-relative value and agent-neutral value distinction (Portmore, 2007) by utilizing FA. In fact, Schroeder (2007: 292) has proposed a characterization of the distinction by modifying the agential component of FA in a very similar way in which I have characterized relational value. According to Schroeder, an object is good-relative-to-A when it is only fitting for A to favor the object and an object is agent-neutrally good when it is fitting for everyone (or anyone) to favor the object.[19]

As Schroeder (2007) has shown, however, there is a devastating objection against this characterization. At least if one wants to use agent-relative value to argue for deontic side-constraints.

On Schroeder's characterization of agent-relative- and agent-neutral value the following is true:

> *X* is neutrally-better than *y* if and only if it is fitting for anyone to favor *x* more than they favor *y*.

However if we assume that *x* is [A murdering one to avoid someone else murdering two], and *y* is [A not murdering one and someone else murdering two]. In such a case, *ex hypothesi y* is better-relative-to-A than *x*.

> *Y* is better-relative-to-A if and only if it is fitting for A to favor *y* more than A favors *y*.

[18] Sen (1983) is said to be among the first to discuss agent-relative value. For a good overview of the discussion see Schroeder (2021).

[19] Some have tried to capture the agent-relative value and agent-neutral value distinction by appealing to the distinction with agent-relative and agent-neutral reasons. However, unless we have a better grasp of what this distinction is it will not illuminate the distinction between agent-relative value and agent-neutral value. For attempts on how to draw this distinction, see (Bykvist, 2012; Nagel, 1970; Parfit 1984; Rønnow-Rasmussen, 2009; Skorupski, 2010). It is of little concern to us if it is possible to draw a tenable distinction between agent-relative reasons and agent-neutral reasons, say, by appeal to a non-reducible free agent-variable. If such a distinction is feasible then relational values are plausible concerned with agent-neutral reasons rather than agent-relative reasons. Anyone would have reason to love my parents if they stood in the same relationship to my parents as I do. In this sense, the reason I am concerned with are more similar to the (agent-neutral) reason you have when you are the only available agent that can save a drowning child than with agent-relative reasons.

Given (1) that "agent A is a part of the set of 'anyone'" and (2) that "agent A should favor *y* more than *x*" it comes out as false (3) that "*x* is neutrally-better than *y*". In other words, we get inconsistent rankings, because we assumed at the start that murdering one to stop someone else murdering two was neutrally better than not murdering one and letting two die. The only thing that is every agent-neutrally better is that which is better-relative-to-any-agent. (Schroeder, 2007).

If I am merely re-inventing the wheel and the distinction between relational- and non-relational values collapses into the distinction between agent-relative/neutral value we are in deep trouble for two reasons. First, this paper would be superfluous. Secondly, Schroeder's detrimental objection would apply in force. Luckily, I believe that what we need to do is quite straightforward. Over and above tinkering with *for whom* an attitude is fitting we also need to look at the attitudinal component to get a proper characterization of relational- and non-relational values. In other words, the first stab at capturing the distinction by merely tinkering with the agential component is mistaken.

That my parents are loveable in relation to me does not mean that it is fitting for me to love them *more* than someone else. That I am blameworthy in relation to you does not mean that you should blame me *more* than you should blame someone else that does something wrong (but does not wrong you). It is a qualitative difference rather than a quantitative one. My love towards my parents is different from my love towards someone that is non-relationally valuable. The wronged's blame is different from, say, a third-party's blame. This could perhaps be cashed out in terms of Strawson's (1962) distinction that it is fitting for the victim feel resentment, third-parties to feel indignation, and the perpetrator to feel guilty. I, however, feel no need to commit to one way or the other on the validity of Strawson's observations. This way, we avoid inconsistent rankings. In order to have something to work with let us make a second stab at characterizing the distinction between relational- and non-relational values. We only need a slight modification, replace 'favor' with 'favor*'. Favor* with an asterisk is a placeholder for the specific relational attitude we have in mind.

> FA-R2: *x* is valuable in relation to A if and only if, and because, it is fitting for A but not necessarily anyone else to *favor* x*.

