The Ethics of Post Hoc interventions:
Three Potential Problems

Mattias Gunnemyr

In Gunnemyr, Mattias & Jönsson, Martin L. (2023) *Post Hoc Interventions: Prospects and Problems.* Lund: Department of Philosophy, Lund University. https://doi.org/10.37852/oblu.184

ISBN: 978-91-89415-60-7 (print)
978-91-89415-61-4 (digital – pdf)
978-91-89415-62-1 (digital – html)

DOI: https://doi.org/10.37852/oblu.184.c507

Published by the Department of Philosophy, Lund University.
Edited by: Mattias Gunnemyr and Martin L. Jönsson
Cover layout by Cecilia von Arnold, Pufendorf Institute for Advanced Studies

7

The Ethics of Post Hoc Interventions

Three Potential Problems

Mattias Gunnemyr[1]

Abstract. The paper investigates three potential ethical problems related to the use of post hoc interventions: that they might infringe on the freedom of the decision makers, that they might correct for bias even when they should not even if all conditions for applications are satisfied, and that they problematically might rely in probabilistic evidence that does not tell us anything about whether the decision at hand is biased. It is argued that while post hoc interventions might infringe on the freedom of the decision makers, they do not do so in a problematic way – especially not if implemented in as decision support system, that we either should add a condition for application of post hoc interventions or apply it in a specific way to avoid incorrect updates of decisions, and that post hoc interventions do not rely on probabilistic evidence in a problematic way. The focus of the paper is a particular post hoc intervention called GIIU (Generalized Informed Interval Scale Update).

The Need for Post Hoc Interventions

Which group we are perceived to belong to often affects our prospects of getting jobs, research funding and good grades. Consider first job applications. In an American study from 2004, Bertrand and Mullainathan showed that a job seeker named Jamal typically needed eight more years of work experience to get the same response from employers as a candidate named Greg. In a similar

[1] Mattias Gunnemyr, Researcher in practical philosophy, Department of Philosophy, Lund University. Post doc in the Financial Ethics Research Group, Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg.

study from 2007, Correll, Benard and Paik showed that a woman who wrote in her CV that she was a member of the American PTA (Parent-Teacher Association) had only half the chance of getting an interview as a woman who did not state this in her CV. Zschirnt and Ruedin (2016) show in their meta-study that ethnic discrimination in hiring decisions is widespread across OECD countries: equivalent minority candidates need to send around 50% more applications to be invited for an interview than majority candidates. In another meta-study, including 97 field experiments and over 200,000 job applications, Quillian et al. (2019) find that discrimination rates concerning ethnicity vary strongly by country, where France and Sweden stand out with the highest discrimination rates, much higher than for instance the U.S.[2]

Which group we are perceived to belong to also affects our prospects of receiving research funding. In an internationally recognized study, Wennerås and Wold (1997) showed that a woman who applied for research funding from the Medical Research Council (now part of the Swedish Research Council) needed an average of three extra scientific publications in a well-known journal such as *Nature* or *Science*, or 20 extra publications in a less well-known but still well-regarded journal such as *Infection and Immunity* or *Neuroscience*. Further, Tamblyn, Girard, Qian, and Hanley (2018), who evaluated all grant applications submitted to the Canadian Institutes of Health Research between 2012 and 2014, found evidence of gender bias of sufficient magnitude to change application scores from fundable to nonfundable. Relatedly, Lincoln, Pincus, Koster, and Leboy (2012) studied U.S. scholarly awards and prizes within STEM research between 1991 and 2010 and found that men receive an outsized share of such awards and prizes compared with their representation in the nomination pool.

Further, which group we are perceived to belong to might influence our likelihood of getting fair grades. For instance, Lavy (2008) evaluated Israeli high school matriculation exams in nine subjects and found a bias against male students, and Kiss (2013) showed that second-generation immigrants in Germany have math grade disadvantages in primary education while girls are systematically graded better in math than boys in upper-secondary school. In addition, Hinnerich, Höglin, and Johannesson (2015) found a sizeable and

[2] On the brighter side, Bygren and Gähler (2021) find no evidence that employers in Sweden statistically discriminate against women. On the less positive side, however, Arai, Bursell, and Nekby (2016) and Bursell (2014) make evident that Swedish employers discriminate against male applicants with Arabic or North African names.

robust discrimination effect against students with foreign backgrounds in grading of Swedish national tests in the Swedish high schools.[3]

Decisions made on the basis of biased judgments are usually both unfair and incorrect. They are unfair because some people are disadvantaged simply because of their group membership while others are advantaged because of theirs. They are incorrect because they do not lead to the most merited person getting the job or the research funds, because they result in students not getting the grade they deserve, and so on. This raises the question of whether it is possible to make decisions fairer and more accurate.

