Edited by
Mattias Gunnemyr and
Martin L. Jönsson

Post Hoc Interventions

Prospects and Problems

Since prejudice harms so many, it is natural to think that it must be stopped at its source. That it must itself be removed, to stop its undesirable effects. Or, at the very least, that it must be contained, to thereby stop its manifestations. But what if we can't? What if prejudice proves too entrenched and its manifestations too difficult to stop?

Then we must intervene where we can. We must look downstream from the manifestations of prejudice, yet upstream from its undesirable outcomes, for places where the stream can be rerouted. This is the key insight that motivates the pursuit of post hoc interventions.

This volume collects nine contributions to the emerging literature on this novel approach to prejudice mitigation, composed by the members of a research incubator which took place in 2022 at the Pufendorf Institute of Advanced Studies in Lund, Sweden. It adopts a range of different perspectives, and draws on expertise in epistemology, ethics, law, psychology, statistics and computer science, as it explores various aspects of post hoc interventions.

Post Hoc Interventions: Prospects and Problems

Post Hoc Interventions

Prospects and Problems

Edited by
Mattias Gunnemyr and
Martin L. Jönsson

LUND
UNIVERSITY

The printing of this book was made possible by the Pufendorf Institute for Advanced Studies, Lund University.

Post Hoc Interventions: Prospects and Problems

Published by the Department of Philosophy, Lund University.

Edited by: Mattias Gunnemyr and Martin L. Jönsson

Cover layout by Cecilia von Arnold

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

Suggested citation: 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

Printed in Sweden by Media-Tryck, Lund University

Contents

PREFACE
Mattias Gunnemyr & Martin L. Jönsson 7

1. A BRIEF INTRODUCTION TO POST HOC INTERVENTIONS
Martin L. Jönsson 11

2. CHALLENGES TO REDUCING SOCIAL BIAS
Predictions for a New Post Hoc Intervention
Una Tellhed 21

3. SOME REFLECTIONS ON THE PRACTICAL APPLICABILITY OF GIIU
Jakob Bergman 47

4. POST HOC INTERVENTIONS IN CRIMINAL SENTENCING
An Empirical Thought Experiment
Erik J. Girvan 55

5. POST HOC INTERVENTIONS AND SWEDISH DISCRIMINATION LAW
Anna Nilsson 77

6. POST HOC INTERVENTIONS AND THE GENERAL DATA PROTECTION REGULATION
Martin L. Jönsson & Jonas Ledendal 93

7. THE ETHICS OF POST HOC INTERVENTIONS
Three Potential Problems
Mattias Gunnemyr 113

8. POST HOC INTERVENTIONS AND MACHINE BIAS
Kasper Lippert-Rasmussen 133

9. SIX WAYS OF FAIRNESS
Thore Husfeldt 155

Preface

This volume is the tangible result of a research incubator (a 'Pufendorf theme') that took place in 2022 at the Pufendorf Institute of Advanced Studies in Lund, Sweden. During this time, members of the theme – Martin Jönsson, Mattias Gunnemyr, Jakob Bergman, Erik Girvan, Thore Husfeldt, Kasper Lippert Rasmussen, Anna Nilsson, and Una Tellhed – met at the Pufendorf Institute, once a week, for a yearlong interdisciplinary conversation on post hoc interventions (see Chapter 1 for a brief introduction). It is a testament to the broad expertise of the research group (containing as it did, epistemological, ethical, statistical, computational, legal, and psychological expertise), to the open-mindedness of the experts, and the fertility of the subject matter, that so much progress could be made in such a short time.

Scientific progress involves both uncovering new *prospects* – identifying embryos to new interventions, places of application, ways to overcome obstacles – as well as uncovering new *problems* – identifying restrictions and limitations, and disheartening dead ends at the end of once promising paths. As its name implies, this volume contains examples of both kinds.

One problem – discussed by Una Tellhed in Chapter 2 of this volume – is that GIIU – the post hoc intervention most often in focus during the year – qua structural social bias intervention, is likely to meet various forms of resistance once introduced into an organization; people might, for instance, deny that social bias exists, trivialize discrimination and segregation issues, and distrust an algorithm that offers suggestions at odds with their personal opinions. Tellhed describes some of these forms of resistance and lists some suggestions on how to proceed.

In Chapter 3, Jakob Bergman begins by offering a prospect. He discusses the possibility of bias that works in a non-linear way and outlines a test to detect different types of biases of this kind, which might be used as a failsafe when applying GIIU. He also briefly discusses situations where candidates are

assessed on several dimensions, or by several evaluators, and shares some ideas on how to proceed in such cases.

In Chapter 4, Erik Girvan considers what would happen if we used GIIU to adjust for potential ethnic biases in criminal sentencing outcomes in the U.S. In particular, he considers whether GIIU's presuppositions could be satisfied in practice and argues that while the requirements may not all be satisfied in their strong form, they can likely be satisfied in many circumstances.

The Law, by its very nature, is restrictive, and it is thus unsurprising that post hoc interventions might face legal problems. In addition to improving the accuracy of evaluations, GIIU can also help to reduce discrimination. Doing so is permitted, and even prescribed, in most legal systems in liberal democracies. Still, many legal systems prohibit affirmative action such as the use of quotas or other preferential treatment, at least in some areas. While GIIU is importantly different from affirmative action, it is still important to decide whether it is different enough from a legal perspective. In chapter 5, Anna Nilsson examines the implications of Swedish and EU discrimination law for the use of post hoc interventions during recruitment processes. She argues that post hoc interventions such as GIIU might be thought of in two different ways: either as a tool that corrects for biases and prejudice or as a form of preferential treatment. If it is best characterized as a tool that corrects for biases and prejudice, there is nothing in the Swedish Discrimination Act that prevents an employer from using GIIU. There is, however, a risk that a Swedish court could find that the use of GIIU is a form of preferential treatment. If so, it would not be allowed to use GIIU to compensate for ethnic discrimination in Sweden. (Still, it would be allowed to use GIIU to compensate for prejudice against persons with disabilities and persons with transgender identity or expression.) Further, because of EU regulations, it would not be allowed to use GIIU to automatically update biased decisions when the bias concerns the sex of the applicants. Instead, in such cases, GIIU must be used as a decision support device.

Chapter 6 also discusses legal questions in relation to post hoc interventions. Martin Jönsson and Jonas Ledendal investigate the possible tension between GIIU and GDPR (the General Data Protection Regulation, which restricts data processing within the EU). They conclude that many applications of GIIU are in compliance with the GDPR, even without the specific consent for this processing of data subjects, but that others might not be, specifically those where the processing includes special categories of personal data that is considered sensitive.

The final three chapters concern the ethics of post hoc interventions, broadly construed. In Chapter 7, Mattias Gunnemyr considers three potential ethical problems in relation to post hoc interventions: that post hoc interventions might infringe on the decision makers freedom to make decisions in morally problematic way, that GIIU might indicate that a certain decision is biased even if it is not, and that GIIU might rely on probabilistic evidence that does not tell us anything about whether the decision at hand is biased. Gunnemyr concludes on a positive note that 1) 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 GIIU is implemented as decision support system, that 2) GIIU actually requires an additional presupposition, or should at least be constrained so that it is only applied in a specific way to avoid incorrect updates of evaluations, and that 3) post hoc interventions do not rely on probabilistic evidence in a problematic way.

In the US, courts have used the algorithmic tool COMPAS to assess potential recidivism risk. Critics have argued that the court's use of COMPAS unfairly disadvantages black offenders since it lacks calibration across groups. In particular, it generates higher false positive rates of predicted recidivism for black offenders than white offenders. Kasper Lippert-Rasmussen argues in Chapter 8 that we do not think that lack of calibration entails unfair bias in non-algorithmic contexts, such as hirings, and that we therefore should reject the view that calibration is necessary for fairness in an algorithmic context.

In the final chapter, Thore Husfeldt considers whether fairness requires calibration across groups, but from another perspective. He starts off formalizing some of our most common notions of fairness and shows that they are incompatible. For instance, we might think that fairness requires that group membership (in terms of sex, ethnicity, etc.) does not influence who gets recruited for a certain position, and that all groups should be represented in proportion to their part of the entire population (at a certain job, in parliament, etc.). These kinds of fairness are sometimes called "equal odds" and "democratic parity". Husfeldt shows that it is impossible to be fair in both ways simultaneously in non-trivial cases.

Most of the contributions in this volume (with the exception of chapters 3 and 6) were presented at a conference which shares its name on October 5th-6th 2022 in Lund. We are thankful to the participants of that conference for their many suggestions, questions and requests that indirectly helped improve this volume. We are particularly indebted to the invited commentators at the conference for their generous feedback. Nazar Akrami, Fredrik Björklund, Eric

Brandstedt, Leila Brännström, Jenny Magnusson, Boel Nelson, Maria Stanfors, and Frej Klem Thomsen. Thank you!

We would also be remiss if we forgot to thank everyone who has visited our weekly seminar during the theme's run: Nazar Akrami, Ramón Alvaro, György Barabás, Fredrik Björklund, Michaela Digneus, Jonas Ledendal, Fredrik Lindstrand, Alex Madva, Sara Martinsson, Gregor Noll, Julian Schuessler, Anders Sjöberg, András Szigeti, Kim Mannemar Sønderskov, and Joan Williams. Thank you for sharing your work and for tuning in to our conversation.

Moreover, we would like to thank everyone at the Pufendorf Institute in Lund for their financial support, and for giving the group the opportunity to spend its Wednesdays at their beautiful premises. Johanna Albihn, Ann-Katrin Bäcklund, Stephan Choquette, Eva Persson, Sune Sunesson, Stacey Ristinmaa Sörensen och Cecilia von Arnold. Thank you for all your help.

Last but not least, we would like to thank the other wonderful participants of the theme: Anna, Erik, Jakob, Kasper, Thore and Una. For your efforts, patience and insights. Thank you.

/ Mattias Gunnemyr and Martin Jönsson, January 2023

1

A Brief Introduction to Post Hoc Interventions

Martin L. Jönsson[1]

Abstract. The paper offers some background to the phrase 'post hoc intervention', and some associated concepts. It defines a narrow concept of a post hoc intervention, and illustrates it in detail by way of GIIU, a particular example of a post hoc intervention of this kind.

Introduction

Since it harms so many, it is natural to think that prejudice must be stopped at its source. That it must itself be removed, to stop its undesirable effects. Or, at the very least, that it must be contained, so that its manifestations are thereby stopped. But what if we can't? What if prejudice proves too entrenched and its manifestations too difficult to stop?

Then we must intervene where we can. We must look downstream from the manifestations of prejudice, yet upstream from its undesirable outcomes, for places where the stream can be rerouted. This is the key insight that motivates the pursuit of post hoc interventions.

Many manifestations of prejudice – e.g. violence and anti-social behaviour – are themselves undesirable. For these, there is no place to intervene, no gap between manifestation and undesirable outcome. But for many others a gap exists, e.g. the gap between prejudiced evaluation and discriminatory hiring, the gap between prejudiced grading and unfair admission, and the gap between prejudiced performance review and unfair promotion. Depending on which gaps we identify, different kinds of things can be done to intervene, and the phrase 'post hoc intervention' can thus be used to denote various kinds of activities.

[1] Martin L. Jönsson, Senior Lecturer in Theoretical Philosophy, Department of Philosophy, Lund University.

Compositionally, a *post hoc intervention* (PHI), just means 'an intervention after some event'. But the phrase can be given more substance by including the purpose of the intervention – a teleological component – as well as the targeted kind of event – an ontological component. The phrase can thus be used to denote, for instance, an *attempt to increase the accuracy of an evaluation* after that evaluation has been carried out (where the teleological component is an attempt to increase the accuracy of something, and the ontological component is an evaluation).[2] But the phrase can be restricted even further by including details about the causal history of the ontological component – an etiological component. The phrase can thus be used – in line with Jönsson (2022) and Jönsson and Bergman (2022) – to denote an attempt to increase the accuracy of an evaluation after that evaluation has been carried *out in an attempt to mitigate the (suspected) effects of human prejudice*. Call this the narrow concept of a PHI. A PHI in this sense is an attempt to remedy inaccuracies in already produced evaluations with a certain (suspected) causal history – human prejudice. As described above, these evaluations can be reviews of job applicants, the grades of a group of students, or a performance review – anything produced by a human which is accuracy-evaluable.

As this volume makes evident however, the idea that prejudice can be mitigated after it has manifested, has application beyond the reach of the narrow concept. For instance, in Lippert Rasmussen's contribution to this volume, a post hoc intervention is an attempt to mitigate undesirable outcomes of prejudice in general. This combines broad teleological, ontological, and etiological components, and this allows Lippert Rasmussen to coherently discuss the possibility of a PHI that improves differential false positive rates due to an algorithm, something that wouldn't make sense on the narrow concept of a post hoc intervention (since differential false positive rates are not evaluations). Combinations of other teleological, ontological, and etiological components will create other concepts, which affords other generalizations and applications.[3]

[2] As noted by Kasper Lippert Rasmussen (in personal communication), it is more accurate to say that the purpose of the PHI decomposes into a teleological and an ontological component rather than to identify the purpose with the teleological component since the purpose of few PHIs is merely to increase the accuracy of *something* (in contrast with a particular kind of thing).

[3] Not anything goes however since the three components conceptually constrain each other. For instance, for it to be possible to increase the accuracy of something *X*, *X* has to be such that it can be accurate or inaccurate. The teleological and ontological components thus constrain each other.

The term 'post hoc intervention' as means to express the narrow concept was introduced by Jönsson in a grant application from 2017, but the underlying idea came from an earlier article by Jönsson and Sjödahl (2017). However, Jönsson and Sjödahl didn't use 'post hoc intervention' explicitly and operated with an even narrower concept, since they were primarily concerned with improving accuracy ('veracity' in their terminology) of *implicitly biased rankings* (rather than *prejudiced evaluations*). They thus used narrower ontological and etiological components than the narrow concept. In hindsight, this further restriction seems unnecessary.

It is likely that the narrow concept was introduced when the phrase for it was coined, but its origin is somewhat obscure, since it at the very least bears family resemblance to several other concepts of independent origin. First and foremost, there is affirmative action such as the use of quotas. These could be understood, for instance, as *attempts to increase diversity in a group, after selecting its members*, and thus understood as PHIs in this sense. Strictly speaking though, quotas are typically applied mid-selection rather than after the selection has taken place.[4] But if this is ignored, quotas can even be understood as crude PHIs in the narrow sense given certain affirmative action rationales (cf. Anderson 2010: ch. 7 on the 'discrimination blocking rationale'). Second, there have been interventions that try to counteract the perceived bias of tests, by handling test scores for various social groups differently. For instance, the Medical University of Vienna evaluated the aptitude scores (for an 'EMS'-test) differently for men and women applying to study medicine after identifying what was believed to be a gender bias of the test itself (see Winkler-Hermaden 2012). Although this method was not aimed at counteracting the prejudice of a prejudiced person (but the perceived bias of a test), and thus not a PHI narrowly construed, it is clearly a PHI on a related construal (with a different etiological component). Similarly, among the machine learning technologies concerned with mitigating bias and increasing fairness (on which research intensified towards the end of the 2010s) there is a small subsection concerned with 'post-processing methods' which have clear affinities with PHIs (see Caton and Haas 2020 for an overview). These methods attempt to offset algorithmic bias rather than human prejudice but, again, are clearly PHIs, if these are understood slightly differently. Finally, there is research on the statistical moderation of school-based assessment (see e.g. Williamson 2016 and Thulasidas 2021) that relate to PHIs. For instance, it

[4] I'm indebted to Mattias Gunnemyr for stressing the timing of the two kinds of interventions.

is suggested in a recent report from the Swedish National Agency for Education (Skolverket 2020: 7), that one way to come to terms with (non-prejudicial) inequalities between the grading in different Swedish schools is through the (post hoc) statistical moderation of grades informed by students' results on standardized tests. Again, this a process that would also clearly count as a PHI, if this phrase was used in a slightly different way than the narrow one (by using comparability rather than accuracy in the teleological component and by dropping the etiological component entirely).

However construed, PHIs contrast with most existing prejudice interventions, which attempt to prevent prejudice from manifesting, so called *ante hoc* interventions, in Lippert Rasmussen's terminology. In particular, if we limit ourselves to evaluation-focused interventions, these interventions try to *prevent* evaluators from evaluating prejudicially. Such interventions include both *structural interventions* that change the circumstances under which the evaluation takes place (e.g. through the introduction of anonymization, or through criteria–based decision making) and thereby decrease prejudiced behavior, and non-structural *individual interventions* that attempt to change the evaluator (e.g. by the evaluator undergoing debiasing training, or being exposed to increased intergroup contact) and thus make *them* less prejudiced (cf. Madva, 2020).[5]

What initially attracted Jönsson and Sjödahl (2017) to narrowly understood PHIs, (henceforth 'PHIs' tout court), was the promise they saw in PHIs to constitute concrete and direct countermeasures to prejudice that both avoids many of the standard objections to quotas (e.g. weakened meritocracy, see Lippert Rasmussen 2020), and circumvents the inertia of prejudiced people that is apparent from the literature on individual ante hoc interventions (cf. Paluck; Lai et al 2017).