This second attempt is not only able to side-step Schroeder's objection against agent-relative values – it allows us to explain why some comparisons are hard and should remain hard.[20]

That it is fitting for A to blame* B does not entail that it is fitting for A to (non-relationally) blame B more than anyone else. The attitude of blame* and blame pick

[20] I am here using 'hard' in a technical sense to imply that we do not want to say that the objects being compared are equally as good nor one better than the other and that it is not obvious that improvements of one object would change how the objects compare.

out two different values. So even if B is blameworthy* in relation to A, it is possible that it is fitting for anyone (including A) to blame C more than B.

Some common comparisons are hard, take for instance "Is it fitting that I love my parents more than Nelson Mandela?", "Would it be fitting to love my wife more than my son?", and "Would it be fitting to love my first child more than my second child?"

Our answer to the first question is that, in the relational sense of love* it is fitting that I love my parents more than I love* Nelson Mandela. Given my lack of relationship to Mandela, it is not fitting for me to love* Mandela at all. In the non-relational sense, despite how great my parents might be, it is probably fitting for me to love Mandela more than my parents. We can here see how we are able to deal with Schroeder's objection. Mandela is more lovable than my parents. It is fitting for anyone to love him more than my parents. This does not conflict with the fact that my parents are more lovable* in relation to me than Mandela. As I said, it is fitting for me to love* my parents more than I love* Mandela.

Our answer to the second question might perhaps take things a step too far and I am unsure whether we should take that step. However, arguably the sense of familial love I have towards my son is qualitatively different from the romantic sense of love I have towards my wife. They denote two different relational values. Once we have distinguished between romantic love and familial love as two different attitudes, a lot of the anxiety behind the question disappears. Maybe there is a wider covering concept of love in which they are comparable, but maybe such a wider concept of love would lose what is perhaps most important to us when it comes to these kinds of attitudes and values, their inherent relational and personal nature.

Answering the last question "Would it be fitting to love my first child more than my second child?" is also hard. Here we cannot dissolve the hardness of the question by an appeal to a distinction between familial love and romantic love. If we want to take the same route in explaining the hardness of the comparison, we would have to claim that love-towards-my-child-A and love-towards-my-child-B are distinct enough attitudes as to be classified as different attitudes and the children having particular different values in relation to me. This would entail an explosion in different kinds of values and attitudes. [21] I am skeptical. On the other hand, the grounds of relational values are very peculiar and particular, so why is it unreasonable to think that the attitudes are also very peculiar and particular? I leave the project of answering and explicating the difference between love and love* for another paper. In the case of blame and blame*, I believe that the Strawsonian observation on distinguishing resentment, guilt and indignation goes some of the way in providing such a story.

[21] Another hard question that invites similar issues would be "Is it fitting for me to love my child more than you love your child?"

5. Conclusion

Philosophers often strive for universality, understanding universal values of justice, fairness and so on. Toni Rønnow-Rasmussen has showed that the analytic philosopher need not (indeed should not) shy away from the personal to understand the values in our world. With this, I hope to have shown that it is open for analytic philosophers to take relations serious and acknowledge the relational aspect inherent in some of the values most important in our everyday lives and our lives as philosophers.[22]

References

Brentano, F. (1889/2009). *The Origin of Our Knowledge of Right and Wrong* (R. M. Chisholm, Trans.). London: Routledge.

Bykvist, K. (2009). No Good Fit: Why the Fitting Attitude Analysis of Value Fails. *Mind, 118*(469), 1-30.

Bykvist, K. (2018). Agent-Relative and Agent-Neutral Reasons. In: *The Oxford Handbook of Reasons and Normativity*, edited by D. Star. Oxford: Oxford University Press.

Darwall, S. L. (2002). *Welfare and Rational Care*. Princeton: Princeton University.

Darwall, S. L. (2006). *The Second-Person Standpoint: Morality, Respect, and Accountability*. Cambridge, MA: Harvard University Press.

Ewing, A. C. (1948). *The Definition of Good*. Westport Conn: Hyperion Press.

Frankfurt, H. (2001). Some Mysteries of Love. *The Lindley Lectures.* The University of Kansas.

Howard, C. (2019). Fitting Love and Reasons for Loving. In *Oxford Studies in Normative Ethics Volume 9*. Oxford: Oxford University Press.

Johansson Werkmäster, M. (ms.). Aspects of Blame.

Kolodny, N. (2003). Love as valuing a relationship. *Philosophical Review, 112*(2), 135-189.