The most common approaches in the literature on prejudice prevention involve preventing the prejudiced decision to occur in the first place (Madva 2020). These include individual interventions aimed at making the evaluator less prejudiced, and structural interventions aimed at changing the circumstances in which the decision takes place with the aim of reducing the number of biased decisions (such as the introduction of anonymization or criteria-based decision-making). While the latter kind of intervention might have some effect, the former typically have little to no effect (Lai et al. 2014; Forscher et al. 2019; Paluck, Porat, Clark, & Green 2021).

A less explored kind of interventions aim to address prejudiced decisions after they have been made but before they have a negative effect. These are the *post hoc interventions,* discussed in this volume. Post hoc interventions might come in many different forms. The texts in this volume focus on GIIU (Generalized Informed Interval Scale Update), and I will do the same. Roughly, the idea behind GIIU is to identify an evaluator's bias towards a certain social group by surveying his or her previous decisions, and then use this information to debias subsequent decisions. On a straight-forward model, debiasing occurs automatically. For instance, GIIU might be implemented in the relevant software, automatically updating the evaluator's submitted rankings of applicants for a certain job, or the evaluator's grading of students.

[3] Still, as Bergqvist Rydén (2022) warns us, assessment practice is always deeply contextual and shaped in an assessment culture, and such cultures often vary locally and disciplinary. Therefore, results from a study on assessment bias and anonymization cannot necessarily be assumed to apply to another context. For instance, while there is evidence of discrimination against students with foreign background in the Swedish high school, Hinnerich, Höglin, and Johannesson (2011) find no evidence of discrimination against boys in grading in the Swedish high school. Further, Bygren (2020) examines group differences in average grades prior to and after an introduction of blinded examinations at Stockholm University and finds no gender bias. However, he finds a weak tendency that examiners discriminate positively for students perceived to have an immigrant background.

On a more subtle model, the debiasing does not occur automatically. Instead, the evaluator is informed that, based on his or her previous rankings or gradings, there are reasons to believe that the current ranking or grading is biased, and that he or she should consider re-evaluating the ranking or some of the grades or ask for a second opinion. This could be followed by a recommendation about what the ranking or grades should be.

While the use of post hoc interventions promises to increase accuracy and fairness in hiring processes, gradings, evaluations of research proposals, etc., it also raises ethical issues. First, it might be objected that evaluating and updating the decision makers' decisions infringe on their freedom to make decisions as they see fit. Second, there is the worry that we should not revise decisions on mere statistical grounds. What matters is the quality of the application or examination at hand, not the mistakes the evaluator previously has made considering other applications or examinations. Third, there is the related worry that the intervention mistakenly changes (or recommends to change) a decision that should not be changed. Possibly, there are also other potential ethical problems with using post hoc interventions, but these are the three worries I will address here.

"Don't Mess with My Evaluation!"

Imagine that you are evaluating applications for a certain position. After having gone through the applications thoroughly, you give each applicant a certain score based on his or her previous experience, education, and so on. Finally, you rank the applicants, and submit the evaluation using the required software. Later, you learn that the software changed the ranking you suggested. Based on your previous evaluations of applicants, the software deemed that you had given some applicants for this position too high a score. How would you react? Preliminary inquiries indicate that many decision makers react negatively to having their decisions evaluated and changed in this way. They have the *lingering feeling* that there is something wrong about subjecting one's evaluation to reworking after the decision is made. As a result, they might resist using GIIU. Tellhed (this volume) calls this *The "Will Not" Challenge*. Is there something to this worry?

There are of course several possible explanations for why some people have this lingering feeling. They might worry that the evaluation might reveal that they harbor implicit biases and make biased decisions. This kind of worry would be similar to the stress students might feel before an exam. It is the

worry that the exam or evaluation might show that they are not good enough. Decision makers might also worry that GIIU might reveal to colleagues and others that they harbor implicit biases, something that also is potentially distressing. These kinds of considerations might explain why some people feel an unease about implementing post hoc interventions like GIIU. While employers who consider implementing GIIU, and researchers researching the effects of such implementations, certainly should take such considerations seriously, they are not the main focus here. People might think that implementing post hoc interventions is justified, but still be worried about what these interventions will reveal, to themselves and to others. Instead, the focus here is whether the lingering feeling that there is something wrong about post hoc interventions reflects the idea that such interventions are not justified; that is, the idea that there is something morally problematic with such interventions.

There are several reasons why one might think that post hoc interventions are not morally justified. One might for instance think that such interventions interfere with one's freedom to make decisions as one sees fit. Alternatively, the idea that post hoc interventions are not justified might be explained in terms of (lack of) autonomy, control, respect, trust, professionalism, etc. For the sake of brevity, I will focus on the question whether post hoc interventions interfere with the decision makers' freedom. I will argue that while post hoc interventions do interfere with the decision makers' freedom, they do not do so in a morally problematic way.

There are two common ways of understanding freedom. First, there is the *liberal* understanding of freedom as the ability to do whatever one wants to do. On this understanding, the opposite of freedom are restrictions of different sorts: laws, regulations, prohibitions, and the like. Usually, liberal freedom is taken to come in two variants: negative and positive. Negative freedom is freedom from external constraints, and positive freedom involves having the ability and resources to do whatever one wants to do in a certain situation. Historically, the liberal notion of freedom can be traced at least to Hobbes, who wrote that "A free man is he that […] is not hindered to do what he has a will to". (1997/1651). Other proponents include Burke (1986/1790), Mill (2008/1859) and Berlin (1958).