To explore the feasibility of PHIs, Jönsson and Sjödahl (2017) introduced and discussed three varieties: SOU (Solipsistic ordinal-scale update), MIRU (Multiplicative Informed Ratio-Scale Update), and GIRU (Generalized Informed Ratio Scale Update). They concluded that only the latter PHI had hopes of increasing accuracy systematically and proposed a number of conditions they argued should be sufficient for it to work as intended.

A few years later Jönsson and Bergman (2022) refined and revised these conditions into the following list:

[5] These two kinds of interventions correspond respectively to attempts to remove and contain prejudice mentioned in the opening paragraph.

G1. Evaluations use, minimally, an interval scale.

G2. The history of evaluations is large enough to reliably find prejudices with a suitable statistical test.

G3. The mean values in the relevant populations of whatever is being evaluated are known (or can be estimated), or are known to be the same.

G4. GIIU makes use of subsets of the groups the evaluator is prejudiced against.

G5. Any fluctuations in the Evaluator's prejudice are small compared to the size of the corresponding prejudice.

G6. The evaluator's prejudice operates in an approximately linear way.

G7. The evaluator's prejudice operates on discrete groups.

They then used a computer simulation to show that these were in fact sufficient to to improve accuracy in the face of a wide variety of forms of prejudice, in an overwhelming majority of cases.[6] They also noted that GIRU should be renamed GIIU (Generalized Informed *Interval* Scale Update) to mark that the intervention doesn't require that the evaluations are made on a ratio scale – as assumed by Jönsson and Sjödahl – but merely an interval scale. G1 is thus a weaker condition than Jönsson and Sjödahl suggested.

Although GIIU is just one option in the logical space of PHIs, it is the one that is in focus in this volume, and to make the rest of the volume more accessible, GIIU will now be illustrated with an example (taken from Jönsson and Bergman 2022).

GIIU

GIIU is meant to be applied to, and thus improve the accuracy of, a set of numerical competence evaluations such as r_0 – which we will call *the target evaluation* – that an evaluator *E* has produced for the purpose of ranking people in terms of their suitability with respect to some end (such as hiring, or promoting, or for some other purpose).

[6] See Jönsson and Bergman (2022) for a detailed discussion of the seven conditions.

r_0		r_0*	
Mike	8	Mike	8
Mark	7	Mark	7
Felicia	6	Felicia	9
Gordon	4	Gordon	4
Sarah	3	Sarah	6
Latifah	3	Latifah	6

The number assigned to each person is *E*'s estimate (on some interval scale) of how competent that person is. For purposes of illustration we assume that *E* is prejudiced – sexist – and that the real competence scores are given by r_0*.

To determine whether GIIU needs to update r_0, GIIU first consults *E'*s history of evaluations, a set of sets just like r_0 consisting of evaluations that *E* has previously made. In *E*'s history of evaluations, GIIU checks the mean scores of members of a number of *target groups* – groups that *E* might be prejudiced against (e.g. groups such as men and women, or different ethnic groups). GIIU then compares *E'*s group means with the means one would expect to find if the *populations corresponding to the target groups* are assumed to be identical (or assumed to differ to a certain known degree) in terms of their distribution of competence (or whatever property one is interested in). In the absence of an alternative explanation, GIIU then proceeds to assume that *E* is prejudiced if there is a statistically significant difference between the mean competence scores of the target groups in the history of evaluations (or if they differ significantly more than the mean competence scores in the corresponding populations). It then identifies a set of corrective functions *v* such that each corrective function individually makes the means in the target groups in the history of evaluations come out the same (or differ to the same degree as the corresponding populations). GIIU then suggests that r_0 should be updated *along the lines of what the corrective functions in v jointly ordinally agree on.*

To make this more vivid, consider the following illustration from Jönsson and Sjödahl's article, where *E'*s history of evaluations is assumed to contain three sets of competence evaluations, h_1, h_2 and h_3, and we are considering whether to update the target evaluation r_0.

h_1		h_2		h_3	
John	8	Mike	7	Brittney	6
Luke	7	Catherine	6	Jamal	5
Sarah	4	Billy	6	Richard	3
Amber	4	Richard	5	Susan	3
Isa	3	Jennifer	4	Aaliyah	3
Jenny	3			Molly	2

If we assume that the target groups are men and women, we can calculate the mean competence scores in the history of evaluation for these groups, which are 5.9 and 3.8 respectively. Given that the difference between the mean scores is statistically significant, and that there is no difference between men's and women's population means, GIIU concludes that *E* is prejudiced and that r_0 needs to be updated. One way to do this (which would remove the mean difference between men and women in the history of evaluations if the function is applied to it) is by adding 2.1 to the score of each woman in r_0 (Felicia, Sarah and Latifah). One corrective function f_1 thus corresponds to '$f_1(x) = x+2.1$'. Another way to update r_0 is to multiply each score by 1.55, and '$f_2(x) = x*1.55$' thus corresponds to another function f_2. If either of these functions is applied to the scores in the history of evaluations, the difference in men's and women's means would disappear. Other corrective functions might have the same result. GIIU then suggests that each such function is individually applied to the values in the target evaluation r_0, resulting in *n* different rankings $f^1(r_0)$, $f^2(r_0)$, etc., one for each corrective function f_i. If we assume that f^1 and f^2 are the only relevant corrective functions in this example, we get the following result:

$f^1(r_0)$		$f^2(r_0)$	
Felicia	8,1	Felicia	9,3
Mike	8	Mike	8
Mark	7	Mark	7
Sarah	5,1	Sarah	4,7
Latifah	5,1	Latifah	4,7
Gordon	4	Gordon	4

To the extent that these rankings converge on the same ordinal results, GIIU recommends that r_0 should be updated accordingly. In this case, GIIU recommends that r_0 should be updated so that Felicia is ranked first, and so that Sarah and Latifah are ranked before Gordon.

In this toy example, this will clearly improve the accuracy of r_0, as indicated by the new ranking being perfectly rank-order correlated with the correct ranking r_0*.

Acknowledgments

The research was funded by a research grant from the Swedish Research Council (Dnr. 2017-02193) and research funding provided by the Pufendorf IAS in Lund. I'm indebted to Åsa Burman, Mattias Gunnemyr, Kasper Lippert Rasmussen, and Lisa Zetterberg for helpful comments on the text or its subject matter.

References

Anderson, E. (2010). *The Imperative of Integration*. Princeton University Press.

Caton, S., & Haas, C. (2020). Fairness in Machine Learning: A Survey. *ArXiv, abs/2010.04053*.

Jönsson, M. L. (2022). On the Prerequisites for Improving Prejudiced Ranking(s) with Individual and post hoc Interventions" *Erkenntnis*.

Jönsson, M. L. and Bergman, J. (2022). Improving misrepresentations amid unwavering misrepresenters, *Synthese*, *200*.

Jönsson, M. L. and Sjödahl, J. (2017). Increasing the veracity of implicitly biased rankings, *Episteme 14*(4), 499–517.

Lippert-Rasmussen, Kasper (2020). *Making Sense of Affirmative Action*. Oxford University Press.

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.

Skolverket (2020) *Analyser av likvärdig betygssättning i gymnasieskolan.*

Thulasidas, M. (2021) Statistical moderation: A case study in grading on a curve. *2021 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE): Wuhan, December 5-8: Proceedings*. 734-739. Research Collection School of Computing and Information Systems.

Williamson, J. (2016) Statistical moderation of school-based assessment in GCSEs. *Research Matters: A Cambridge Assessment publication*.

Winkler-Hermaden, R. (2012) Medizin-Uni Wien: Frauen werden bei Aufnahmetest milder beurteilt**.** https://www.derstandard.at/story/1331207289145/medizin-uni-wien-frauen-werden-bei-aufnahmetest-milder-beurteilt, Accessed 2023-01-22.

2

Challenges to Reducing Social Bias

Predictions for a New Post Hoc Intervention

Una Tellhed[1]

Abstract. This chapter discusses a new tool for social biases interventions in work-contexts called Generalized Informed Interval Scale Update (GIIU) from a social psychological perspective. The aims are to categorize GIIU in relation to previous social bias interventions in the literature and analyze potential challenges that it may face when implemented in work contexts. Conclusions include that GIIU is a structural social bias intervention and as such, will likely meet predominately challenges that are motivational in character (the so called will not - challenge).

Aims of This Chapter

This chapter is included in an anthology that discusses a new tool for social bias interventions in work-related judgments, such as decisions for recruitment, promotion, and resource distribution. The social bias invention is called Generalized Informed Interval Scale Update (GIIU, Jönsson, 2022; Jönsson & Bergman, 2022; Jönsson & Sjödahl, 2017). Its method is described in the first chapter to this anthology, but I will also briefly describe it here, to enable a freestanding reading of this chapter. GIIU investigates large data materials of quantitative work-related judgments (e.g. on a 1-5 scale). An algorithm tests for mean differences related to social group categories in the material. For example, GIIU can detect if an assessor has systematically rated

[1] Una Tellhed, Senior Lecturer at the Department of Psychology, Lund University.

women's job performance as significantly lower than men's in a data set. GIIU also calculates to what degree the quantitative rating of the underrated social group could be raised, to adjust for the detected systematic difference. Assuming that performance is the same or known to be different with a certain magnitude in the two populations, GIIU can thereby correct bias "post hoc" and is therefore classified as a post hoc intervention (Jönsson, 2022; Jönsson & Bergman, 2022; Jönsson & Sjödahl, 2017).

This anthology takes an interdisciplinary approach to analyzing GIIU, with authors from Philosophy, Psychology, Law, Computer Science and Statistics. The current chapter analyzes GIIU from a social psychological perspective. The aim is to describe previous social bias interventions in the literature and reflect upon how GIIU fits within this literature. The main focus is to analyze common challenges that social bias interventions meet, that limits their effectiveness. Based on this previous research, I will make predictions for the type of challenges that GIIU may meet when it is implemented in organizations. I will also present some ideas for how future psychological research can investigate the implementation of GIIU. But first, I will describe what GIIU and other social bias interventions are designed to combat, namely the segregated labor market, and its roots in social bias.

The Segregated Labor Market

The Swedish labor market is strongly segregated according to sex/gender and ethnicity (Allbright, 2022; European Institute for Gender Equality, 2017; Nordic Council of Ministers, 2022; Tellhed, 2022; Wolgast & Wolgast, 2021). The lion share of research has focused on these social categories (gender and ethnicity), and therefore, so will I in this text. However, the Swedish Discrimination Act concerns the following seven social categories which are relevant for segregation in the labor market; sex, transgender identity or expression, ethnicity, religion or other belief, disability, sexual orientation and age (Government Offices of Sweden (2008:567).

The labor market is segregated both "vertically" and "horizontally" (European Commission, 2009; 2014). Vertical segregation refers to the circumstance that some social groups are disproportionally represented, relative to their population statistics, in positions of high power and status

(European Commission, 2009; 2014). For example, "white"[2] men are overrepresented at high power positions in the labor market (Allbright, 2022, Statistics Sweden, 2022). Concerning overall employment, men and women are employed in equal numbers in Sweden, but people with an immigrant background are employed to a much lower degree as compared to people born in Sweden (Statistics Sweden, 2021).

The labor market is also horizontally segregated, which means that social groups are disproportionally represented in different types of occupations. For example, more men than women work in STEM (Science, Technology, Engineering, and Mathematics) and more women than men work in HEED (Health care, Elementary Education and Domestic (Block, et al., 2018; Nordic Council of Ministers, 2022; Tellhed, 2022). This segregation is also partly related to status, where STEM-occupations have higher status than HEED-occupations (Croft et al. 2015; Svensson & Ulfsdotter Eriksson 2009).

The Role of Social Bias

Much research has focused on exploring why labor segregation emerges and persists. The answer is complex and outside the scope of this text, (see Tellhed, 2022 for an overview of explanations for the horizontal gender segregation), but research has shown that social bias plays one part in it (see Caleo & Heilman, 2019; Williams, 2021; Wilson, 2017; Wolgast & Wolgast, 2022 for reviews). Social bias has been defined as a

> "…systematic tendency to evaluate one's own membership group (the ingroup) or its members more favorably than a nonmembership group (the outgroup) or its members" (Hewstone et al., 2002).

Social bias is expressed as stereotypes, prejudice and discrimination, which represent cognitive, affective and behavioral aspects of bias (Hewstone et al., 2002). For example, it may include the belief that ingroup members have more meritorious qualities than outgroup members (i.e. stereotypes), having

[2] See Wolgast & Wolgast (2022) and Åkerlund (2022) for descriptions of how race/ethnicity is perceived in a Swedish context and the commonly perceived overlap of "whiteness" with "Swedishness". Also, official statistics in Sweden only registers the categories "Swedish" contra "Foreign" background in demographic statistics (Statistics Sweden, 2002), which is a crude categorization as compared to race/ethnicity categorization in for example the USA.

negative attitudes or feelings towards outgroup members (i.e. prejudice), and subsequently treating ingroup members more favorable than outgroup members (i.e. discrimination and /or ingroup favoritism, Gilovich, et al., 2019; Hewstone et al., 2002). Discrimination is the behavioral aspect of the bias spectrum and is regulated in law (see chapter 5 in this anthology). Applied to a work context, examples of discriminatory behavior are selecting individuals out in the recruitment process based on their social group membership, overlooking outgroup members for promotions, or awarding outgroup members lesser rewards than ingroup members. Importantly, discrimination is more strongly related to in-group favoritism than outgroup derogation (Brewer, 1999; Greenwald & Pettigrew, 2014).

Social bias can be intentional and include conscious elements, but it can also operate without conscious access and lacking ill intent (Gawronski & Payne, 2010). Especially, when under high stress, people tend to behave automatically (without reflection or the conscious experience of intent) and may base judgments partly on stereotypes and prejudice, without taking notice (Gawronski & Payne, 2010; Pendry, & Macrae, 1994).

Social bias contributes to maintaining structural inequalities in the labor market, when its current gatekeepers select, promote and reward members of their own groups to a larger extent than outgroups. Because social bias perpetuates structural inequalities and causes career obstacles for underrepresented groups, much research has been dedicated to finding ways to mitigate social bias (e.g. Caleo & Heilman, 2019; Lai et al., 2014; Paluck & Green, 2009; Paluck et al., 2021; Williams, 2021). There are a multitude of interventions designed to reduce social bias, where some have focused specifically on increasing social diversity (i.e. reduce segregation) in the labor market (Caleo & Heilman, 2019; Lai et al., 2014; Paluck & Green, 2009; Paluck et al., 2021; Williams, 2021). I here categorize these efforts as "psychological" versus "structural" social bias interventions and will discuss their effectiveness and where GIIU fits as a new addition to social bias interventions.

Psychological Social Bias Interventions

I define psychological social bias interventions as actions which aim to change individuals' psychology in some respect, with the explicit goal to make people less biased. "Psychology" is generally described as the study of humans' thoughts (cognition), feelings (affect), and behavior (Holt et al., 2019).

Correspondingly, psychological social bias interventions aim to reduce individuals' stereotyping (cognitive bias) or prejudice (affective bias), which is then assumed to cause reductions in discrimination (behavioral bias).

Effectively reducing individuals' social bias has proven challenging. Empirical tests of psychological social bias interventions show at best moderately reduced bias post-intervention (Caleo & Heilman, 2019; Paluck & Green, 2009; Paluck et al., 2021). The longevity of the effects is rarely tested, and few studies examine behavioral outcomes (e.g. discrimination, Caleo & Heilman, 2019; Lai et al., 2017; Paluck & Green, 2009; Paluck et al., 2021). One salient concern is that some psychological social bias interventions have been found to *increase* social bias in individuals. I will describe examples of common social bias interventions and discuss their effectiveness. The categorization of the interventions is based upon previous psychological work (Caleo & Heilman, 2019; Paluck & Green, 2009; Paluck et al., 2021) and it should be noted that the concepts represent "fuzzy sets" rather than precise definitions (Mc Closkey & Glucksberg, 1978).

Diversity Training

Sociologists Frank Dobbin and Alexandra Kalev have repeatedly warned that the most common types of social bias interventions in North American work organizations, called "diversity training", tend to have no effect, or even increase bias and segregation in organizations (e.g. Dobbin & Kalev, 2016; 2021). "Diversity training" is a fuzzily defined concept and its content varies, but one common element is informing participants about discrimination and its economic and legal consequences (Dobbin & Kalev, 2016; 2021; Paluck et al., 2021; Williams 2021). Some have proposed that shaming or threatening messages in diversity training may cause participants to react negatively to the intervention, and thereby limiting its effect (Dobbin & Kalev, 2016; 2021; Flood et al., 2021; Wiggins-Romesburg & Gibbins, 2018; Williams, 2021). Also, after being told not to discriminate certain target groups, some individuals from more privileged groups (such as white people or men) may believe that there is "reverse discrimination" in their organization, and that their ingroup is now disadvantaged relative to the groups targeted by the intervention (Dobbin & Kalev, 2016; 2021; Flood et al., 2021; Wiggins-Romesburg & Gibbins, 2018; Williams, 2021). This perception may instigate a resistance towards diversity work and increase social bias for these individuals, which I will elaborate more on later.