McHugh, C and Way, J. (2016). Fittingness First. *Ethics 126*(3): 575–606.

Nagel, T. (1970). *The Possibility of Altruism.* Princeton: Princeton University Press.

Orsi, F. (2015). *Value Theory*. London: Bloomsbury Academic.

Parfit, D. (1984). *Reasons and Persons*, Oxford: Clarendon Press.

[22] This paper was presented at Thomas Schmidt's Colloquium in Practical Philosophy, Humboldt University and at Swedish Congress of Philosophy, 2022. Thanks are owed to its participants. Thanks are also owed to Andrés Garcia, Mattias Gunnemyr, and Marta Johansson Werkmäster for comments on earlier drafts of this paper. This research was funded by the Swedish Research Council, grant number: 2020-06383.

Plato. (1998). *The Symposium*. Trans. Archimedes Icarus and Alexandra Nehemas. Florida, IN: Jewel Publishing Company.

Portmore, D. (2007). Consequentializing moral theories. *Pacific Philosophical Quarterly 88*(1):39–73.

Rabinowicz, W. (2013). Value: Fitting-Attitude Account of. In H. LaFolette (Ed.), *International Encyclopedia of Ethics* (pp. 1-12). Wiley-Blackwell

Rabinowicz, W., & Rønnow-Rasmussen, T. (2004). The Strike of the Demon: On Fitting Pro-attitudes and Value. *Ethics, 113*(3), 391-423.

Rowland, R. (2019). *The Normative and the Evaluative: The Buck-Passing Account of Value*. Oxford: Oxford University Press.

Rønnow-Rasmussen, T. (2009). Normative reasons and the agent-neutral/relative dichotomy. *Philosophia, 37*(2), 227-243.

Rønnow-Rasmussen, T. (2011). *Personal Value*. Oxford: Oxford University Press.

Rønnow-Rasmussen, T. (2022). *The Value Gap*. Oxford: Oxford University Press.

Rønnow-Rasmussen, T. (2008). Love, Value and Supervenience. *International Journal of Philosophical Studies, 16*(4), 495-508.

Rosati, C. (2008). Objectivism and Relational Good. *Social Philosophy & Politics. 25*(1): 314-49.

Scanlon, T. M. (1998). *What We Owe to Each other*. Cambridge: Belknap.

Scanlon, T. M. (2008). *Moral Dimensions: Permissibility, Meaning, Blame*. Cambridge, Belknap Harvard University Press.

Schroeder, M. (2007). Teleology, agent-relative value, and 'good'. *Ethics, 117*(2), 265-000.

Schroeder, M. (2010). Value and the right kind of reason. *Oxford Studies in Metaethics, 5*, 25-55.

Schroeder, M, (2021). Value Theory. *The Stanford Encyclopedia of Philosophy* (Fall 2021 Edition), Edward N. Zalta (ed.). <https://plato.stanford.edu/archives/fall2021/entries/value-theory/>

Skorupski, J. (2010). *The Domain of Reasons*. Oxford: Oxford University Press.

Sen, A. (1983). Evaluator relativity and consequential evaluation. *Philosophy and Public Affairs, 12*(2), 113-132.

Strawson, P.F. (2008/1962). Freedom and Resentment. In: *Freedom and Resentment and Other Essays.* London: Routledge.

Wallace, R. J. (2019). *The Moral Nexus.* Princeton: Princeton University Press.

Werkmäster, J (2019). *Reasons and Normativity*. PhD Dissertation, Lund University.

Ode to Three Apprentices

In days of yore, three scholars great,
Sat huddled close, by flickering fate,
With pen in hand and heart alight,
They toiled away, day and night.

Their task, a festschrift, grand and true,
Dedicated to mentors, cherished few,
Their minds, ablaze with wisdom bright,
Seeking the words, to set it right.

But oh, the struggle, so intense,
As they battled with each sentence,
And faced the toil, that it demands,
With pen in hand, and grit of hands.

Yet still they pressed, with all their might,
Guided by the flame, that burned so bright,
For their mentors, they'd see it through,
And make their mark, with words so true.

And in the end, they emerged with pride,
Their festschrift, complete, a joyous ride,
And as they smiled, with hearts so light,
Their mentors' legacies, shining bright