If this is how we understand freedom, it seems that at least post hoc interventions of the more straight-forward type do interfere with the decision makers' freedom to do whatever he or she wants to do. For illustration, imagine once more that you after careful deliberation have suggested a ranking of candidates for a job position, and learn that the software through which you submitted your ranking has changed it. Imagine further that the decision of

who gets the position will be based on the updated ranking. In such a case, the ranking you suggested was hindered; you lacked the ability to put forward the ranking you deemed was the correct one. This might explain why some decision makers are reluctant to post hoc interventions: these interventions infringe on their freedom.

Still, it is far from clear that this kind of restricted freedom is morally problematic. There are limits to freedom, often expressed in *the harm principle*: *People should be free to act however they wish unless their actions cause harm to others.* In the words of Mill, "The only purpose for which power can be rightfully exercised over any member of a civilized community, against his will, is to prevent harm to others." (Mill 2008/1859). You are not allowed to, for instance, hit someone just for fun; not against their will. Civilized society might rightfully enact laws against such behavior, even though doing so infringes on people's freedom. A similar thing might be said about post hoc interventions. Even though such interventions interfere with the freedom of the decision makers, this interference might be justified if it hinders them from causing harm to others. Further, since we have reasons to believe that their decisions, if unaltered, will cause harm to others, the interference might very well be justified. A balancing of reasons must be made. We have to compare the harm done by implementing GIIU in terms of interfering with the freedom of the decision makers, to the harm done in terms of the most merited applicant not getting the position, of students not getting fair grades, etc. As we have seen, these latter harms are all too common and severe, and we have reasons to believe that they outweigh the harm done in terms of interfering with the freedom of the decision makers. Further, the former kind of harm is most likely lesser. Decision makers are typically expected to make correct decisions. If they fail in this, the harm of correcting them – that is, the harm of infringing their freedom to make biased and incorrect decisions – is probably not great.

Someone might object that the harm principle does allow us to infringe on the freedom of the decision makers in this case; they might point out that it does not concern all causings of harm. Upon closer scrutiny, and implicitly, it only concerns proximate harms. It forbids things like beating and killing others. In contrast, the harm principle does not forbid causing harm to distant others. For instance, it does not forbid you to hire someone to beat someone else up. If you do, it is not you who harm this person, it is the thug you hired. Hiring a thug to beat someone up might be wrong for other reasons, but it is not forbidden by the harm principle (see e.g. McLaughlin 1925-26; Grady 2002). Having this in mind, someone might object that making a biased ranking of applicants for a job or giving students the wrong grades because of

implicit bias are not instances of causing proximate harm, and so is not forbidden by the harm principle.

This line of reasoning is mistaken. Even granting that the distinction between proximate and distant causes is morally relevant (which we have reasons to doubt, see e.g. Moore 2009), making a biased ranking of applicants for a job position or giving students the wrong grades are plausibly seen as the proximate cause of harm, and so forbidden by the harm principle. That is, you are not free to make such rankings or gradings as you please. Further, even if it turns out that making such rankings or gradings are not the proximate cause of harm according to some plausible definition of what it is for a cause to be proximate – and by extension that the harm principle does not forbid such rankings or gradings – there might still be reasons to think that you are not free to cause such harms. For comparison, plausibly, you are not free to hire someone to beat someone up just because you want to even though doing so is not the proximate cause of harm.

The upshot of the discussion on the liberal understanding of freedom is that while post hoc interventions like GIIU might interfere with your freedom to make rankings and gradings as you see fit, this interference is warranted insofar as it hinders you from causing harm to others.

Second, there is the *republican* understanding of freedom as non-domination or independence from the arbitrary will of others. On this understanding, the opposite of freedom is not restrictions, but slavery. Within this tradition, it is debated what the conditions of being independent from the arbitrary will of others amounts to. Locke (1980/1690) argues that you are subjugated to the arbitrary will of others when they have the power to control all aspects of your life. This is for instance true if you live in an autocracy where the king or dictator of the autocracy at any time could imprison you or send you to war. Children provide another example. Their parents control more or less all aspects of their lives. Others, like Wollstonecraft (1988/1792), argue that you are subjugated to the arbitrary will of others if this will is unreasonable. On this view, children are not necessarily subjugated to the *arbitrary* will of their parents. Insofar as the parents' decisions are reasonable, the children are not unfree. Similarly, at least in theory, you might live a free life in an autocracy if the dictator makes reasonable decisions, as in a benevolent dictatorship. (However, Wollstonecraft does not think this is a tenable form of government. Power always corrupts, she argues, with the result that the benevolent dictatorship, if there is such a thing, eventually will turn into an oppressive one.) More contemporary proponents of republican freedom include Pettit (1997) and Skinner (1998).