Although the weak results from decades of diversity training is disheartening, Joan Williams (2021) has pointed out that a new generation of diversity training programs (e.g. "habit breaking" or "bias interrupters") show promising effects in reducing bias and increasing diversity in the workplace. These new programs teach about the psychology behind social bias, and one important difference, as compared to the previous interventions, is that it encourages staff members to come up with their own ideas for breaking bias. Enabling autonomous thinking in an intervention is more intrinsically motivating than being told not to discriminate (Devine et al., 2012; Williams, 2021).

A final thought about diversity training is that despite being taught what social bias is, and trying our best to control it, we may still discriminate against others outside of awareness (Gawronski & Payne, 2010; Pendry, & Macrae, 1994). This means that although education is important, it is not sufficient to combat segregation. There is however evidence suggesting that people who believe that they are not biased tend to discriminate more than others (Begeny et al., 2020; Régner et al., 2019). Thus, learning that social bias is common and that we all may be biased to some extent should be important. However, to complicate things further, learning that implicit (automatic) bias is common may also strengthen individuals' bias, based on the logic that if everyone stereotypes, it is the norm (Duguid & Thomas-Hunt, 2015). This suggests that education about social bias should also try to motivate people to control their bias. Research suggests that implicit bias is controllable to some extent (Calanchini et al., 2021), but control has limitations, which I will discuss more later.

Contact Interventions

A well-researched type of psychological social bias interventions is contact interventions (Allport, 1954). Studies in this field have traditionally arranged for people from different social groups (typically ethnic groups) to meet and collaborate toward some common goal, preferably while on equal standings (Jones, & Rutland, 2018; Paluck & Green, 2009; Paluck et al., 2021). Contact interventions have been shown to moderately reduce prejudice, and the effect is related to increases in perspective taking and reduced intergroup anxiety (Aberson & Haag, 2007; Jones, & Rutland, 2018; Paluck & Green, 2009; Paluck et al., 2021).

Recently, researchers have expanded contact interventions to also include "extended" or "imagined" contact. In these interventions, participants do not meet in real life, but simply read or watch material where an ingroup member is described as positively interacting with an outgroup member (Jones, & Rutland, 2018; Paluck et al., 2021). Even this minimal research design has shown prejudice reducing effects, particularly strong with children as participants, although some of the more impressive results have failed to replicate (Paluck et al., 2021).

Applied to work settings, contact interventions suggest that working collaboratively and on equal terms in diverse work teams, should reduce staff member's social bias (Dobbin & Kalev, 2016). However, one limitation is that it demands that the organization is already diverse enough to allow for diverse work groups. Also, contact theory states that for contact to effectively reduce prejudice, group members should have equal status in the collaboration (Allport, 1954), which contrasts the current vertical segregation in the labor market that I previously described.

Social Categorization Interventions

Another common type of social bias intervention aims to alter individuals' social categorization (Paluck & Green, 2009; Paluck et al., 2021). For example, instead of categorizing others as outgroup members ("us" versus "them"), we can try to see others as unique individuals (i.e. a lower-level categorization) or as members of a common ingroup (i.e. a higher-level categorization, e.g. "We are all employed by this company", Paluck & Green, 2009; Paluck et al., 2021).

These types of interventions have mostly been tested in laboratory settings and tend to show effects on both implicit (indirect) and explicit measures of prejudice, although the effect sizes are typically small (Fitzgerald et al., 2019; Paluck & Green, 2009; Paluck et al., 2021).

Relatedly, interventions may also attempt to change group stereotypes by displaying counter-stereotypical examples of outgroup members (e.g. Calanchini et al., 2021; Fitzgerald et al., 2019; Lai et al., 2014; Paluck & Green, 2009; Paluck et al., 2021). However, one limitation is that when we meet an individual that counter a stereotype (such as woman with a successful career in tech), we may perceive this outgroup member as an exception, or "subtype" them, and thus preserve our stereotype intact (Kunda, & Oleson, 1995).

This implies that we need to encounter a (sufficiently) high number of outgroup members in counter-stereotypical work roles, to permanently alter

stereotypes. One interesting study showed that beginner college students associated men more than women with leadership on a computerized stereotype measure (the Implicit Association Test, Dasgupta & Asgari, 2004). However, in a follow up measure one year later, students that had encountered many women professors during their college year, now associated leadership equally strong with women as with men (Dasgupta & Asgari, 2004). This suggests that stereotypes may change when representation changes, but with the current segregation in the labor market, counter-stereotypical examples are still rare for many work roles.

Another problem with social categorization interventions has been raised in the literature of "color-blind racism" (e.g. Dovidio et al., 2015; Whitley Jr., & Webster, 2019). When majority group members claim to not "see ethnicity/race", they may have good intentions (e.g. aspire to not be racist, Whitley Jr., & Webster, 2019). However, this strategy may have adverse effects for minority groups, which is why it is called color-blind "racism". For example, John Dovidio and colleagues (2015) describe how color-blind approaches can create an "illusion of harmony", where attention is distracted away from existing bias structures, while the discrimination of ethnic minority groups continues.

Social Influence Interventions

The last type of psychological social bias intervention I will discuss utilizes social influence. Research has shown that we are quite easily influenced by the opinions of our ingroup-group members, especially if they have high status, and that their comments can affect how much prejudice we express (Munger, 2017; Paluck & Green, 2009; Paluck et al., 2021; Zitek & Hebl, 2007).

Studies have tested how we are influenced by our peers with experimental design (see Paluck et al., 2009; 2021 for reviews). One interesting study used twitter bots (a software program that fakes a twitter account, Munger, 2017) to vary peer influence. The twitter bot wrote a message to white men on twitter, that had just used the n-word. The message read: "Hey man, just remember that there are real people who are hurt when you harass them with that kind of language". When the bot portrayed a white man with many followers (high status), the users reduced their use of racist slurs. As a contrast, when the twitter bot portrayed a black man with few followers, the white men on twitter increased their racist language (Munger, 2017). Studies like this indicate that it is important that leaders and members of more privileged groups actively

take part in the quest to reduce social bias in society; work that has mostly been performed by women and ethnic minorities (e.g. Caleo & Heilman, 2019).

Applying the social influence approach to working life has several limitations though. It takes courage to stand up against coworkers who express prejudice, and it may unfortunately come with a price to take part in company diversity work, due to the "resistance", that I will discuss more later (e.g. see Caleo & Heilman, 2019; Dobbin & Kalev, 2016 for overviews).

It is also important to realize that social influence can be utilized (sometimes intentionally) to both increase as well as to decrease prejudice and inequality (Bates, 2020; Zitek & Hebl, 2007). This implies that when racist and sexist expressions are accepted in the workplace and "political correctness" is ridiculed, social bias and segregation is likely to increase.

Structural Social Bias Interventions

The research on psychological social bias interventions teaches us that it is not impossible to reduce individuals' social bias, but that the current methods have limitations. To effectively reduce segregation in the labor market, psychological social bias interventions may be combined with "structural" social bias interventions. Structural social bias interventions are not designed to reduce individuals' social biases per se. Instead, they can be defined as actions which change recruitment or promotion structures in some respect, with the aim to make assessors' biases less pervasive for organizational diversity outcomes. Actions can include working towards ensuring that judgments are objective and based on merit, and/or compensating for current or historical biases that certain groups have encountered. I will exemplify two commonly described structural bias intervention methods in the literature.

Systematic Recruitment Process

Research in work and organizational psychology has developed systematic recruitment strategies which minimize the reliance on gut feeling (that is prone to bias) in decision making (Ryan & Ployhart., 2014). It is beyond the scope of this text to describe this vast field of research, but recommendations include using evidence-based test instruments (e.g. ability- and personality tests), and structured interviews in recruitment, and to analytically weigh the results from the evaluation factors into an overall judgment (Ryan & Ployhart., 2014).

When possible, anonymizing applicants further reduces the risk of social bias in judgments.

Although much research has shown that applying these recruitment methods maximizes performance outcomes in organizations, research on the "science practitioner gap" has shown that recruiters and managers are often hesitant to implement it (Highhouse, 2008; Neumann et al., 2021). Reasons include a reduced sense of autonomy over the recruitment process and limited possibilities to display one's competence as a skillful recruiter/manager (Highhouse, 2008; Neumann et al., 2021).

The systematic recruitment methods also have limitations for the specific goal of increasing diversity, which is the focus of this chapter. Past (historical) disadvantages and discrimination that minority groups have encountered means that applicants from underprivileged groups sometimes have fewer merits, such as in education or work experience, as compared to more privileged groups (Rupp et al., 2020). Additionally, psychological phenomenon such as "stereotype threat", means that negatively stereotyped groups may underperform on cognitive tests, due to the association of negative ability stereotypes with their ingroup (e.g. see Spencer et al., 2016; Wilson, 2017, for reviews). Previous suggestions for resolving this dilemma include weighting recruitment criteria (Rupp et al., 2020) or using quotas, to ensure representation from all target groups (Jones, et al., 2021; Roos, et al., 2020), which I will discuss next.

Affirmative Action and Quotas

In Sweden, "positive" action (i.e. affirmative action) to combat segregation in the labor market is only legal when competing candidates have equal merits, and only on the basis of sex/gender (The Equality Ombudsman, 2022a). The Swedish parliament has also voted against the EU proposal to introduce mandatory gender quotas in corporate boards (Europaportalen, 2022). Proponents for using quotas or positive/affirmative action argue that it is an efficient method to rapidly decrease segregation and that it may be necessary to compensate for historical discrimination (Jones, et al., 2021; Roos, et al., 2020). Opponents' arguments include that it reduces corporations' autonomy and implies decreased meritocracy (e.g. Jones, et al., 2021; Roos, et al., 2020). It has further been pointed out that quotas are also limited by the social categories it targets (e.g. sex/gender), fails to recognize intersectional power

relations (e.g. how ethnicity interacts with gender) and may exclude non-binary individuals (Roos, et al., 2020).

Research has also investigated consequences for individuals recruited by quotas. One experimental study showed that when women were appointed leadership roles based on gender quotas, it lowered their perceived sense of competence, as compared to women that were recruited based on merit (Heilman et al., 1991). This problem did not occur for men that were recruited by gender quotas, possibly due to society's tendency to associate high competence with men (Storage et al., 2020). On the other hand, the same study showed that women sustained their sense of competence when recruited on quotas, if given confirmation that they have the right merits for the job (Heilman et al., 1991). This research suggests that it is important to communicate to those recruited or promoted by quotas (and to their co-workers) that they fulfill predetermined criteria for the position.

My description of psychological and structural social bias interventions has pointed to limitations with both types of interventions. I see common themes in these limitations and will next describe how they may relate to two categories of psychological functions: one motivational and one cognitive.

The "Will not" Challenge

The motivational challenge that social bias interventions face boils down to different forms of resistance towards social bias interventions and diversity work (Faludi¸1992; Wiggins-Romesburg & Githens 2018). This type of challenge affects both psychological and structural social bias interventions. Resistance towards social bias interventions has been described as one aspect of the broader term "backlash", which is defined as resistance towards progressive social changes (Faludi¸1992; Flood et al., 2021). Much research has shown that efforts to counteract discrimination, reduce segregation and increase diversity in organizations tend to face objections, which makes progress towards increasing diversity slow (e.g. Flood et al., 2021; Lansu, et al., 2020; Wiggins-Romesburg & Githens 2018¸ Wilson, 2017).

Examples of identified diversity resistance strategies include denial that social bias exists or claims of reverse discrimination (that the majority group is discriminated), victim blaming or trivialization of the segregation issues, passivity and lack of engagement in anti-discrimination efforts, hidden or overt attempts to undermine anti-discrimination work, and even harassment, aggression and violence against feminists and anti-racists (Bates, 2020; Flood

et al., 2021; Jones, et al., 2021; Tildesley, et al., 2021; Wiggins-Romesburg & Githens, 2018; Wilson, 2017; Åkerlund, 2022).

Resistance to diversity work is mostly performed by individuals from normative or numerical majority groups, in organizations, such as white men (Flood et al., 2021; Williams, 2021; Åkerlund, 2022). This circumstance has been related to power motives (e.g. social dominance) and a sense of aggrieved entitlement (Flood et al., 2021; Sidanius & Pratto, 1999; Tildesley, et al., 2021; Wiggins-Romesburg & Githens 2018; Wilson, 2017). The argument is that some individuals from more privileged groups perceive that their ingroup benefits from preserving the status quo, rationalize inequalities and feel threatened when they perceive that social hierarchies are changing (Dover et al., 2016; Flood et al., 2021; Tildesley, et al., 2021; Wiggins-Romesburg & Githens 2018¸ Wilson, 2017).

Another strain of research describes how resistance to social bias interventions and diversity work may also stem from autonomy motivation, where some individuals reject others influencing their decision making, for example in recruitment (Dobbin & Kalev, 2016; Jones, et al., 2021; Williams, 2021). This relates to research on the science practitioner gap that I described above (Highhouse, 2008; Neumann, et al., 2021). It may also be classified as a form of power motivation (desire to control outcomes), but is more individualistic in nature, as compared to the motivation to preserve ingroup privilege.

There is a lack of research into what strategies may effectively reduce dominance-motivated resistance to social bias interventions, especially when it originates from high-power individuals (e.g. managers) in an organization (Wiggins-Romesburg & Githens 2018). If leaders of an organization passively allow for resistance to diversity work, or actively participate in it, reducing segregation is more difficult (Flood et al 2021; Lansu, et al., 2020). This implies that involving leaders and high-status individuals in diversity work is important to mitigate the resistance. Especially motivating white men with high status to participate in the diversity work should help diminish resistance from other white men in the organization, since research shows that we are mostly influenced by high-status ingroup members (e.g. Caleo & Heilman, 2019; Munger, 2017; Paluck et al., 2021). Involving employees to help find solutions to diversity problems, rather than telling them what to do, should also reduce resistance motivated by autonomy motivation (Williams, 2021). I will next turn to the other major type of limitation I see in social bias interventions.

The "Cannot" Challenge

The second psychological factor which limits the effectiveness of social bias interventions relate to cognition, and mostly concern psychological social bias interventions that aim to reduce individuals' social bias. An important insight from social psychological research is that even when people are motivated to control their bias, they may fail to do so (Gawronski & Payne, 2010; Pendry, & Macrae, 1994). Research shows that our conscious awareness is very limited and some even argue that we have no conscious awareness of our cognitive processes, only some awareness of their output (Earl, 2014). Applied to recruitment decisions, this implies that we may have a gut feeling that we like a certain applicant better than another but limited (or no) access to what thought processes, including potential bias, that have caused these attitudes. If we are unaware of our social bias, we cannot control it.

Even when we do realize that we may have negatively stereotyped an individual, research has also shown that it is difficult to suppress stereotypes, and that attempting to do so may even increase stereotypical thinking (Macrae et al., 1994). Controlling prejudice is especially difficult when under stress, which is common in most workplaces (Pendry, & Macrae, 1994). However, actively engaging in counter-stereotypical thought is more effective than simply trying not to stereotype (Fitzgerald et al., 2019; Paluck et al., 2021) and research suggests that control attempts explain part of the reducing effect counter-stereotypical examples have on implicit bias (Calanchini et al., 2021). To change stereotypes, we also need to learn new associations, which takes practice, which is true for retraining automatized behavior in general (e.g. Calanchini et al., 2021; Gawronski & Payne, 2010).

The cognitive limitations I've described implies that social bias interventions that teaches participants how social bias works and instructs and motivates them to not stereotype, to mitigate their prejudice and not discriminate, may show limited success. In addition to some participants actively resisting to comply with the interventions (the "will not" challenge), reasons also include the principles of our cognitive functioning (the "cannot" challenge).

So where does GIIU fit in this range of social bias interventions? What limitations do I predict for GIIU and what potential may GIIU have to help reduce segregation in the labor market? Also, how should future psychological research test the effectiveness of GIIU, in my opinion? I will conclude with some thoughts on this matter.

Implications for GIIU

GIIU was designed as a structural intervention (Jönsson & Sjödahl, 2017). To remind readers, it searches through data sets with quantitative ratings of past job applications or employee performance ratings, with the aim to detect mean differences between targeted social groups. GIIU also calculates to what degree the ratings of a comparatively lower rated social group should be "corrected" assuming that the mean differences reflect assessors' social bias. Since GIIU is not a preventative intervention (like systematic recruitment is), but detects and corrects potential biases after the fact, it has been called a "post hoc" bias intervention (Jönsson, 2022; Jönsson & Bergman, 2022; Jönsson & Sjödahl, 2017).

Predicted Limitations: Resistance

I predict that introducing GIIU in organizations will meet similar challenges as for other structural social bias interventions. Since GIIU is not intended to reduce assessors' bias, but correct for it post hoc, the most relevant type of limitation is motivational resistance (the "will not" challenge).

As an example, I predict that applying GIIU in organizations will meet objections in the form of denial that detected mean social group differences reflect assessors' social bias. That is, if GIIU for example shows that women have generally received lower ratings than men in promotion decisions, some may attribute this to women having lower competence than men, or being less career motivated than men, rather than indicating that the assessor undervalued women's competence or merits due to bias. This assumption contradicts research which show only small gender differences or "gender similarity" in most psychological traits, including ability tests and in "agentic" (e.g. status-pursuing) career-motivation (Diekman et al., 2016; Hyde et al., 2005, 2019; Tellhed et al., 2018; Zell et al., 2015). However, lacking evidence that an assessor has been biased by stereotypes or prejudice in their candidate ratings, other attributions are possible. For example, one may assume that a mean difference in a sample depends upon methodological limitations when GIIU was applied. Small participant samples may be skewed and not representative of population characteristics in the target categories. Further, resistance may also relate to autonomy motivation, where staff members disapprove of having their work corrected by an algorithm (Highhouse, 2008; Neumann et al., 2021).

If application of GIIU will meet these types of resistance, it may imply that organizations will not use GIIU to correct for detected mean social group differences in their work-related judgments. However, organizations which hesitate to correct ratings post hoc, may still perceive the feedback from GIIU as valuable information. In Sweden, companies are obliged by law to take active measures to prevent discrimination (The Equality Ombudsman, 2022b) and GIIU may be seen as a helpful tool in this diversity work. Possibly, organizations may want to use GIIU for examining mean differences in their work-related ratings, which form the basis for their recruitment processes, promotion strategies and resource allocation. If mean differences are detected for target categories in these evaluations, it should probe for further investigation into the origins of these differences. The ratings could for example be reevaluated to ensure that they reflect differences in merits in the sample, and not assessor bias.

Future research should investigate if implementing GIIU in organizations does meet resistance and what form this potential resistance takes. It could also compare attitudes towards using GIIU as an investigative tool in organizational discrimination prevention work, versus changing the ratings post hoc in accordance with GIIU's suggestions. I suggest using both quantitative method (rating scales) and qualitative method (argument analysis) to assess attitudes and potential resistance strategies towards implementation of GIIU in organizations.

Predictors and Moderators of Attitudes

Attitudes towards GIIU, and different forms of resistance strategies, is likely to vary between staff members in organizations. Drawing on past research on resistance towards other structural interventions, that I described above, staff members from higher-status groups in the organization (such as white men) should on average display more negative attitudes towards GIIU as compared to groups with lower status (Flood et al., 2021; Williams, 2021). There should also be individual differences in attitudes within these groups. For example, attitudes could vary in relation to social dominance orientation (e.g. group based power motivation, Sidanius & Pratto, 1999; Wilson, 2017), certain personality factors such as openness and agreeableness (Akrami et al., 2009), empathy (Aberson, et al., 2007), political ideology, for example regarding the GAL (Green/Alternative/Liberal)-TAN (Traditionalist/Authoritarian/Nationalist) dimensions (Solevid et al., 2021) and autonomy motivation (Highhouse,

2008; Neumann et al., 2021; Williams, 2021). Future research could test these factors as potential predictors or moderators of attitudes towards GIIU.

Another potential moderator of attitudes towards GIIU concern who performed the evaluated past ratings; oneself, someone else, or perhaps artificial intelligence (AI), where the latter is becoming increasingly common in Sweden (The Equality Ombudsman, 2022c). I predict that having one's own past ratings assessed for suspected bias generates the most negative attitudes, since it risks exposing past discriminatory behavior one has committed. Ensuring confidentiality in GIIU applications could help reduce the risk for this type of resistance. Contrastingly, I expect the most positive attitudes if GIIU is used to correct ratings made by AI. This since using AI for decisions has been criticized for the lack of transparency into the basis for some types of AI decisions, the recent insights that also AI discriminates (The Equality Ombudsman, 2022c), and for the circumstance that AI (supposedly) has no feelings that can be hurt.

Psychological Intervention?

Lastly, although GIIU is designed as a structural social bias intervention, there might also be reason to evaluate if it can be used as a psychological social bias intervention, that is if GIIU can reduce assessor's bias. My argument is that if assessors learn that they have systematically rated a target group lower than other groups in the past, some may become motivated to reduce their social bias in future ratings, at least if they rate high on factors that relate to low resistance towards social bias interventions (Calanchinini et al., 2021). However, as for other psychological social bias interventions, the effectiveness of GIIU to reduce assessors' social bias should then also depend on the limits of cognitive control that I have previously described (The "cannot" challenge).

Future research could also evaluate if GIIU: s possible potential to reduce assessors social bias may be strengthened if the GIIU output is presented in combination with education on topics such as the size of mean differences in ability in target groups and how unconscious bias operates. One could also test if adding GIIU to an existing psychological social bias intervention, such as "habit breaking training" (Devine et al., 2012) or "bias interrupters" (Williams, 2021), increases their potential to reduce individuals' social bias and organizational segregation.

If GIIU is evaluated for its potential to reduce individuals' social bias, I recommend using large enough participant samples to allow for testing of moderators of the intervention's effectiveness. Study design in psychological

social bias interventions rarely include sufficiently large participant samples to do this, but it might be that the intervention has strong effects for certain individuals, but zero or even reversed effects for others (e.g. that show high resistance). When this is the case, opposite effects can cancel each other out in statistical analysis and the overall result looks weak.

Concluding Thoughts

Much research has been devoted to finding ways to reduce social bias or reducing its effect on segregation in the labor market. GIIU is a new tool that adds to this array of interventions. It is designed as a structural social bias intervention, such that it does not aim to reduce individual's biases per se but detects patterns in past work-related judgments that may have been caused by social bias, and calculates how ratings should be changed, to correct for assumed bias.

I see GIIU as a promising new tool for the quest of increasing diversity in the labor market. I predict that it will be most warmly received in the role of a potential bias detector in organizational diversity evaluations. I also predict that the function of GIIU to not only detect suspected biases in work-related decisions, but also correct for them will face resistance in organizations. However, if GIIU examines ratings made by AI, I predict that correcting suspected bias in these ratings will be more readily accepted.

To test these predictions and more, GIIU should be empirically investigated, preferably in implementation in real-world organizations. Psychological factors which may be of interest to study include attitudes towards GIIU, in-depth qualitative analysis of potential resistance in staff members, statistical testing of individual and collective factors that may relate to variation in attitudes towards GIIU, and exploration of circumstances which affects attitudes. It may also be of interest to study if GIIU is a helpful addition to current psychological social bias interventions, and may thereby contribute to reducing individuals' stereotyping, prejudice and discrimination.

GIIU is likely to meet challenges in its implementation, particularly in the form of resistance to social bias interventions. This does not mean that GIIU is defective since movement towards progressive social change will always encounter resistance. For the goal of developing a society where social group belongingness does not hinder individuals' career development, we need both psychological and structural social bias interventions. GIIU may play a role in this quest.

References

Aberson, C. L., & Haag, S. C. (2007). Contact, perspective taking, and anxiety as predictors of stereotype endorsement, explicit attitudes, and implicit attitudes. Group Processes and Intergroup Relations, 10(2), 179–201. https://doi.org/10.1177/1368430207074726

Akrami, N., Ekehammar, B., Bergh, R., Dahlstrand, E., & Malmsten, S. (2009). Prejudice: The person in the situation. *Journal of Research in Personality, 43*(5), 890–897. https://doi.org//10.1016/j.jrp.2009.04.007

Allbright (2022). Trångsynt i toppen. [Narrow at the top]. https://static1.squarespace.com/static/5501a836e4b0472e6124f984/t/626d7a0bd81a355bc5b409c4/1651341841584/trangsynt-i-toppen_2022.pdf

Allport G. 1954. *The nature of prejudice*. Boston, MA: Addison-Wesley.

Bates, L. (2020). Men who hate women. The extremism nobody is talking about. Simon & Schuster.

Begeny, C. T., Ryan, M. K., Moss-Racusin, C. A., & Ravetz, G. (2020). In some professions, women have become well represented, yet gender bias persists-Perpetuated by those who think it is not happening. *Science Advances, 6*(26), 1–10. https://doi.org/10.1126/sciadv.aba7814

Block, K., Croft, A. & Schmader, T. (2018). Worth less? Why men (and women) devalue care-oriented careers. *Frontiers in Psychology, 29*(9), 1–20. https://doi.org/10.3389/fpsyg.2018.01353

Brewer, M. B. (1999). The psychology of prejudice: Ingroup love and outgroup hate? *Journal of Social Issues, 55*(3), 429–444. http://dx.doi.org/ 10.1111/0022-4537.00126

Calanchini, J., Lai, C. K., & Klauer, K. C. (2021). Reducing implicit racial preferences: III A process-level examination of changes in implicit preferences. *Journal of Personality and Social Psychology, 121*(4), 796–818. http://dx.doi.org/10.1037/pspi0000339

Caleo, S & Heilman, M. E. (2019). What could go wrong? Some unintended consequences of gender bias interventions. *Archives of Scientific Psychology, 7*(1), 71–80. https://doi.org/10.1037/arc0000063

Croft, A., Schmader, T., & Block, K. (2015). An underexamined inequality: Cultural and psychological barriers to men's engagement with communal roles. *Personality and Social Psychology Review, 19(*4), 343–370. http://doi.org//10.1177/1088868314564789

Dasgupta, N., & Asgari, S. (2004). Seeing is believing: Exposure to counterstereotypic women leaders and its effect on the malleability of automatic gender stereotyping. *Journal of Experimental Social Psychology, 40*(5), 642–658. https://doi.org//10.1016/j.jesp.2004.02.003

Devine, P. G., Forscher, P. S., Austin, A. J., & Cox, W. T. L. (2012). Long-term reduction in implicit race bias: A prejudice habit-breaking intervention. *Journal of Experimental Social Psychology, 48*(6), 1267–1278. https://doi.org//10.1016/j.jesp.2012.06.003

Diekman, A. B., Steinberg, M., Brown, E. R., Belanger, A. L., & Clark, E. K. (2016). A goal congruity model of role entry, engagement, and exit: Understanding communal goal processes in STEM gender gaps. *Personality and Social Psychology Review, 21*(2), 142–175. https://doi.org/10.1177/1088868316642141

Dobbin, F., & Kalev, A. (2016, July-August). Why diversity programs fail, and what works better. *Harvard Business Review*. https://hbr.org/2016/07/why-diversity-programs-fail

Dobbin, F., & Kalev, A. (2021). The civil rights revolution at work: What went wrong. *Annual Review of Sociology, 47,* 281–303. https://doi.org//10.1146/annurev-soc-090820-023615

Dover, T. L., Major, B., & Kaiser, C. R. (2016). Members of high-status groups are threatened by pro-diversity organizational messages. *Journal of Experimental Social Psychology, 62*, 58–67. https://doi.org//10.1016/j.jesp.2015.10.006

Dovidio, J. F., Gaertner, S. L., & Saguy, T. (2015). Color-blindness and commonality: Included but invisible? *American Behavioral Scientist, 59*(11), 1518–1538. https://doi.org//10.1177/0002764215580591

Duguid, M. M. (1), & Thomas-Hunt, M. C. (2). (2015). Condoning stereotyping? How awareness of stereotyping prevalence impacts expression of stereotypes. Journal of Applied Psychology, *100*(2), 343-359. https://doi.org//10.1037/a0037908

Earl, B. (2014). The biological function of consciousness. *Frontiers in Psychology, 5*, 1–18. https://doi.org//10.3389/fpsyg.2014.00697

Europaportalen (2022, June 8). EU ett steg närmare könskvotering i bolagsstyrelser. https://www.europaportalen.se/2022/06/eu-ett-steg-narmare-regler-om-konskvotering-i-bolagsstyrelser

European Commission. (2009). *Gender segregation in the labour market: Root causes, implications and policy responses in the EU*. Publications Office of the European Union. https://data.europa.eu/doi/10.2767/1063

European Commission. (2014). *A new method to understand occupational gender segregation in European labour markets*. Publications Office of the European Union. http://ec.europa.eu/justice/gender-equality/files/documents/150119_segregation_report_web_en.pdf

European Institute for Gender Equality. (2021). *Index score for European Union for the 2021 edition*. https://eige.europa.eu/gender-equality-index/2021/country

Faludi, S. (1992). Backlash: The undeclared war against women. London: Vintage.

FitzGerald, C., Martin, A., Berner, D. & Hurst, S. (2019). Interventions designed to reduce implicit prejudices and implicit stereotypes in real world contexts: A systematic review. *BMC Psychology, 7*(1), 1–12. https://doi.org//10.1186/s40359-019-0299-7

Flood, M., Dragiewicz, M., & Pease, B. (2021). Resistance and backlash to gender equality. *Australian Journal of Social Issues*, 56(3), 393–408. https://doi.org//10.1002/ajs4.137

Gawronski, B., & Payne, K. B. (2010). Handbook of implicit social cognition: measurement, theory, and applications. New York: Guilford Press.

Gilovich, T., Keltner, D., Chen, S., Nisbett, R. E (2019). Stereotyping, prejudice and discrimination. In Gilovich, T., Keltner, D., Chen, S., Nisbett, R. E (Eds.) *Social Psychology* (5th ed., pp. 359-407). New York: W. W. Norton.

Government Offices of Sweden (2008:567). *Discrimination Act.* https://www.government.se/4a788f/contentassets/6732121a2cb54ee3b21da9c628b6bdc7/oversattning-diskrimineringslagen_eng.pdf

Greenwald, A. G., & Pettigrew, T. F. (2014). With malice toward none and charity for some: Ingroup favoritism enables discrimination. *American Psychologist, 69*(7), 669–684. http://dx.doi.org/10.1037/a0036056

Heilman, M. E., Brett, J. F., & Rivero, J. C. (1991). Skirting the competence issue: Effects of sex-based preferential selection on task choices of women and men. *Journal of Applied Psychology, 76(*1), 99–105. https://doi.org/10.1037/0021-9010.76.1.99

Hewstone, M., Rubin, M., & Willis, H. (2002). Intergroup Bias. *Annual Review of Psychology, 53*(1), 575–604. https://doi.org/10.1146/annurev.psych.53.100901.135109

Highhouse, S. (2008). Stubborn reliance on intuition and subjectivity in employee selection. *Industrial and Organizational Psychology-Perspectives on Science and Practice, 1*(3), 333–342. https://doi.org//10.1111/j.1754-9434.2008.00058.x

Holt, N., Vliek, M., Sutherland, E., Bremner, A., Passer, M., & Smith, R. E. (2019). *Psychology: the science of mind and behaviour* (Fourth edition). Maidenhead: McGraw-Hill Education

Hyde, J. S. (2005). The gender similarities hypothesis. *American Psychologist, 60*(6), 581–592. http://dx.doi.org/10.1037/0003-066X.60.6.581

Hyde, J. S., Bigler, R. S., Joel, D., Tate, C. C., & van Anders, S. M. (2019). The future of sex and gender in psychology: Five challenges to the gender binary. *American Psychologist, 74*(2), 171–193. https://doi.org/10.1037/amp0000307

Jones, S., & Rutland, A. (2018). Attitudes toward immigrants among the youth. *European Psychologist, 23*(1), 83-92. https://doi.org//10.1027/1016-9040/a000310

Jones, S.O., Taylor, S., & Yarrow, E. (2021). 'I wanted more women in, but..': oblique resistance to gender equality initiatives. *Work, Employment and Society, 35*(4), 640–656. https://doi.org//10.1177/0950017020936871

Jönsson, M. L. (2022). On the prerequisites for improving prejudiced ranking(s) with individual and Post Hoc interventions. *Erkenntnis*. Advanced online publication. https://doi.org//10.1007/s10670-022-00566-2

Jönsson, M. L., & Bergman, J. (2022). Improving misrepresentations amid unwavering misrepresenters. *Synthese: An International Journal for Epistemology, Methodology and Philosophy of Science, 200*. https://doi.org//10.1007/s11229-022-03744-5

Jönsson, M., & Sjödahl, J. (2017). Increasing the veracity of implicitly biased rankings. *Episteme, 14*(4), 499–517. https://doi.org/10.1017/epi.2016.34

Kunda, Z., & Oleson, K. C. (1995). Maintaining stereotypes in the face of disconfirmation: Constructing grounds for subtyping deviants. *Journal of Personality and Social Psychology, 68*(4), 565–579. http://dx.doi.org/10.1037/0022-3514.68.4.565

Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J.-E. L., Joy-Gaba, J. A., Ho, A. K., Teachman, B. A., Wojcik, S. P., Koleva, S. P., Frazier, R. S., Heiphetz, L., Chen, E. E., Turner, R. N., Haidt, J., Kesebir, S., Hawkins, C. B., Schaefer, H. S., Rubichi, S., … Nosek, B. A. (2014). Reducing implicit racial preferences: I A comparative investigation of 17 interventions. *Journal of Experimental Psychology: General, 143*(4), 1765–1785. https://doi.org//10.1037/a0036260

Lai, C. K., Skinner, A. L., Cooley, E., Murrar, S., Brauer, M., Devos, T., Calanchini, J., Xiao, Y. J., Pedram, C., Marshburn, C. K., Simon, S., Blanchar, J. C., Joy-Gaba, J. A., Conway, J., Redford, L., Klein, R. A., Roussos, G., Schellhaas, F. M. H., Burns, M., … Nosek, B. A. (2016). Reducing implicit racial preferences:

II Intervention effectiveness across time. *Journal of Experimental Psychology: General, 145*(8), 1001–1016. https://doi.org//10.1037/xge0000179