There is something to the idea that the implementation of post hoc interventions interferes with the freedom of decision makers, where freedom is understood in the republican way. Their decisions are dominated, or overruled, by others; the decision makers are not independent from the will of others. This might explain why some decision makers are reluctant to the implementation of post hoc interventions. Still, it is far from clear that the implementation of post hoc interventions subjugates the decision makers to the *arbitrary* will of others. In Locke's view, you are only subjugated to the arbitrary will of others if all (or most) aspects of your life are subjugated to the will of others. This is not the case when it comes to the implementation of post hoc interventions. Post hoc interventions do not concern all aspects of the decision makers' lives. Then again, Locke's view of freedom as non-domination does not seem to apply well to the question under consideration. It is tailor-made to apply to questions about how the state should be governed; as a dictatorship or a republic. Perhaps Wollstonecraft's view is better suited for evaluating post hoc interventions. According to her, you are not subjugated to the arbitrary will of others if this will is reasonable. We must then ask if it is reasonable for an employer, for instance, to implement GIIU at the workplace. Wollstonecraft does not give much guidance for how to evaluate whether a will is reasonable, but I take it that there is a good case to be made for thinking that it is. GIIU, if correctly used, will improve the accuracy of rankings of applicants for job positions, and thus in the end result in more merited personnel being hired. Similarly, they will improve the accuracy of teachers' gradings, referees' rankings of research proposals, etc.

Still, as Wollstonecraft sees it, there is a certain inherent value in being independent. It is better to make reasonable decisions yourself than to be subjugated to the will of others, even if their will is reasonable (*ceteris paribus*). Applied to post hoc interventions, this idea seems to entail that those who evaluate applications should advocate the implementation of post hoc interventions or implement them themselves, that the teachers themselves should advocate the implementation of post hoc interventions or implement the interventions themselves, etc. At least, this is the case insofar as implementing post hoc interventions is the reasonable thing to do. I will not pursue this idea here, but I think there is something to it. Professionals that find out that their actions bring about harmful outcomes should find ways to improve their ways of working. Just as journalists in many countries with freedom of the press have adopted codes of practice to reduce the possibility of causing harm to others in their course of work, professionals who make

decisions that importantly influence the lives of others should take measures to see to it that these decisions are fair.

Finally, also on the topic of freedom, there are reasons to prefer the more subtle version of GIIU where the updating of rankings or grades does not occur automatically to the more straight-forward version of GIIU discussed here where it does. Informing the decision makers that there are reasons to believe that the ranking or grading they just made is biased and encourage them to reevaluate some applications or exams, but giving them the final say about what the final ranking or grading should be, arguable interferes less with their freedom than what an automatic update does.

The Possibility of Incorrect Interventions

We have reasons to believe that the use of GIIU is justified provided that it helps us make more accurate and fair decisions. However, sometimes it seems to provide less accurate and fair decisions. Consider Recruiter, who has a long history of evaluating applicants' competence. Their actual competence is shown in the following table: (For ease of exposition, I only consider 6 applicants and 1 ranking).

Applicant	Education	Social skills	Experience	Average
Anthony	8	8	2	6
Benjamin	6	6	3	5
Charles	3	3	3	3
Deborah	7	7	7	7
Emma	6	6	6	6
Fiona	3	3	6	4

If Recruiter correctly evaluates the applicants' competence, he will rank them in the following order: Deborah, Anthony, and Emma (tie), Benjamin, Fiona, and Charles. Given this ranking, Deborah would get the position. However, Recruiter suggests a quite different ranking, namely: Anthony, Deborah, Benjamin, and Emma (tie), Charles and Fiona (tie). Here, the men are ranked higher than they are in the correct ranking. Anthony is for instance ranked

higher than Deborah instead of lower, Benjamin is ranked as tie with Emma instead of lower than Emma, and so on. So, it seems that Recruiter is biased against women, and this is also what GIIU would say. In the next recruitment process, GIIU would recommend updating Recruiter's ranking; it would recommend giving female applicants a higher score than Recruiter does.

However, there is a possibility that Recruiter is not biased against women. There is another possible explanation for why his ranking is different from the expected one. He might not think that experience matters for the position at hand. In fact, if we disregard experience, he has suggested the correct ranking. In this sample, the women have higher experience than the men, and if experience is not taken into account, they get lower average scores while the men get higher average scores, as follows:

Applicant	Education	Social skills	Experience	Average
Anthony	8	8	2	8
Benjamin	6	6	3	6
Charles	3	3	3	3
Deborah	7	7	7	7
Emma	6	6	6	6
Fiona	3	3	6	3

Given these average scores, Recruiter's ranking is correct. Anthony has the highest average score, followed by Deborah's, and so on.