Lansu, M., Bleijenbergh, I., & Benschop, Y. (2020). Just talking? Middle managers negotiating problem ownership in gender equality interventions. *Scandinavian Journal of Management, 36*(2), 1-9. https://doi.org//10.1016/j.scaman.2020.101110

Macrae, C.N., Bodenhausen, G.V., Milne, A.B., Jetten, J. (1994). Out of mind but back in sight: stereotypes on the rebound. *Journal of Personality and Social Psychology, 67*(5): 808–17. https://doi.org/10.1037/0022-3514.67.5.808

Mc Closkey, M., & Glucksberg, S. (1978). Natural categories: Well defined or fuzzy sets? *Memory and Cognition,* 6, 462-472. https://doi.org/10.3758/BF03197480

Munger, K. (2017). Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior, 39*(3), 629–649. https://doi.org//10.1007/s11109-016-9373-5

Neumann, M., Niessen, A.S.M., & Meijer, R.R. (2021). Implementing evidence-based assessment and selection in organizations: A review and an agenda for future research. *Organizational Psychology Review, 11(*3), 205–239. https://doi.org//10.1177/2041386620983419

Nordic Council of Ministers. (2021). Genusperspektiv på framtidens högteknologiska arbetsliv: En nordisk forskningsöversikt om utbildningsval inom STEM (science, technology, engineering and Mathematics). [A gender perspective on the high-tech work life of the future: A Nordic literature review of educational choice of STEM]. https://www.norden.org/sv/publication/genusperspektiv-pa-framtidens-hogteknologiska-arbetsliv

Paluck, L. E., & Green, D. P. (2009). Prejudice reduction: What works? A review and assessment of research and practice. *Annual Review of Psychology, 60*, 339–367. https://doi.org//10.1146/annurev.psych.60.110707.163607

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

Pendry, L. F., & Macrae, C. M. (1994). Stereotypes and mental life: The case of the motivated but thwarted tactician. *Journal of Experimental Social Psychology, 30*(4), 303–325. http://dx.doi.org/10.1006/jesp.1994.1015

Régner, I., Thinus-Blanc, C., Netter, A., Schmader, T., & Huguet, P. (2019). Committees with implicit biases promote fewer women when they do not believe gender bias exists. *Nature Human Behaviour, 3*(11), 1171–1179. https://doi.org/10.1038/s41562-019-0686-3

Roos, H., Mampaey, J., Huisman, J., & Luyckx, J. (2020). The failure of gender equality initiatives in academia: Exploring defensive institutional work in Flemish universities. *Gender and Society, 34*(3), 467–495. https://doi.org//10.1177/0891243220914521

Rupp, D. E., Song, Q. C., & Strah, N. (2020). Addressing the so-called validity–diversity trade-off: Exploring the practicalities and legal defensibility of pareto-optimization for reducing adverse impact within personnel selection. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 13*(2), 246–271. https://doi.org/10.1017/iop.2020.19

Ryan, A. M., & Ployhart, R. E. (2014). *A century of selection*. Annual Review of Psychology, *65*, 693–717. https://doi.org/10.1146/annurev-psych-010213-115134

Sidanius, J., & Pratto, F (1999). *Social dominance: An intergroup theory of social hierarchy and oppression*. Cambridge: Cambridge University Press

Solevid, M., Wängnerud, L., Djerf-Pierre, M., & Markstedt, E. (2021). Gender gaps in political attitudes revisited: the conditional influence of non-binary gender on left–right ideology and GAL-TAN opinions. *European Journal of Politics and Gender, 4*(1), 93–112. https://doi.org//10.1332/251510820X15978604738684

Spencer, S. J., Logel, C., & Davies, P. G. (2016). Stereotype Threat. *Annual Review of Psychology, 67*, 415–437. https://doi.org//10.1146/annurev-psych-073115-103235

Statistics Sweden (2002). *Statistics on persons with foreign background: Guidelines and recommendations*. https://www.scb.se/contentassets/60768c27d88c434a8036d1fdb595bf65/mis-2002-3.pdf

Statistics Sweden (2021). *Stora skillnader i arbetslöshet mellan utrikes och inrikes födda* [large differences in employment between citizens born in Sweden and abroad]. https://www.scb.se/hitta-statistik/statistik-efter-amne/arbetsmarknad/arbetskraftsundersokningar/arbetskraftsundersokningarna-aku/pong/statistiknyhet/arbetskraftsundersokningarna-aku-1a-kvartalet-2021/

Statistics Sweden (2022). *Women and men in Sweden - Facts and figures 2022*. https://www.scb.se/contentassets/b3ba3d3ad7a74749936c7fd2e3b4bee6/le0201_2021b22_x10br2201.pdf

Storage, D., Charlesworth, T. E. S., Banaji, M. R., & Cimpian, A. (2020). Adults and children implicitly associate brilliance with men more than women. *Journal of Experimental Social Psychology, 90*, 1–14. https://doi.org//10.1016/j.jesp.2020.104020

Svensson, L. G., & Ulfsdotter Eriksson, Y. (2009). *Occupational status. A sociological study on the perceptions and valuations of occupations*. (Research report no. 140). Department of Sociology, Gothenburg University, Sweden. https://gupea.ub.gu.se/bitstream/2077/19737/1/gupea_2077_19737_1.pdf

Tellhed, U. (2022). *Val efter eget kön: Könsskillnader i utbildningsval: teori och empiri från den socialpsykologiska litteraturen*. Jämställdhetsmyndigheten. https://jamstalldhetsmyndigheten.se/media/q0nfrnl2/bilaga-3-till-huvudrapporten-val-efter-eget-k%C3%B6n.pdf

Tellhed, U., Backström, M., & Björklund, F. (2018). The role of ability beliefs and agentic vs. communal career goals in adolescents' first educational choice. What explains the degree of gender-balance? *Journal of Vocational Behavior, 104*, 1–13. https://doi.org/10.1016/j.jvb.2017.09.008

The Equality Ombudsman (2022a, July 22). *Arbetslivet* [Work life]. https://www.do.se/diskriminering/diskriminering-olika-delar-samhallet/diskriminering-pa-jobbet

The Equality Ombudsman (2022b, September 15). *Active measures*. https://www.do.se/choose-language/english/active-measures

The Equality Ombudsman (2022c). Transparens, träning och data: myndigheters användning av AI och automatiserat beslutsfattande samt kunskap om risker för diskriminering. https://www.do.se/download/18.56175f8817b345aa7651be9/1646982570826/rapport-transparens-traning-och-data.pdf

Tildesley, R., Lombardo, E., & Verge, T. (2021). Power struggles in the implementation of gender equality policies: The politics of resistance and counter-resistance in universities. *Politics & Gender*, 1–32. https://doi.org//10.1017/S1743923X21000167

Whitley, B. E., & Webster, G. D. (2019). The relationships of intergroup ideologies to ethnic prejudice: A meta-analysis. *Personality and Social Psychology Review, 23*(3), 207-237. https://doi.org//10.1177/1088868318761423

Wiggins-Romesburg, C. A., & Githens, R. P. (2018). The psychology of diversity resistance and integration. *Human Resource Development Review, 17*(2), 179-198. https://doi.org//10.1177/1534484318765843

Williams, J. C (2021). *Bias interrupted. Creating inclusion for real and for good*. Brighton, Mass.: Harvard Business Review Press

Wilson, E. K. (2017). Why diversity fails: Social dominance theory and the entrenchment of racial inequality. *National Black Law Journal, 26*(1), 129–153. https://escholarship.org/uc/item/2zn704q4

Wolgast, M., & Wolgast, S. (2021). *Vita privilegier och diskriminering: processer som vidmakthåller rasifierade ojämlikheter på arbetsmarknaden*. Länsstyrelsen i Stockholms län. https://www.lansstyrelsen.se/download/18.635ba3017c11a69d575fdb/1632984190659/R2021-23-Vita%20privilegier%20och%20diskriminering-webb-slutlig.pdf

Zell, E., Krizan, Z., & Teeter, S. R. (2015). Evaluating gender similarities and differences using metasynthesis. *American Psychologist, 70*(1), 10–20. http://dx.doi.org/10.1037/a0038208

Zitek, E. M., & Hebl, M. R. (2007). The role of social norm clarity in the influenced expression of prejudice over time. *Journal of Experimental Social Psychology, 43*(6), 867–876. https://doi.org//10.1016/j.jesp.2006.10.010

Åkerlund, M. (2022). Far right, right here: Interconnections of discourse, platforms, and users in the digital mainstream. Akademiska Avhandlingar: Sociologiska Institutionen, Umeå Universitet.

3

Some Reflections on the Practical Applicability of GIIU

Jakob Bergman[1]

Abstract. This chapter discusses developments of GIIU in different ways. As the chapter tries to outline possible lines of future research, it is by nature exploratory, on the verge of being speculative. We discuss the issue when the bias depends on the score in a non-linear way and outline a test to detect different type biases of this kind. We also discuss issues where candidates are awarded multiple scores, when and how to apply GIIU.

Introduction

The post-hoc intervention General Informed Interval scale Update (GIIU) was originally suggested by Jönsson and Sjödahl (2017) as GIRU. It was further developed by Jönsson and Bergman (2022) and Jönsson (2022), and applied by Bergman and Jönsson (2022) on a grant application data.

In this paper we discuss some of the limitations of GIIU with focus on its practical applicability, and how these limitations may be mitigated.

Non-Linear Biases

Jönsson and Bergman (2022) state as a presupposition of GIIU that 'the prejudice operates in an approximately linear way' and also give examples of a non-linear bias where e.g. students of a certain ethnic group are always failed (regardless of their performance) or women are never awarded the highest grade(s). This type of bias would obviously be very hard to correct, as there is usually no way of knowing which of the students who should not have been

[1] Jakob Bergman, Senior Lecturer in Statistics at the Department of Statistics, Lund University.

failed or which women among those receiving the highest grade should have received an even higher grade. If auxiliary information is available, e.g. the opinion of a second evaluator or the scores of some other test, this could be used to indicate which individuals might be subject to a biased evaluation score. However, since such a situation with auxiliary information is not the one for which GIIU is designed and we furthermore also expect it to occur rarely in practice, we will not consider it further in this paper.

Another type of non-linear bias can arise when the bias is a function of the evaluation score. It has been shown in the literature that a biased evaluation is more common when there is greater uncertainty in evaluation. If the candidate is clearly very good or very bad, and the score is obvious, there is less room for subjectivity and hence biased assessments. However, if the candidate's performance is (partially) contradictory or ambiguous, it was shown by e.g. Dovidio and Gaertner (2000) and Hodson et al. (2002) that evaluators awarded lower scores to black candidates than white candidates, which was taken as indication of aversive racism. From our point of view, 'ambiguous' would in general mean a mid-range score. The bias is then a function of the score, where the bias is the greatest for mid-range values and smaller (or even non-existent) for the smallest and largest values.

As an example, we assume that scores are awarded as real numbers on scale from one to nine. We also assume that an evaluator is negatively biased towards one group adding a negative bias to the scores of members of that group. We can easily construct two such functions where the bias is greater for scores close to five and smaller for values close to one or nine, using the absolute and squared deviation from five, respectively:

$$bias(Score) = \frac{|Score - 5|}{2} - 2, \quad Score \in [1, 9] \tag{1}$$

$$bias(Score) = \frac{(Score - 5)^2}{2} - 2, \quad Score \in [1, 9] \tag{2}$$

In both cases, the bias functions will equal minus two when the score is five, and zero when the score is one or nine. The difference lies in how rapidly the bias decreases. Table 1 illustrates the two bias functions for integer values from one to nine. Note that for the quadratic function, the bias decreases more slowly than for the absolute deviation.

The type of non-linear biases introduced in (1) and (2), may easily be mitigated using GIIU, if we know the form of the bias, i.e. what the bias function looks like. Without any prior knowledge, this is an extremely hard, if

Table 1: Two examples of non-linear bias, where for each example the bias is the greatest for mid-range scores and smaller for the extreme scores. The two biases exemplified are calculated using (1) and (2).

Score	(1)		(2)	
	Bias	Biased score	Bias	Biased score
1	0	1.0	0	1.000
2	−0.5	1.5	−0.875	1.125
3	−1.0	2.0	−1.500	1.500
4	−1.5	2.5	−1.875	2.125
5	−2.0	3.0	−2.000	3.000
6	−1.5	4.5	−1.875	4.125
7	−1.0	6.0	−1.500	5.500
8	−0.5	7.5	−0.875	7.125
9	0	9.0	0	9.000

not impossible, task. Even if there is prior knowledge or auxiliary information, it is still a difficult task to estimate a bias function, unless one has very detailed knowledge of the form of the function or one knows exactly which individuals that have received biased scores. Otherwise one would need to make very strong assumptions about the distribution of scores within the groups, e.g. that all groups have the same distribution of scores. We find such assumptions to be not very realistic, and hence questionable.

A potential way forward could be to partition relevant social groups into three or more groups according to their scores. Assuming there are two salient social groups, A and B, we would thus create one group consisting of the third of the members of group A with lowest scores, one group of the third of the members of group A with the mid-scores, and one group of the third of the members of group A with the highest scores, and similarly partition the members of group B. We would then compare the mean scores for the A and B groups with the lowest scores, for the A and B groups with mid-scores, and for the A and B groups with the highest scores. If there is no bias, the difference in mean value would be close to zero for all three comparisons. If there is a constant bias, the mean difference would be positive (or negative) and about the same for all three comparisons. And if there is a bias of the type sketched above, we would expect the mean

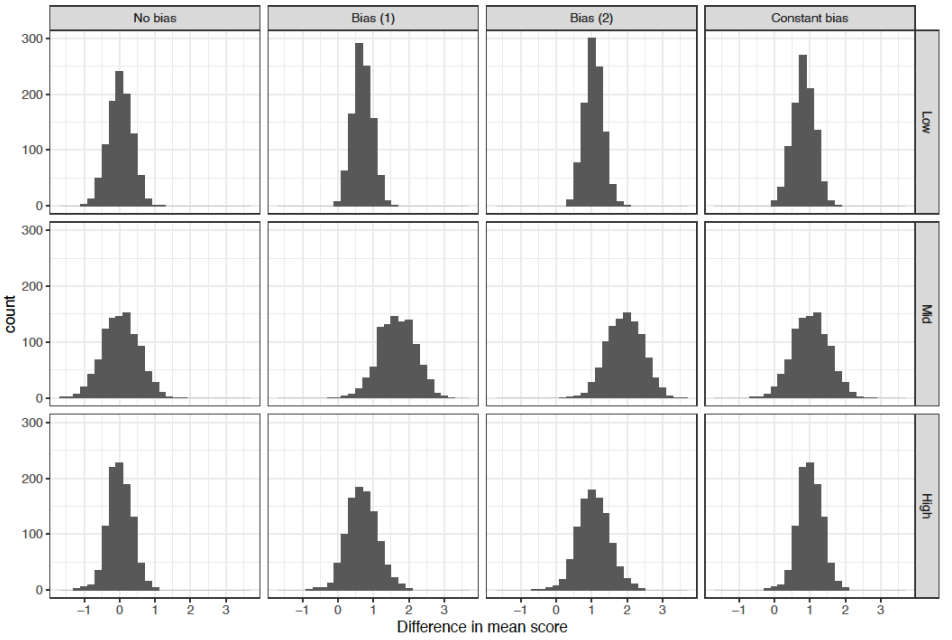


Figure 1: Histograms of the difference in mean scores from 1000 simulations of two groups with 99 candidates each. The top row shows the difference in mean score for the candidates with lowest scores, the middle row shows the difference in mean score for the candidates with the middle scores, and the bottom row shows the candidates with the highest scores.

difference to be greater in the mid-score group and smaller in the two other groups. One might also expect the variation to be smaller in the biased group with the lowest scores, as this group would consist of those candidates with genuinely low scores and those with low scores as a result of bias, thus pushing the scores towards the lower boundary, and conversely the variation to be slightly greater in the biased group with highest scores.

To investigate the feasibility of such a test, we conducted a small simulation study. In the study two groups of 99 candidates were constructed. Each candidate received a random score from a uniform distribution between 1 and 9. We calculated biased scores for the candidates from one of the groups using both (1) and (2). For comparison we also calculated biased scores with a constant bias, by subtracting 1 from all scores for one of the groups. (Biased scores below 1 were set to 1, to stay within score range.) The difference in mean score was then

Table 2: Standard deviations of the difference of the means from simulation study.

Partition	No bias	Bias (1)	Bias (2)	Constant bias
Low	0.3300	0.2641	0.2557	0.2995
Mid	0.4982	0.5020	0.4987	0.4982
High	0.3317	0.4219	0.4379	0.3317

calculated for the 33 candidates with lowest, the middle, and the highest scores, respectively. Figure 1 shows the results of the simulation study. As expected, the mean differences are greater on average for the biased scores. One can see a clear difference between bias 1 and 2 on the one hand, and the constant bias on the other, in that for the latter, the mid-range values have a similar mean as the low and high values, while for the former there is a shift in mean value for the mid-range values. One may also note that the ordering of the values seems to impact the distribution of the mean of the mid-range values the most, for both the unbiased and the biased cases, as this partition has the greatest variation. The standard deviations of the differences in mean scores are presented in Table 2. These may be compared to the standard deviation of the difference of the means of two random samples of 33 observations each from a uniform distribution which is $\sqrt{2(9-1)^2/12/33} = 0.5685$. As may be expected, the ordering reduces the variation, especially for the low and high values. As anticipated, the variation is smaller for the low biased values and slightly increased for the high biased values.

We believe that a test created along these lines could be used to distinguish between different types of biases. However, several important issues remain to be studied. A fundamental task is to find an appropriate test statistic, and to determine its (approximate) distribution under the null and alternative hypotheses. A complicating factor is the fact that the samples are ordered, so the observations are conditioned on being the e.g. smallest third. Relating to this of course also the task to investigate the power of such tests. A more general question is to study the number of partitions. Is the number of partitions dependent on the shape of the bias or is there an optimal number of partitions? How does the number of partitions relate to the size of the history? From a power point of view, one could expect a practical minimum number of observations per partition, but is there a practical maximum number? Should the partitions increase as the size of the history increases?

Multivariate Grades

A situation where one could easily imagine there being ambiguity, is where candidates are assessed in several ways, or on several dimensions, or by several evaluators. In all these cases there will be several scores on which the final evaluation must be based. Jönsson (2019) discusses the need for a stringent and formalised weighting of scores when admitting students to PhD programmes in a Swedish context. It seems reasonable that this would also be the case in general. From a statistical point of view, several scores per candidate is a multivariate (or multidimensional) score.

An important question when applying GIIU, is at what stage one should apply GIIU. Typically, one would apply GIIU to the univariate final scores, but one could also imagine applying GIIU univariately to the underlying scores. The latter would be particularly relevant if the scores are evaluations by different evaluators, where some, but maybe not all, are prejudiced. An alternative to treating each evaluator separately, one could generalize GIIU to a multivariate setting, where the mean difference between the two social groups is assumed to be the null vector 0 (or some other specified vector d). This hypothesis could, assuming multivariate normal distributed scores, be tested using Hotelling's T^2 (a multivariate generalisation of Student's t). This would require finding a relevant set of multivariate functions for updating the scores. A fourth option would be the case, when the scores are (assumed to be) unbiased, but the weighting is biased, i.e. the evaluator uses different weighting functions for different groups. Depending on the circumstances, this could be a type of mixture problem (Aitchison, 1986).

Conclusion

As has been briefly outlined in this chapter, there are a number of potential developments for GIIU. Some of these are possibly application specific, and might even need to be tailored to specific situations. There are also developments which will require more research.

References

Aitchison, J. (1986). *The statistical analysis of compositional data*. London: Chapman and Hall. (Reprinted with additional material in 2003 by Blackburn press.).

Bergman, J. and Jönsson, M. L. (2022). Gender bias in grant applications: inquiry and the potential for a post-hoc remedy. Submitted.

Dovidio, J. F. and Gaertner, S. L. (2000). Aversive racism and selection decisions: 1989 and 1999. *Psychological Science*, *11*(4), 315–319.

Hodson, G., Dovidio, J. F., and Gaertner, S. L. (2002). Processes in racial discrimination: Differential weighting of conflicting information. *Personality and Social Psychology Bulletin*, *28*(4), 460–471.

Jönsson, M. (2019). Allt sammantaget den bästa kandidaten: Om möjligheten till en reglerad sammanvägning av meriter som ett sätt att undvika osaklig meritvärdering vid antagning till forskarutbildningen. *Högre utbildning*, *9*(2), 65–80.

Jönsson, M. (2022). On the prerequisites for improving prejudiced ranking(s) with individual and post hoc interventions. *Erkenntnis*, page in press.

Jönsson, M. L. and Bergman, J. (2022). Improving misrepresentations amid unwavering misrepresenters. *Synthese*, *200*.

Jönsson, M. L. and Sjödahl, J. (2017). Increasing the veracity of implicitly biased rankings. *Episteme*, *14*(4), 499 – 517.

4

Post Hoc Interventions in Criminal Sentencing

An Empirical Thought Experiment

Erik J. Girvan[1]

Abstract. Post Hoc Interventions (PHIs) are approaches for reducing the impact of discriminatory ratings, evaluations, or other decisions by correcting statistically for the impermissible discrimination before applying the decisions. Scholars have proposed a set of empirical criteria that are theoretically necessary for implementation of PHIs. In this paper, I conduct an empirical thought experiment to examine how the criteria, along with a normative consideration derived from U.S. anti-discrimination law, relate to conditions in an actual case: Application of PHIs to adjust for potential ethnic biases in criminal sentencing outcomes. Results suggest that, while the criteria may not all be present in their strong form, allowing for reasonable inferences, in many circumstances they can be likely satisfied in practice.

Introduction

A core tenant of the rule of law is that legal decisions ought not to be decided arbitrarily. Rather, following the Aristotelian notion of justice, like cases should be decided alike and different ones differently. Deciding which attributes of cases determine whether they are like or different is a normative question. Assessing whether the attributes are present in the circumstance of a particular case is an empirical one.

In the United States and elsewhere, there is an anti-discrimination norm embodied in legal doctrine (Girvan, 2020; Liebman, Butler, Buksunski, 2021). The norm provides that one class of attributes that ought not to be used to determine if cases are like is the race, ethnicity, or sex of those involved, along

[1] Erik J. Girvan, Associate Professor at the University of Oregon School of Law, University of Oregon.

with other protected attributes. Individuals who can show that they were treated differently by government officials, employers, or businesses based on one of these attributes can thus obtain equitable or financial relief.

Decision-makers who share the anti-discrimination norm or who wish not to be legally liable for violating it adopt a range of strategies to avoid making decisions based on the protected attributes (see e.g., Hassen et al., 2021; Madva, 2020). Most commonly these efforts are preventative, targeting factors (e.g., explicit and implicit bias) thought to contribute to impermissible, discriminatory decision-making. However, the preventative efforts have a mixed record of success (Lai et al, 2014; Lai et al, 2016; Mcintosh, Smolkowski, Gion, et al, 2020). Sometimes the efforts reduce disparities related to the protected attributes. Often, they do nothing. Occasionally they produce backlash effects, making the disparities worse (for a review see Tellhed, this volume).

In addition or as an alternative to preventative approaches, Jönsson and colleagues (Jönsson & Bergman, 2022; Jönsson & Sjödahl, 2017) suggest that harm from discriminatory decision-making may be mitigated after the fact using statistical methods to directly correct decisions for the extent to which protected attributes like race, ethnicity, or sex influenced their outcome, an approach they refer to as *post hoc interventions* (PHIs). In addition, they identify a set of empirical conditions thought to be necessary for use of PHIs.

The goal of this paper is to conduct an empirically grounded thought experiment into the viability of PHIs in practice. In particular, building on the findings reported in Girvan and Marek (2023), I use PHIs to correct for racial and ethnic disparities in the extent to which White and Hispanic individuals are sentenced to prison, as compared to jail or probation, for violations of criminal laws. In doing so, I apply the empirical requirements for PHIs discussed by Jönsson and Bergman (2022), along with an additional normative limitation on steps one may take to correct for disparities based on protected attributes, and discuss the implication for PHIs and flexibility in the specified conditions in practice.

Conditions for Post Hoc Interventions

PHIs are adjustments to ratings, evaluations, or other assessments, r, of a latent characteristic, c, that has been identified as a legitimate basis for decision-making. Their use involves three basic steps. First, prior ratings, r_0, are examined to determine whether they differ impermissibly based on protected attributes of the individuals being evaluated. If individuals with certain of the attributes, e.g.,

Men, have been assigned higher evaluations on r_0 than those with comparison attributes, e.g., Women, under conditions in which members of the two groups can be assumed to have the same distribution of the latent characteristic c_0, then the group difference in r_0 is assumed to reflect impermissible use of the attribute. Second, a precise, incremental, quantitative correction, v, is statistically identified that, when applied to r_0, produces the same evaluations for individuals irrespective of the attribute. Third, v is applied to future evaluations ($r_{1\text{-}n)}$ and the result, $r_{1\text{-}n} + v$, used to determine the decision outcome.

Jönsson and colleagues (Jönsson & Bergman, 2022; Jönsson & Sjödahl, 2017, see also Jönsson, this volume) discuss the empirical conditions that must, in theory, be present in order to use PHIs. They are, restated and summarized in my terms:

1. *Interval scale ratings*. To be able to calculate and apply correction, v, to ratings, r_0, r_0 must be on an interval scale.

2. *Low error*. To be able to justify application of correction, v, underlying estimates of differences in r_0 must be based on a large enough sample of r_0 such that the extent of error in estimates of group differences is sufficiently narrow.

3. *No unknown differences*. To be able to justify the inference that differences in r_0 are attributable to impermissible consideration of a protected attribute, there must either no or known differences in c_0 based on that attribute.

4. *Constant bias over time*. To be able to justify the inference that application of v to $r_{1\text{-}n}$ is corrective of impermissible consideration of a protected attribute in those future evaluations, the magnitude of the differences in r_0 and $r_{1\text{-}n}$ based on the attribute must not vary systematically.

5. *Same categorization as bias*. To be able to correct for impermissible consideration of a protected attribute using v, individuals being rated must be categorized in the same way on the attribute in the PHI process as they were by the evaluators who produced r_0.

6. *Same contingencies as bias*. To be able to correct for impermissible consideration of a protected attribute using v, the PHI process must incorporate any contingencies regarding differences in r_0 based on the attribute (e.g., intersectionality between two or more protected attributes, interactions between a protected and permissible attributes).

7. *Same relationship as bias*. To be able to correct for impermissible consideration of a protected attribute using v, v must reflect the relationship (e.g., linear, non-monotonic) between the attribute and r_0 in the evaluations.

In addition to the empirical limitations, there are numerous potential normative considerations regarding the conditions under which one ought or ought not to directly correct for disparities related to protected characteristics in evaluations, ratings, or other assessments. Here I consider one.

1. *Cure not worse than the disease*. The anti-discrimination norm provides that decisions about people ought not be directly impacted by their status with respect to their race, ethnicity, sex, or other protected attributes. By adjusting r_0 based on such characteristics, PHIs are arguably doing just that. Under the norms embodied in U.S. anti-discrimination law, the adjustments are justified as a corrective measure only to the extent that we are sure that the group difference was caused by impermissible consideration of the attributes, e.g., racism or sexism of decision-makers (Girvan, 2020). They may not, however, be justified if the differences are attributable to random error in the sample or extrinsic factors that are causally related to c_0 and also happen to be correlated with the protected attributes (Chemerinsky, 2014; Rutherglen, 2009). To the extent that r_0 differs based on protected attributes of the individuals being evaluated for a reason other than impermissible consideration of the attributes, deliberately applying v to $r_{1\text{-}n}$ based on an individual's status with respect to protected attributes may thus be regarded as itself a violation of the anti-discrimination norm.

Application of PHIs to Criminal Sentencing Decisions

Could PHIs be used to correct for racial disparities in criminal sentencing decisions? As an empirical thought experiment, I apply the PHI approach to adjust for ethnic disparities in a set of actual sentencing decisions. In doing so, I compare and contrast the conditions of the cases of criminal sentencing to the empirical conditions identified as necessary for PHIs as well as the normative consideration and identify implications of any similarities or differences.

Sample of Criminal Sentencing Decisions

For the empirical thought experiment, I used a sample of records of sentencing decisions regarding 222,035 unique sentenced offenses (USOs)[2] committed by

[2] USOs are the most serious concurrently sentenced offenses for each individual. An individual who was simultaneously sentenced for four offences the sentences for each of which were to be served concurrently would have only one – the most serious sentenced offense – in the sample as one USO. If an individual completed a sentence and then committed and were sentenced for

195,854 people who were ultimately incarcerated in the State of Oregon at any point between 2004 and 2018 and belonged to one of three racial/ethnic categories. White-White individuals (N=162,742; USOs=184,976) were those identified by actors in the legal system as White (non-Hispanic) and predicted, using validated estimates, to self-identify as White. Hispanic-Hispanic individuals (N=21,101; USOs=23,983) were identified by actors in the legal system as Hispanic (any race) and predicted, using validated estimates, to self-identify as Hispanic. White-Hispanic individuals (N=12,011; USOs=13,076) were identified by actors in the legal system as White and predicted, using validated estimates, to self-identify as Hispanic.

In the United States, sentences for more severe crimes are generally served in state-run prisons (i.e., longer-term, more secure facilities). By comparison, sentences for less serious offenses are generally to jail (i.e., short-term, locally operated facilities), probation, or a combination of the two. Consistent with the distinction, in the sample, 60,240 USOs resulted in sentences to prison and 161,795 sentences to jail, probation, or both.

In Girvan and Marek (2023), my collaborator and I analyzed this sample of criminal sentencing decisions to determine whether race and ethnicity of the individuals sentenced impacted the likelihood of their being sentenced to prison as compared to jail/probation. To summarize, psychological theory indicates that, for group-based biases to impact decisions, decision-makers must first identify and categorize target individuals as members of the relevant group. Accordingly, we reasoned that, to the extent group-based biases impacted sentencing decisions, there would only be sentencing differences based on perceived race/ethnicity, not self-identified race/ethnicity where the two differed: After accounting for legally relevant factors, individuals perceived by those in the criminal justice system as Hispanic would be more likely to be sentenced to prison than similarly situated individuals perceived to be White. However, sentences of individuals misperceived as White but who self-identified as Hispanic would not differ from those of individuals accurately perceived as White. Our findings were consistent with the predictions. Even after controlling for crime severity and criminal history, individuals who were accurately labeled as Hispanic in criminal justice records (Hispanic-Hispanic) were nearly twice as likely to be sentenced to prison as those who were accurately labeled as White [White-White; Odds Ratio: 1.95 (95% CI: 1.86, 2.04)]. By comparison, individuals who were mis-perceived in

another offence, or if they committed two offenses the sentences for which were served consecutively, then they would appear in the dataset twice, once for each USO.

criminal justice records as White but who, based on validated estimates, self-identified as Hispanic (Hispanic-White) had the same likelihood of prison sentences as those who were accurately perceived to be White [Odds Ratio: 1.01 (95% CI: 0.94, 1.07)].

The empirical thought experiment takes a step further from the findings in Girvan and Marek (2023) by asking: What would it look like to use PHIs to attempt to correct for the observed disparity? However, the PHI approach uses past estimates of disparities in ratings r_0 to create a corrective function, v, to be applied to future ratings, $r_{1\text{-}n}$. Accordingly, rather than using all of the data for r_0 and $r_{1\text{-}n}$, for the thought experiment, I split the sample into sentences of USOs up to and including 2014 (N=168,290), which I treated as r_0, and those after 2014 (N=53,745), which served as $r_{1\text{-}n}$.

PHI Steps

PHI Step 1: Identification of Disparities in r_0

The first step in PHIs involves use of extant data regarding ratings, r_0, of a latent characteristic, c_0, that has been identified as a legitimate basis for decision-making to determine whether they differ impermissibly based on protected attributes of the individuals being evaluated. Here, as is typical in the U.S., criminal sentencing decisions in Oregon are made with reference to a set of sentencing guidelines designed to assign longer and more punitive sentences to what I will refer to as more reprehensible behavior, c_0. The guidelines operationalize reprehensibility and provide for the duration of criminal sentences using two underlying considerations: The severity of the offence committed and the extent of the criminal history of the individual being sentenced (Or. Admin. R. 213-004-0001). At the intersection of any level of offense severity and criminal history, the guidelines provide a presumptive sentencing range within which the sentencing judge has discretion to choose the appropriate sentence (Or. Admin. R. 213-004-0001; Or. Admin. R. 213-005-0007). The sentencing judge may depart from a presumptive sentence range, but only upon a finding of "substantial and compelling reasons" to do so (Or. Rev. Stat. § 137.671; Or. Admin. R. 213-008-0001). Such departures may be dispositional (imposing probation when the presumptive sentence is prison or vice versa) or durational (diverging from the presumptive sentence as to the term; Or. Admin. R. 213-003-0001(6), (8)). Thus, in theory, adhering to sentencing guidelines, individuals with comparable criminal histories who commit similarly severe offenses should receive like sentences, r_0 (Mitchell, 2017).

Table 1: Logistic Regression Coefficients and Odds Ratios Indicating Likelihood of Sentences to Prison Compared to Jail and/or Probation by Offender Race and Ethnicity.

	Coefficients		Odds Ratios	
Intercept	-3.356	[-3.536, -3.176]	.04	[0.03, 0.04]
Race/Eth. (White-White)				
White-Hispanic	-.026	[-.108, .056]	.97	[0.90, 1.06]
Hispanic-Hispanic	.724	[.672, .776]	2.06	[1.96, 2.17]
Psudo-R2		.956		

Note. Cell values are logistic regression coefficients (first column) or corresponding odds ratios (third column) followed, in brackets, by the 95% confidence intervals. All p-values are less than .001 except that for White-Hispanic (p = .530). Coefficients and odds ratios for legally relevant factors omitted from table.

To assess whether there was an impermissible difference in sentences of USOs, r_0, based on the perceived race and ethnicity of the individuals being sentenced, I fit a logistic regression model that includes the legally relevant factors that should, under the law, determine the type of sentence: Indicators of the individuals' offense history and offense severity. To this I added the sentenced individuals' sex and race/ethnicity, described above (see Girvan & Marek, 2023). To account for potential impacts of lack of independence, p-values and confidence intervals for coefficients were calculated using cluster-robust standard errors.