One might suspect that Recruiter has engaged in motivated reasoning when deciding that experience does not matter for this position. He might have disregarded experience in order to arrive at the desired verdict that Anthony should get the position and not Deborah. However, say that this is not the case. Recruiter does in fact not have any bias against women. If things would have been different, and the men in his evaluation history had had more experience than the women, he would still have disregarded these merits when making his ranking. In such a case, we would not want GIIU to infer that Recruiter is biased against women. Rather, we would want to get the verdict that GIIU does not apply, and we would want to get this verdict since GIIU is not designed to correct mistakes other than those that are based on biases against certain social groups. We would also possibly want an indication that Recruiter wrongly disregards experience when making his rankings.

There are situations when GIIU does not apply. Jönsson and Bergman (2022) suggest the following conditions for GIIU to apply:

(1) Evaluations are carried out using, minimally, an interval scale.

(2) The history of evaluations is large enough to reliably find prejudices with a suitable statistical test.

(3) The mean values in the relevant populations of whatever is being evaluated are known, or are known to be the same.

(4) GIIU makes use of subsets of the groups the evaluator is prejudiced against.

(5) Any fluctuations in E's prejudice are small compared to the size of the corresponding prejudice.

(6) The evaluator's prejudice operates in an approximately linear way.

(7) The evaluator's prejudice operates on discrete groups.

In the case at hand, (2) is not satisfied. The history of evaluation is not large enough. However, we can disregard this problem. It is possible that the indicated problem would occur even if the history of evaluations would be large enough. Here, I used a small history for the sake of exposition.

One suggestion for avoiding the problem at hand is to add a condition similar to (4), namely the following:

(4*) GIIU makes use of the same competences as the evaluator does when calculating the evaluator's bias, or subsets thereof.[4]

This condition is not satisfied in the case under consideration. When evaluating the evaluator's bias, GIIU presumes that education, social skills, *and* experience are relevant for the position, while the evaluator only deems that education and social skills are important for the position. So, given that (4*) is required for GIIU to apply, we find that it does not apply on this particular occasion, and so will not wrongly deem that the evaluator is biased against women, and wrongly compensate for this bias in future recruitment processes. Moreover, when checking whether (4*) is satisfied, we will find indications that Recruiter wrongly disregards experience when making his evaluations.

Still, it is not obvious that the extra condition (4*) is needed. Upon closer reflection, it turns out that condition (6) is not satisfied. If we only consider

[4] Condition (4) could also be updated to include this requirement.

average scores, it might seem that (6) is satisfied in Recruiter's history. Women consequently get a lower average score than they should, and men consequently get a higher average score than they should, so it might seem that Recruiter has bias against women that operates in an approximately linear way. However, this illusion disappears if we look at Recruiter's evaluation of each competence instead of the average scores. We then see that Recruiter evaluates women's and men's competences correctly, but disregards experience. This amounts to setting the experience for all men and women to the same value, such as zero, regardless of what their experience is. Doing so is not a linear function, and therefore we can conclude that condition (6) is not satisfied.

So, we can conclude that we face a choice: Either, we can continue applying GIIU to the applicant's average competence score and add a further condition of application for GIIU, such as (4*). Or, we can apply GIIU to each relevant competence score rather than to the average competence score. Either way, we avoid the problem that GIIU might suggest inaccurate and unfair updatings of rankings in cases where an unbiased evaluator disregards a certain competence when making his evaluations, and where this competence is unequally distributed among the salient social groups.

Before we leave this topic, there is a final issue that should be mentioned. As the example is construed, Recruiter is not biased against women, and does not discriminate against them directly. However, this is likely a case of indirect discrimination. Indirect discrimination occurs when there is a policy that applies in the same way for everybody but disadvantages a group of people who share a protected characteristic. Importantly, it makes no difference whether anyone intended the policy to disadvantage you or not. To go free from charges of indirect discrimination, you must show that there are good reasons for the policy. At least, this is the case in many jurisdictions, such as Sweden and the UK. Still, there seems to be no good reasons to disregard experience in a typical hiring procedure. So, the case under consideration is most likely a case of indirect discrimination, which in turn means that the unaltered version of GIIU (i.e. GIIU applied to average scores and without 4*) compensates for indirect discrimination. Therefore, the harm done if we would use the unaltered version of GIIU is limited. Indeed, in some respects, it is an advantage that GIIU might compensate for indirect discrimination.

Verdicts Based in Statistics

Basing verdicts on mere statistical evidence is problematic. Consider for instance the following case:

> *Blue Bus:* A bus causes harm. There is no eyewitness, but we have uncontested data regarding the distribution of buses in the relevant area. The Blue Bus Company runs roughly 80 percent of the buses there.

Even though we have statistical evidence that it was a Blue Bus that caused harm, the evidence does not seem to be enough to support the belief that it was a Blue Bus that caused harm. Moreover, the law would typically not find the Blue Bus Company liable on statistical evidence alone. In some jurisdictions, such evidence would not even be considered relevant.

This poses a potential problem for GIIU. GIIU involves revising decisions – or recommendations to revise decisions – on the basis of statistical evidence. Could this ever be justified?

It might. Evidence that comes with a certain probability is not always problematic. Consider for instance the following case:

> *Blue Bus with Eyewitness:* A bus causes harm. There is an eyewitness. The eyewitness reports that a bus belonging to the Blue Bus Company caused harm. The witness, however, is unreliable. Let us say that she is roughly 80 percent reliable in this case.

In this case, it seems appropriate to form the belief that it was a bus belonging to the Blue Bus Company that caused harm. Further, the law will typically find the Blue Bus Company liable for harm in such circumstances.

The question is whether the evidence GIIU uses is more like the statistical evidence in *Blue Bus,* or more like the evidence in the form of an eyewitness in *Blue Bus with Eyewitness*? On the one hand, it might seem that the evidence GIIU uses is more like the former. It uses statistics about an evaluator's previous decisions as evidence for (recommending) updating her decisions about rankings, gradings, or the like. If this is the case, it seems that GIIU uses evidence in a problematic way when forming recommendations or revising decisions; the belief that the updated decisions are the right ones does not seem supported. On the other hand, it might seem that the evidence GIIU uses is like the latter. GIIU does not use statistics over how biased decision makers in general are as grounds for (recommending) updating. Rather, GIIU uses that

particular evaluator's history of decisions as grounds for calculating that evaluator's bias (if any); a calculation that then is used to determine whether the current decision is biased and in need of revision. If this is the case, it seems that GIIU does not use evidence in a problematic way. The belief that the updated decisions are the right ones seems supported.

Is there a principled way of deciding cases where it is fitting to form a certain belief on probabilistic evidence from cases where it is not? There are several suggestions in the literature for how to do this (see e.g. Redmayne 2008). Enoch, Spectre, and Fisher (2012) suggest the perhaps most promising principle. The basic idea is simple: Our belief that something is the case should be appropriately sensitive to the truth. They suggest the following principle:

> *Sensitivity: S*'s belief that *p* is sensitive = df. Had it not been the case that *p*, *S* would (most probably)[5] not have believed that *p*. (Enoch et al. 2012: 204)

When a belief is not sensitive, it is of the problematic kind. Consider again *Blue Bus,* where it does not seem fitting to form the belief that it was a bus from the Blue Bus Company that caused harm, and say that someone, *S*, forms the belief that it was a blue bus that caused harm on the basis of the statistical evidence. This belief is not sensitive. Had it not been the case that it was a bus from the Blue Bus Company that caused harm – say that it actually was a red bus – the statistical evidence would still have been just the same, and *S* would (most probably) still have believed that it was a Blue Bus. The statistical evidence at hand is not sensitive to whether it was a blue bus or a red bus on this particular occasion.

Things are different in *Blue Bus with Eyewitness.* Consider someone, *S*,* who forms the belief that it was a blue bus that caused harm on the basis of the witness' report. This belief is sensitive. Had it not been the case that it was a blue bus – say that it was a red bus instead – the witness would (most probably) not have reported that it was a blue bus, and so *S** would (most probably) not have believed that a blue bus caused harm. These results generalize to most similar cases. While we should grant that *Sensitivity* is not the only plausible

[5] They add the most-probably qualification to bypass a technical problem, having to do with the common way of fleshing out counterfactual semantics in terms of possible worlds. The problem is that worlds that are less likely to be the actual one (such as the one where the eyewitness is mistaken) are not guaranteed to be further from the actual world than more likely worlds. I am not sure the most-probably qualification helps us avoid the technical problem. Still, this is not the place to sort out these technical details. I will assume that it is possible to avoid the technical issue Enoch et al gestures at, and that we safely can go on using *Sensitivity.*

way to distinguish probabilistic evidence of the problematic kind from the unproblematic kind, it gives reliable enough guidance to do so.

We can now return to the question of whether GIIU problematically bases its verdicts on statistical evidence. It turns out that it does not. Say that *S*** bases her belief that a certain ranking given by evaluator *E* is biased and should be updated based on GIIU's recommendations (which in turn is based on *E'*s history of evaluations). This belief is sensitive. Had it not been the case that the ranking was biased and should be updated, GIIU would (most probably) not have indicated so, and *S*** would (most probably) not have believed that the ranking is biased and should be updated. Therefore – at least insofar as we can trust *Sensitivity* – we can conclude that *S***'s belief is not of the problematic kind, and that GIIU does not base its verdicts on probabilistic evidence in a problematic way.

Someone might object that while evaluator *E'*s history of biased rankings gives us reasons to believe that *E* has been biased previously, we cannot infer that he was biased on this particular occasion. Maybe he has changed for the better. The only way to know for certain that *E* was not biased on this last occasion, they might argue, is to measure his bias on this particular occasion. We must use some device – maybe a brain scanner of sorts – to decide whether he is biased when making his decision.