The relevant portion of the results of the analysis are given in Table 1. Effectively replicating the results of Girvan and Marek (2023), they indicate that sentencing decisions made from 2004 to 2014 regarding USOs of individuals who were perceived to be Hispanic (i.e., Hispanic-Hispanic) were approximately twice as likely to result in a sentence to prison than decisions regarding USOs by legally similarly situated individuals accurately perceived to be White (i.e., White-White). By comparison, decisions about USOs committed by individuals perceived to be White but who, based on validated estimates, would self-identify as Hispanic (i.e., White-Hispanic) did not differ from those of White-White individuals.

PHI Step 2: Calculate Correction v

The second step of the PHI-process is to calculate a precise, incremental, quantitative correction, v, that, when applied to r_0, produces the same ratings

for individuals irrespective of their status on the protected attribute. Here we can use the value of the coefficient for Hispanic-Hispanic individuals in the model from step 1: .724.

PHI Step 3: Apply Correction v to $r_{1\text{-}n}$

The third step is to apply v to future evaluations, $r_{1\text{-}n}$ and use the result, $r_{1\text{-}n} + v$, to determine the decision outcome. To do so, I used the coefficients from the logistical model in Step 1 to generate predicted log-odds of a sentence to prison for each USO decided after 2014, i.e., the as $r_{1\text{-}n}$ dataset. I then subtracted the adjustment v of .724 from the log-odds of predicted sentences for USOs by Hispanic-Hispanic individuals. The adjusted log-odds were then used to predict sentencing decisions for all USOs, with those having a log-odds greater than 0, the equivalent to an odds greater than 1, being to prison.

To illustrate the impact of the adjustment, Table 2 provides the actual sentences (top two rows), sentences that would be predicted by the unadjusted r_0 model coefficients (middle two rows), and sentences predicted by the adjusted coefficients (bottom two rows). Notably, comparison of the actual sentences to the un-adjusted predicted sentences shows that the only group for which the un-adjusted coefficients over-predict prison sentences is Hispanic-Hispanic offenders. Application of the adjustment corrects this, bringing the predicted sentences more in line with those for the other groups. Comparing the predictions of the unadjusted and adjusted models thus suggests that application of the adjustment results, depending on the baseline, in

Table 2: Outcomes of Actual, Unadjusted Predicted, and Adjusted Predicted Sentences of USOs Made after 2014

		White-White	**Hispanic-Hispanic**	**White-Hispanic**
Actual Sentences	Prison	11,457 (.213)	1,947 (.036)	927 (.017)
	Jail/Probation	33,324 (.620)	3,422 (.064)	2,668 (.050)
Predicted Sentences	Prison	11,116 (.207)	2,192 (.041)	804 (.015)
	Jail/Probation	33,665 (.626)	3,177 (.059)	2,791 (.052)
PHI Adjusted Predicted Sentences	Prison	11,116 (.207)	1,695 (.032)	804 (.015)
	Jail/Probation	33,665 (.626)	3,674 (.068)	2,791 (.052)

Note. Cell values are counts followed by proportions of all sentences.

approximately 250 to 500 fewer prison sentences for USOs by Hispanic-Hispanic individuals sentenced post-2014, equivalent to about 5 to 10% of the USOs sentenced for this group.

Comparison of Example PHI to Identified Empirical and Normative Requirements

Interval Scale Ratings

Jönsson and Bergman (2022) specify that PHIs must be applied to interval ratings in order to be able to calculate a corrective function. As with many threshold decisions, the latent characteristic *c* upon which sentencing decisions are based, level of reprehensibility, may be thought of as a continuous, interval- (or even ratio-) level construct. Even so, when conceptualized as ratings, *r*, decisions regarding the nature of a criminal sentence are ordinal, representing a dichotomous decision to sentence an individual to prison, if sufficiently reprehensible, or jail/probation, if not. Consistent with the logistic regression approach used in the example, when such dichotomous decisions are aggregated, calculating a corrective function for them based on the log-odds of prison compared to jail or probation, which is on an interval scale, is straightforward. In practice, however, application of the corrective function to individual future decisions, $r_{1\text{-}n}$, is challenging. Judges do not issue their sentencing decisions in log-odds of prison and thus their decisions cannot be directly corrected in this way.

One alternative approach, used in the example, is to use the coefficients from the model fit on prior sentencing decisions, adjusted with *v*, to predict types of sentences for individual cases as they arise. This approach differs from the prototypical PHI, however, in that adjustments are not made directly to judges' ratings in the new cases. Indeed, if the predicted sentencing decisions are viewed as "correct," then, once the coefficients and adjustments are calculated, the decision process can be automated and judicial ratings in new cases are not actually required at all. To the extent that this is viewed as methodologically or normatively problematic, one could use a hybrid system in which judges continue to make sentencing decisions in parallel with adjusted predicted decisions generated from all prior sentencing decisions. Where the two differ, the predicted decision will be used, the judge notified that a protected attribute may have influenced the decision and invited to reconsider, or another layer of processes added such as supplemental review by a panel. In any of these scenarios, judicial decisions would govern in most cases and, where they did

not, at a minimum, they would continue to influence sentences indirectly through inclusion in the sample used to generate coefficients in future estimation (Step 1) or adjusted prediction (Step 3) models.

Low Error

A second requirement for PHIs is that the underlying estimates of differences in r_0 must be based on a large enough sample such that the extent of error in estimates of group differences is sufficiently narrow to be usable. In the example, there are ample prior sentencing decisions to make reliable estimates of the influence of protected attributes on decisions. Indeed, the range of the 95% confidence interval around the coefficient estimate for the influence of perceived Hispanic ethnicity is equivalent to just .02 of the standard deviation in log-odds.

No Unknown Differences

The third requirement is that, for the inference that observed differences in r_0 are attributable to impermissible consideration of a protected attribute to be valid, there must be either no or known differences in c_0 based on that relevant protected attribute. In practice, this condition will not be met, except possibly in circumstances in which no judgment is required to operationalize the latent constructs that form the bases of the ratings being examined and no discretion afforded to raters interpreting or making ratings based on measures of them. In practice, however, there is also error and uncertainty in the measurement of nearly all latent characteristics and discretion in processes used to generate ratings from them. Accordingly, the requirement should turn on either (a) an assessment of whether the judgment conditions are such that an attribution of impermissible use of a protected attribute is a reasonable inference regarding the observed group difference or (b) application of a norm of presuming no unknown differences between groups, absent sufficient evidence to the contrary.

With respect to the sentencing decision example, the weaker requirement of a reasonable inference is satisfied in two ways. The first and perhaps most generalizable of the ways is that the institution on whose behalf the judges are making the decisions, the Oregon criminal justice system, had the opportunity to and did specify in advance the factors that ought to determine the outcome of the sentencing decisions: The severity of the USOs or the criminal record of the offender. Moreover, the influence of these factors was accounted for in the first step of the PHI, which showed that approximately 95% of the variance in

prison sentences was explained by them. Accordingly, it is a reasonable inference that observed differences between groups based on their status on a protected attribute stems from impermissible consideration of the attribute or an associated characteristic that ought not to impact the decision. Second and perhaps not as generalizable, the difference in sentencing decisions in the example is consistent with the predictions of psychological theory regarding the conditions under which group-based stereotypes and attitudes tend to influence decisions. And, while the correlational nature of the analysis precludes a strong inference of causality, any alternative explanation for the sentencing differences would also have to consider the fact that they are associated with perceived, but not self-identified, ethnicity.

Constant Bias over Time

Fourth, for a corrective function based on past ratings to accurately adjust for the impacts of impermissible consideration of protected attributes in future ratings, the magnitude of the impact must be relatively consistent over time. As with the requirement of no or known group differences, in practice it will often be impossible to know exactly the extent to which impermissible consideration of protected attributes changed over time. Given sufficient longitudinal data, however, it is relatively easy to determine whether differences in decisions associated with protected attributes remain relatively consistent, supporting a reasonable inference that the impact of potential impermissible influence of consideration of them on the decisions is also consistent.

To illustrate, I separately re-ran the logistic regression model on sentencing decisions made during four different time frames: 2004 to 2006, 2007 to 2009, 2010 to 2012, and 2013 to 2014. Results, given in Table 3 (next page), suggest that the magnitude of the increased likelihood of USOs by Hispanic-Hispanic individuals being sentenced to prison as compared to those by legally similarly situated White-White individuals rose over the first three periods and then decreased. Moreover, the change is sufficiently large that the highest of the coefficients and associated adjustments, .881, falls outside of the 95% confidence intervals for the other time periods.

To illustrate the impact of the change, we can repeat our PHI process three times, treating the earlier time period, e.g., 2004 to 2006, as r_0 from which we computer the adjustment v and the subsequent one, e.g., 2007 to 2009, as the $r_{1\text{-}n}$ to which we apply it. The result would be that, in the first two times we used PHIs, the adjustment would under compensate for the influence of

Table 3: Logistic Regression Coefficients and Odds Ratios Indicating Likelihood of Sentences to Prison Compared to Jail and/or Probation for USOs by Hispanic-Hispanic Individuals by Groups of Years Sentenced.

	Hispanic-Hispanic Coefficients		**Hispanic-Hispanic Odds Ratios**	
2004 - 2006	.634	[.532, .737]	1.89	[1.70, 2.09]
2007 - 2009	.724	[.631, .817]	2.06	[1.88, 2.26]
2010 - 2012	.881	[.779, .984]	2.41	[2.18, 2.68]
2013 - 2014	.672	[.540, .803]	1.96	[1.72, 2.23]

Note. Cell values are logistic regression coefficients (first column) or corresponding odds ratios (third column) followed, in brackets, by the 95% confidence intervals. All p-values are less than .001. Other coefficients and odds ratios omitted from table.

protected attributes on ratings in the subsequent period. By comparison, the third time the adjustment would over-compensate. How much such under- or over-compensation is acceptable may be context specific, depending on empirical considerations related to factors like the overall stability in the ratings themselves and normative considerations regarding the implications of incremental changes in the ratings. For example, where ratings fluctuate considerably over time but where incremental changes to ratings have a low impact on outcomes of others, e.g., where the decisions outcomes are independent as when assigning grades based on absolute performance, then instability may be less of a concern. In the context of the thought experiment, because sentencing decisions are relatively independent, i.e., adjusting the decision so that someone who would have gone to prison instead goes to jail or serves probation does not require that someone else who would have gone to jail or served probation to now serve a prison sentence, the level of instability observed here may be acceptable.

Same Categorization as Bias, Same Contingencies as Bias, and Same Relationship to Bias

The fifth, six, and seventh requirements for PHIs I identify each capture a type of complexity in the ways in which, as a result of social psychological processes, raters' impermissible consideration of protected attributes may impact r_0: Raters may categorize individuals based on protected attributes differently than would others or the individuals themselves, raters' decisions

may be influenced by the interactions between several protected attributes or protected attributes and characteristics of the rating situation, and the influence of protected attributes on r_0 may otherwise be non-linear. The more accurately the complexities are modeled in the PHI process, the more accurately v will be able to correct for bias when applied to $r_{1\text{-}n}$. As with the other requirements, if interpreted strictly, in practice, given variation in human perceptions, differences in the subjective salience of particular socially defined attributes, and the conditional nature of some biases, it will rarely be completely satisfied in circumstances involving room for interpretation or discretion. However, if viewed as requirements that PHIs may be done in circumstances where it is reasonable to infer that the primary influence of the protected attributes on ratings can be sufficiently similarly captured in the PHI process, then it may only limit some applications of PHIs in which there is particular reason for concern. To illustrate, for each requirement, I consider the example PHI in criminal sentencing in light of some potential ways in which estimation and application of v may differ from the original impacts of impermissible consideration of protected attributes on r_0.

First, for raters' attitudes and stereotypes regarding protected attributes to impact their decisions, the raters must identify someone based on their status in relation to the attributes (Rees, Ma, & Sherman, 2020). Where someone's status as to a protected attribute is difficult to observe reliably, data sources based on self-reported status regarding the attribute may differ from those based on perceived status. For example, in the U.S., research results indicate that individuals who identify as Hispanic or Native American tend to be mistaken for White (Girvan & Marek, 2023). In the Oregon Department of Corrections database, race and ethnicity of offenders were based on the perceptions of race and ethnicity by officials in the criminal justice system, such as the arresting law enforcement officers. In such circumstances, using self-identified race and ethnicity in PHIs, by, for example, asking inmates, job applicants, or others subject to decision-making to identify their own race and ethnicity, rather than recording the attributes perceived by the raters themselves, may result in inaccurate adjustments.

To illustrate, I re-ran the logistic regression model for Step 1 of the PHI using only the validated estimates of self-identified race and ethnicity rather than perceived values. The result indicated less of a difference between sentences of USOs by Hispanic and White individuals [$\beta = .494$ (95% CI: .524, .625); Odds Ratio: 1.78 (95% CI: 1.69, 1.87)], and thus that less of an adjustment would be needed than with the model using perceived race and ethnicity. If the coefficient and self-categorization approach were used to make

Table 4: Logistic Regression Coefficients and Odds Ratios Indicating Likelihood of Sentences to Prison Compared to Jail and/or Probation by Offender Race and Ethnicity, Sex, and the Interaction Between Them.

	Coefficients		Odds ratios	
Intercept	-3.363	[-3.543, -3.183]	.04	[0.03, 0.04]
Race/Eth. (White-White)				
White-Hispanic	.345	[.167, .524]	1.41	[1.18, 1.69]
Hispanic-Hispanic	.364	[.167, .561]	1.44	[1.18, 1.75]
Sex (Female)				
Male	.553	[.499, .607]	1.74	[1.65, 1.84]
White-Hispanic x Male	-.470	[-.670, -.271]	.63	[.51, .76]
Hispanic-Hispanic x Male	.384	[.180, .588]	1.47	[1.20, 1.80]
Psudo-R2		.956		

Note. Cell values are logistic regression coefficients (first column) or corresponding odds ratios (third column) followed, in brackets, by the 95% confidence intervals. All p-values are less than .001. Coefficients and odds ratios for legally relevant factors omitted from table.

adjustments in step 3, the result would be that the PHI would under correct $r_{1\text{-}n}$ for self-identified individuals who were perceived to be Hispanic and overcorrect those who were not.

The same potential problem may be extended further to protected attributes that are treated as categorical but the identification and influence of which varies continuously. One example of this is Afrocentric features, i.e., the extent to which people appear closer to stereotypes of the phenotype of individuals of African descent. Research on racial bias in sentencing in the U.S. indicates that people who have more Afrocentric features tend to receive harsher sentences than those with less Afrocentric features (Burch, 2015; King & Johnson, 2016). Under some circumstances, other potentially protected attributes like age may also influence judgments primarily continuously, thus making it important to capture raters' subjective perceptions of the characteristic directly.

Turning to contingencies, results of a substantial body of research suggests that the operation of stereotypes and attitudes regarding people based on their

Table 5: Outcomes of Actual, Unadjusted Predicted, and Adjusted Predicted Sentences of USOs Made after 2014

		White-White	Hispanic-Hispanic	White-Hispanic
Actual Sentences	Prison	11,457 (0)	1,947 (0)	927 (0)
	Jail/Probation	33,324 (0)	3,422 (0)	2,668 (0)
Predicted Sentences	Prison	11,114 (-2)	2,174 (-18)	781 (-23)
	Jail/Probation	33,667 (+2)	3,195 (+18)	2,814 (+23)
PHI Adjusted Predicted Sentences	Prison	8,262 (-2,854)	1,473 (-222)	700 (-104)
	Jail/Probation	36,519 (+2,854)	3,896 (+222)	2,895 (+104)

Note. Cell values are counts followed by change from original PHI values in Table 2.

attributes often interact or intersect with one another (Crenshaw, 2017; McCall, 2005). Stereotypes of or attitudes towards men and women, for example, may be qualitatively different than those for White men, Black men, White women, and Black women. Depending on the circumstances, understanding which cluster of attributes were salient to raters can be difficult because of the complexity of the interactions.

To illustrate the potential effects of intersecting attributes, I re-ran the logistic regression model for Step 1 of the PHI, adding the interaction terms between race and ethnicity and sex. The relevant results are given in Table 4. They indicate that the race and ethnicity differences from the original model are largely driven by the likelihood of USOs by Hispanic-Hispanic men resulting in a sentence to prison rather than jail or probation, which is higher than that for USOs by individuals of any other combination of race and ethnicity or sex.