This objection is mistaken. There might of course be cases where the evaluator's prejudices have changed. However, GIIU is designed not to apply to those cases. It only applies when any fluctuations in *E*'s prejudice are small compared to the size of the corresponding prejudice. This is the fifth application condition for GIIU. Granted, it might be hard to decide whether *E*'s prejudice has changed over time. Still, as Jönsson (this volume) argues, there are ways of deciding this; ways that do not involve brain scanning. The reason why GIIU does not base its decisions in statistical evidence in a problematic way, then, is roughly the following. We have empirical evidence that *E* previously has made biased decisions in the form of a history of biased decisions. This evidence is not based on statistics in a problematic way. That is, it is sensitive to whether *E* was biased. Had he not been biased, the decisions he made would not have been biased. Further, we have evidence that *E'*s prejudice remains significantly unchanged, and that it still influences his decisions. Therefore, we have evidence that the current decision is also biased. This is not evidence of the problematic statistical kind. It is not merely arguing that since *E* previously made biased decisions, he must have made a biased decision this time as well. It is arguing that since *E* was biased before, and since he has not changed, he is biased now.

Conclusions

To sum up, I have considered three potential ethical problems with implementing post hoc interventions, focusing on GIIU. First, post hoc interventions like GIIU might be morally problematic since they infringe on decision makers freedom. I argued that while some forms of such interventions – the more straight-forward ones that automatically update the decision makers' decision – do infringe on the decision makers' freedom, this is most likely not morally problematic. There is no reason why we should grant decision makers the liberty to make biased and inaccurate decisions that cause harm to others. Moreover, the restricted freedom of the decision makers is much less of a problem if we implement more subtle post hoc interventions, that is, interventions that do not automatically update the decisions of the decision makers, but instead identifies the decisions that are likely to be biased and recommends updating these decisions. Further, I suggested that it would be in the interest of the decision makers to implement some kind of post hoc interventions themselves. GIIU might for instance provide a useful tool, potentially increasing the accuracy of their decisions and thereby help avoiding making discriminatory ones.

Second, in some cases, GIIU might indicate that a certain decision should be updated even though it should not. I argued that this problem might be avoided if we either add a further condition for application of GIIU, or that we use GIIU to evaluate each competence score (or equivalent) instead of using it to evaluate average scores.

Finally, GIIU might objectionably rely on probabilistic evidence. I argued that it sometimes is perfectly fine to rely on probabilistic evidence, that there is a principled way of deciding when it is, and that GIIU does not rely on probabilistic evidence in an objectionable way.

Acknowledgments

This paper was presented at the conference "Post Hoc Interventions: Prospects and Problems" at the Pufendorf Institute for Advanced Studies in Lund in the fall -22. I want to thank the participants at the conference for insightful comments. In particular, I want to thank my designated commentator, Eric Brandstedt. I also wish to thank Martin L. Jönsson and Kasper Lippert-Rasmussen for detailed comments on a previous version of the paper. Last, but not least, this work was supported by the Pufendorf Institute for Advanced Studies.

References

Arai, M., Bursell, M. & Nekby, L. (2016) The reverse gender gap in ethnic discrimination: Employer stereotypes of men and women with arabic names. *International Migration Review, 50*(2), 385–412. https://doi.org/10.1111/imre.12170

Bergqvist Rydén, J. (2022) Anonymiserade examinationer: En problematiserande forskningsöversikt. *Forskningsrapport beställd av fakultetsstyrelsens vid HT-fakulteterna arbetsutskott, Lunds universitet*.

Berlin, I. (1958) *Two concepts of liberty: An inaugural lecture, delivered before the university of Oxford on 31 october 1958.* Oxford: Clarendon Press.

Bertrand, M. & Mullainathan S. (2004) Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *The American Economic Review, 94*(4), 991-1013. https://doi.org/10.1257/0002828042002561

Burke, E. (1986/1790) *Reflections on the revolution in France: And on the proceedings in certain societies in London relative to that event* (C. C. O'Brien Ed.). London: Penguin.

Bursell, M. (2014) The multiple burdens of foreign-named men—evidence from a field experiment on gendered ethnic hiring discrimination in Sweden. *European Sociological Review, 30*(3), 399–409. https://doi.org/10.1093/esr/jcu047

Bygren, M. (2020) Biased grades? Changes in grading after a blinding of examinations reform. *Assessment & Evaluation in Higher Education, 45*(2), 292-303. https://doi.org/10.1080/02602938.2019.1638885

Bygren, M. & Gähler, M. (2021) Are women discriminated against in countries with extensive family policies? A piece of the "welfare state paradox" puzzle from Sweden. *Social Politics: International Studies in Gender, State & Society, 28*(4), 921–947. https://doi.org/10.1093/sp/jxab010

Correll, S.J., Benard, S. & Paik, I. (2007) Getting a job: Is there a motherhood penalty?. *American Journal of Sociology, 112*(5), 1297-1338. https://doi.org/10.1086/511799

Enoch, D., Spectre, L. & Fisher, T. (2012) Statistical evidence, sensitivity, and the legal value of knowledge. *Philosophy & Public Affairs, 40*(3), 197-224. https://doi.org/10.1111/papa.12000

Forscher, P. S., Lai, C. K., Axt, J. R., Ebersole, C. R., Herman, M., Devine, P. G., & Nosek, B. A. (2019) A meta-analysis of procedures to change implicit measures.

Journal of Personality and Social Psychology, 117(3), 522–559. https://doi.org/10.1037/pspa0000160

Grady, M.F. (2002) Proximate cause decoded. *UCLA Law Review, 50*, 293-335.

Hinnerich, B.T., Höglin, E., & Johannesson, M. (2011) Are boys discriminated in Swedish high schools?. *Economics of Education review, 30*(4), 682-690. https://doi.org/10.1016/j.econedurev.2011.02.007

Hinnerich, B.T., Höglin, E. & Johannesson, M. (2015) Discrimination against students with foreign backgrounds: Evidence from grading in Swedish public high schools. *Education Economics, 23*(6), 660-676. https://doi.org/10.1080/09645292.2014.899562

Hobbes, T. (1997/1651) *Leviathan* (R. E. Flathman & D. Johnston Eds.). New York: Norton.

Jönsson, M.L., & Bergman, J. (2022) Improving misrepresentations amid unwavering misrepresenters. *Synthese, 200*. https://doi.org/10.1007/s11229-022-03744-5

Kiss, D. (2013) Are immigrants and girls graded worse? Results of a matching approach. *Education Economics, 21*(5), 447-463. https://doi.org/10.1080/09645292.2011.585019

Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J. L., Joy-Gaba, J. A., Ho, A. K., Teachman, B. A., Wojcik, S. P., Koleva, S. P., Frazier, R. S., Heiphetz, L., Chen, E., Turner, R. N., Haidt, J., Kesebir, S., Hawkins, C. B., Schaefer, H. S., Rubichi, S., Sartori, G., Dial, C. M., Sriram, N., Banaji, M. R., & Nosek, B. A. (2014) Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. *Journal of Experimental Psychology: General, 143*(4), 1765-1785. http://dx.doi.org/10.2139/ssrn.2155175

Lavy, V. (2008) Do gender stereotypes reduce girls' or boys' human capital outcomes? Evidence from a natural experiment. *Journal of Public Economics, 92*(10), 2083-2105. https://doi.org/10.1016/j.jpubeco.2008.02.009

Lincoln, A. E., Pincus, S., Koster, J. B., & Leboy, P. S. (2012) The Matilda Effect in science: Awards and prizes in the US, 1990s and 2000s. *Social Studies of Science*, *42*(2), 307–320. https://doi.org/10.1177/0306312711435830

Locke, J. (1980/1690) *Second treatise of government* (C. B. Macpherson Ed.). Indianapolis, Ind.: Hackett Pub. Co.

Madva, A. (2020) Individual and structural interventions. In E. Beeghly & A. Madva (Eds.) *An introduction to implicit bias: Knowledge, justice, and the social mind*. New York: Routledge

McLaughlin, J.A. (1925-26) "Proximate cause". *Harvard Law Review, 39*(2), 149-199. https://doi.org/10.2307/1328484

Mill, J.S. (2008/1859) On liberty. In J. Gray (Ed.) *On liberty and other essays*. Oxford: Oxford University Press.

Moore, M.S. (2009). *Causation and responsibility: An essay in law, morals, and metaphysics*. Oxford: Oxford University Press.

Paluck, E.L., Porat, R., Clark, C.S., & Green, D.P. (2021) Prejudice reduction: Progress and challenges. *Annual review of psychology, 72*(1), 533-560. https://doi.org/10.1146/annurev-psych-071620-030619

Pettit, P. (1997) *Republicanism: A theory of freedom and government*. Oxford: Clarendon.

Quillian, L., Heath, A., Pager, D., Midtbøen, A.H., Fleischmann, F. & Hexel, O.(2019) Do some countries discriminate more than others? Evidence from 97 field experiments of racial discrimination in hiring. *Sociological Science, 6*, 467-496. https://doi.org/10.15195/v6.a18

Redmayne, M. (2008) "Exploring the proof paradoxes". *Legal Theory, 14*(4), 281 – 309. https://doi.org/10.1017/S1352325208080117

Skinner, Q. (1998) *Liberty before liberalism*. Cambridge: Cambridge University Press.

Tamblyn, R., Girard, N., Qian, C.J. & Hanley, J. (2018) Assessment of potential bias in research grant peer review in Canada. *Canadian Medical Association Journal, 190*(16), 489-499. https://doi.org/10.1503/cmaj.170901

Tellhed, U. (2023) Challenges to Reducing Social Bias: Predictions for a New Post Hoc Intervention. *This volume.*

Wollstonecraft, M. (1988/1792) *A vindication of the rights of woman* (C. H. Poston Ed.). New York: Norton.

Zschirnt, E. & Ruedin, D. (2016) Ethnic discrimination in hiring decisions: A meta-analysis of correspondence tests 1990–2015. *Journal of Ethnic and Migration Studies, 42*(7), 1115-1134. https://doi.org/10.1080/1369183X.2015.1133279