To assess how much difference the outcomes of PHIs using a model that adjusts for significant effects of race and ethnicity, sex, and their interaction terms rather than just race and ethnicity, I used the model with interaction terms to predict un-adjusted and adjusted sentences. The results, in Table 5, provide the number of actual, predicted, and adjusted predicted sentences along with the change from these values in the original PHI (see Table 2). Review of the table shows that addition of the interaction of race and ethnicity and sex to the model did not dramatically change the unadjusted predictions of the model

(middle two rows). However, application of the adjustments resulted in a substantial reduction in the overall number of prison sentences for each group. In addition to interactions between attributes, social psychological theory also indicates that the influence of stereotypes and attitudes also tend to be moderated by features of a decision situation. For example, attitudes or stereotypes associated with protected attributes are more likely to affect decisions that are discretionary or in which the "correct" outcome is unclear, such as when there is some ambiguity or uncertainty in the decision criterion or an exercise of judgement required in order to make a decision (Girvan, 2016; see also Bushway & Forst, 2013; Bushway & Piehl, 2001). Sentencing guidelines were enacted, in part, to reduce racial disparities by limiting judicial discretion (Stith & Koh, 1993). Even with sentencing guidelines, however, judges retain some discretion on the margins to depart from guidelines and can impose sentences of different severity or divert individuals into alternatives like probation and rehabilitative programs. Judges deciding to depart from the guidelines generally do so based on some consideration of subjective factors such as rehabilitative potential or ties with the community (Painter-Davis & Ulmer, 2020).

To illustrate with the example sentencing decisions, Table 6 gives the distribution of the raw number of sentencing outcomes and that would be adjusted by the PHI process (see Table 2) in terms of the sentencing guidelines. Consistent with psychological theory, the adjustments are not randomly distributed across the sentencing grid but rather tend to be concentrated in areas of the guidelines near the threshold for prison or jail and probation sentences. For example, the largest proportion of adjustments (26% of the total) occurred for USOs classified as fairly severe, i.e., 8 out of 11 on the Crime Seriousness Scale, with 11 being the most serious, committed by Hispanic-Hispanic individuals who were at the lowest two lowest levels of the Criminal History Scale, i.e., individuals with no record of serious crime as an adult. The second largest proportion of adjustments (16%) were for USOs classified as moderately severe (6 on the Crime Seriousness Scale) committed in individuals who had at least one prior felony involving harm to a person. By comparison, the PHI adjustments did not change the outcomes of any sentences for USOs involving the most serious crimes, i.e., those at 10 or 11 of the Crime Seriousness Scale, for which prison sentences are effectively the "correct" outcome. Similarly, the PHI adjustments resulted in only a small number of changes to USOs for crimes very low on the Crime Seriousness Scale, those for which the "correct" outcome is a combination of jail and probation.

Table 6: Number and Proportion of Total Sentencing Decisions Regarding Hispanic-Hispanic Offenders that Differ Between Actual and PHI-Adjusted Outcomes by Location on the Sentencing-Guidelines Grid.

		Criminal History Scale									
		A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**X**
Crime Seriousness Scale	**11**	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
	10	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
	9	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (0)	10 (.02)	16 (.03)	0 (0)
	8	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	5 (.01)	19 (.04)	32 (.06)	107 (.20)	0 (0)
	7	0 (0)	8 (.01)	5 (.01)	18 (.03)	3 (.01)	6 (.01)	5 (.01)	8 (.01)	17 (.03)	0 (0)
	6	14 (.03)	14 (.03)	26 (.05)	57 (.11)	3 (.01)	5 (.01)	3 (.01)	14 (.03)	26 (.05)	0 (0)
	5	1 (0)	1 (0)	2 (0)	1 (0)	4 (.01)	4 (.01)	3 (.01)	1 (0)	1 (0)	0 (0)
	4	6 (.01)	6 (.01)	2 (0)	2 (0)	1 (0)	0 (0)	2 (0)	0 (0)	1 (0)	1 (0)
	3	0 (0)	1 (0)	2 (0)	0 (0)	4 (.01)	2 (0)	2 (0)	3 (.01)	0 (0)	0 (0)
	2	0 (0)	2 (0)	9 (.02)	0 (0)	4 (.01)	10 (.02)	8 (.01)	2 (0)	0 (0)	0 (0)
	1	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (0)	0 (0)	0 (0)	1 (0)	0 (0)
	X	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	22 (.04)

Note. Cell values are raw numbers of sentences for USOs of Hispanic-Hispanic individuals at the indicated level of the Crime Seriousness Scale and the indicated level of the Criminal History Scale that were changed by the PHI adjustment followed, in parenthesis, by the proportion that number constitutes of the total number of adjusted sentences for USOs of Hispanic-Hispanic individuals. Light shading indicates percentage is between .05 and .09, inclusive; darker shading indicates percentages greater than .10. X indicates Unknown/Other.

Under circumstances like the sentencing decisions, where decisions are dichotomous and legitimate grounds for decision-making well specified, the concentration of adjustments to particular rating is likely not problematic for PHIs (although it may suggest specific targets for more effective preventative interventions; see e.g., McIntosh, Girvan, Fairbanks Falcon, et al, 2022). Where ratings are continuous, significant factors that raters are using to make decisions unknown or unincorporated into the PHI process, or both, however, contingent effects of rater' stereotypes and attitudes on r_0 may appear to be concentrated among certain rating ranges or otherwise non-linear. In such circumstances, efforts should be made to account for the moderating factors in the models used to calculate and apply adjustments.

Cure not Worse than the Disease

The final factor, embodied in certain anti-discrimination norms and legal doctrine, cautions generally against making direct adjustments to decision outcomes based on protected attributes of those involved as, itself, constituting discrimination. In effect, such adjustments are justified as a corrective measure only to the extent that we are sure that the group difference was caused by impermissible consideration of the attributes by raters (Chemerinsky, 2014; Girvan, 2020; Rutherglen, 2009). With respect to the application of PHIs in practice, whether the adjustment is an acceptable correction or unacceptable discrimination turns on the empirical strength of inference that the ratings were impacted by impermissible consideration of protected attributes as opposed to structural or other factors that happen to be correlated with those attributes. How strong the inference needs to be is itself a normative and legal question. As such, it could limit use of PHIs to the relatively narrow circumstances in which there is evidence of purposeful discrimination by the raters or extend PHIs to the relatively broad set of circumstances in which an objective observer could conclude from the available information that the protected attribute was a likely factor in the ratings (Sloan, 2020).

Application of PHIs to correct the sentencing decisions here likely falls between the two and perhaps satisfies both. There is no direct evidence that the judges who made the sentencing decisions did so in order to punish individuals that they perceived to be Hispanic more harshly, or, equivalently, those that they perceived as White more leniently. And I have made no effort to collect any. Even so, the combination of controls for the legally relevant information and finding that perceived, rather than self-identified race and ethnicity impacts sentencing outcomes for USOs on the margin is very consistent with

psychological theory regarding when raters' stereotypes and attitudes are most likely to impact their decisions. Accordingly, objective observers could certainly conclude that, consciously or unconsciously, the race and ethnicity of those being sentenced likely were a factor in the sentencing outcomes.

Conclusion

The goal of this paper is to use a sample of sentencing decisions to illuminate, explore, and examine the implications of a set of specified requirements for PHIs in practice. Among other things, the empirical thought experiment identified a common type of rating, dichotomous decisions such as whether an individual meets a certain threshold, that may be a challenging one in which to implement PHIs directly. In addition the example highlighted the potential importance of assessing stability in bias over time and modelling the specific nature of the biases in ratings, such as use of the same method to identify groups as did the raters. Finally, while, in practice, it may often be difficult to assess whether several of the requirements are strictly met, it may be possible to draw inferences about them. In those circumstances, the extent to which the inferences are sufficient to justify use of PHIs will likely turn on a normative and potentially legal question related to whether the correction itself is more problematic than its benefits.

Acknowledgements

This research was made possible through funding from the Pufendorf Institute for Advanced Studies at Lund University and the cooperation of the Oregon Criminal Justice Commission. The analysis and opinions expressed here are those of the author and do not necessarily represent the views of the Institute, Oregon Criminal Justice Commission, or the State of Oregon. Consistent with Open Data practices, a deidentified version of the data used for this study is available at https://osf.io/dr2wc. The author has no conflicts of interests to disclose.

References

Burch, T. (2015). Skin Color and the Criminal Justice System: Beyond Black-White Disparities in Sentencing. *Journal of Empirical Legal Studies*, *12*(3), 395-420.

Bushway, S. D., & Forst, B. (2013). Studying Discretion in the Processes that Generate Criminal Justice Sanctions. *Justice Quarterly*, *30*(2), 199-222.

Bushway, S. D., & Piehl, A. M. (2001). Judging Judicial Discretion: Legal Factors and Racial Discrimination in Sentencing. *Law and Society Review*, *35*(4), 733-764.

Chemerinsky, E. (2014). Making Schools More Separate and Unequal: Parents Involved in Community Schools v. Seattle School District No. 1. *Michigan State Law Review*, 633-665.

Crenshaw, K. W. (2017). *On intersectionality: Essential writings*. New York: The New Press.

Girvan, E. J. (2016). Wise Restraints?: Learning Legal Rules, Not Standards, Reduces the Effects of Stereotypes in Legal Decision-Making. *Psychology, Public Policy, and Law*, *22*(1), 31.

Girvan, E. J. (2020). Towards a Problem-Solving Approach to Addressing Racial Disparities in School Discipline Under Anti-Discrimination Law. *University of Memphis Law Review*, *50*, 995-1090.

Girvan, E. J. & Marek, H. (2023). Eye of the Beholder: Increased Likelihood of Prison Sentences for Those Perceived to Have Hispanic Ethnicity, *Law & Human Behavior* (in press).

Hassen, N., Lofters, A., Michael, S., Mall, A., Pinto, A. D., & Rackal, J. (2021). Implementing anti-racism interventions in healthcare settings: a scoping review. *International Journal of Environmental Research and Public Health*, *18*(6), 2993-3008.

Jönsson, M. L. and Bergman, J. (2022). Improving misrepresentations amid unwavering misrepresenters. *Synthese*, *200*.

Jönsson, M. L. and Sjödahl, J. (2017). Increasing the veracity of implicitly biased rankings. *Episteme*, *14*(4), 499 – 517.

King, R. D., & Light, M. T. (2019). Have Racial and Ethnic Disparities in Sentencing Declined?. *Crime and Justice*, *48*(1), 365-437.

Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J. E. L., Joy-Gaba, J. A., ... & Nosek, B. A. (2014). Reducing implicit racial preferences: I. A comparative

investigation of 17 interventions. *Journal of Experimental Psychology: General*, *143*(4), 1765-1785.

Lai, C. K., Skinner, A. L., Cooley, E., Murrar, S., Brauer, M., Devos, T., ... & Nosek, B. A. (2016). Reducing implicit racial preferences: II. Intervention effectiveness across time. *Journal of Experimental Psychology: General*, *145*(8), 1001-1016.

Liebman, J. S., Butler, K. C., & Buksunski, I. (2021). Mine the Gap: Using Racial Disparities to Expose and Eradicate Racism. *Southern California Review of Law & Social Justice*, *30*, 1-88.

Madva, A. (2020). Individual and Structural Interventions, in *An Introduction to Implicit Bias: Knowledge, Justice, and the Social Mind* (eds. Beeghly, E. and Madva, A.). New York: Routledge.

McCall, L. (2005). The complexity of intersectionality. *Signs: Journal of Women in Culture and Society*, *30*(3), 1771-1800.

McIntosh, K., Girvan, E. J., Fairbanks Falcon, S., McDaniel, S. C., Smolkowski, K., Bastable, E., ... & Baldy, T. S. (2021). Equity-focused PBIS approach reduces racial inequities in school discipline: A randomized controlled trial. *School Psychology*, *36*(6), 433-444.

McIntosh, K., Smolkowski, K., Gion, C. M., Witherspoon, L., Bastable, E., & Girvan, E. J. (2020). Awareness is not enough: A double-blind randomized controlled trial of the effects of providing discipline disproportionality data reports to school administrators. *Educational Researcher*, *49*(7), 533-537.

Rees, H. R., Ma, D. S., & Sherman, J. W. (2020). Examining the Relationships Among Categorization, Stereotype Activation, and Stereotype Application. *Personality and Social Psychology Bulletin*, *46*(4), 499-513.

Rutherglen, G. (2009). Ricci v DeStefano: Affirmative action and the lessons of adversity. *The Supreme Court Review*, 2009(1), 83-114.

Sloan, A. (2020). "What to Do About Batson?": Using a Court Rule to Address Implicit Bias in Jury Selection, *California Law Review*, *108*, 233-266.

5

Post Hoc Interventions and Swedish Discrimination Law

Anna Nilsson[1]

Abstract. This chapter discusses the implications of Swedish discrimination law for the use of post hoc interventions during recruitment processes that involve the ranking of job candidates. It argues that such interventions may assist employers in preventing direct and indirect discrimination by alerting recruiters, and others responsible for hiring decisions, to the fact that biases may have influenced the recruitment process. In doing so, such interventions at the very least provide recruiters with a good reason to take a second look at their ranking choices and to reflect on whether the choices can be justified. The chapter also examines the circumstances in which employers that rely on incorrect recommendations from post hoc interventions can be held liable for discrimination.

1. Introduction

Imagine that you apply for a management position at a Swedish company. During the recruitment process, you are informed that the company uses a statistical tool called 'GIIU' to prevent bias and prejudice from influencing the recruiters' decisions, including decisions about which candidates to interview and about the final ranking of candidates for the job. Initially, you find this approach professional and understandable. There is no shortage of studies that reveal discrimination in hiring decisions in Sweden. Studies have, for example, shown that Swedish employers tend to view people who are overweight as significantly less productive than people of average weight, and Arabs as less diligent than Swedes (Agerström and Rooth, 2007; Rooth, 2010; Agerström et

[1] Anna Nilsson, Associate Lecturer in Health Law, Faculty of Law, Lund University.

al., 2012). Employers also tend to reject applicants over 55, in particular women over 60, and people with more than two children (Eriksson, Johansson, and Langenskiöld, 2012, pp. 13–17; Carlsson and Eriksson, 2017, pp. 12–14). Correspondence test studies[2] have shown that homosexuals are less likely than heterosexuals with identical CVs to receive a positive response to a job application or to get invited for an interview (Ahmed, Andersson, and Mats Hammarstedt, 2013). For women, difficulties typically arise when seeking promotion or applying for managerial positions (Boschini, 2017, pp. 53–58). Studies from the United States have shown that women face a catch-22 situation when applying for managerial positions. When they present themselves as confident, competitive, and ambitious, they are viewed as highly competent, but they are nevertheless disliked, and therefore less likely to be hired (Rudman and Glick, 2001; Toneva, Heilman, and Pierre, 2020).

At the end of the recruitment process, you receive an email from the company informing you that you did not get the job. You start to wonder whether this negative outcome has anything to do with the GIIU tool. Did it really protect you against discrimination? Perhaps it saw biases that were not there and distorted the process. Wouldn't that be discrimination?

This chapter discusses GIIU, the Generalized Informed Interval Scale Update, a prejudice-reducing intervention developed by Jönsson and colleagues in a series of articles (Jönsson and Sjödahl, 2017; Jönsson and Bergman, 2022; Jönsson, 2022). As the fictional example illustrates, interventions of this kind raise several legal questions. One set of questions relates to discrimination law. Do post hoc interventions such as GIIU facilitate better compliance with the Discrimination Act (2008:567)? If so, what specific legal wrongdoings do post hoc interventions address? And if GIIU makes a mistake, does the employer who bases decisions on that mistake engage in discrimination? This chapter discusses these questions. To facilitate the discussion, the next section provides a brief introduction to post hoc interventions. Sections three and four explore the possibility of using post hoc interventions to address direct and indirect discrimination, and section five examines the circumstances in which employers that rely on incorrect recommendations from post hoc interventions can be held liable for discrimination.

[2] In these studies, researchers submit job applications for real job openings. The applications are often sent out in pairs, with CVs and cover letters that differ only with respect to the ethnicity and/or gender of the fictitious applicants. Researchers then measure the call-back rates for the different candidates and aim to identify differences in call-back rates relating to whether the fictious candidate was a man or woman, had a Swedish sounding name or not, etc.

2. Post Hoc Interventions

As mentioned above, social science and psychology research has demonstrated that biased thinking and decision-making continue to be problems in the Swedish labour market. Post hoc interventions are new methods of preventing such malpractice. Behind them is the idea that we can identify biased rankings of job candidates through the statistical analysis of recruiters' past rankings. Very briefly, this kind of intervention starts with an analysis of a specific recruiter's past ranking with the aim of identifying patterns, such as, for example, a tendency to rank men higher than women, or people with Swedish-sounding names higher than people with Arabic names – or, indeed, vice versa.[3] Such patterns are identified through the calculation and comparison of mean scores. First, we calculate the mean scores of members of the social group, or groups, that the recruiter may hold biases against (e.g. women or Arabs). Then, we compare these means with the mean scores that one would expect to find for these groups. If there is a statistically significant difference between the recruiter's means and the expected means, the assumption is that the discrepancy is due to the recruiter's ranking being influenced by prejudice or bias. The magnitude of the difference in mean scores is then used to propose a way of improving later rankings to better reflect the actual competences of the candidates (Jönsson and Bergman, 2022, pp. 5–7).

To conduct such an analysis, we need data about the distribution of job-relevant competences across relevant social groups. In the absence of such information, we must rely on assumptions about such distributions. Suppose that, in the fictional example in the introduction, the recruiter has a history of recruitments involving about 100 candidates and that the mean score for male candidates is significantly higher than that of male candidates. Such a difference would, of course, be less worrying if we knew that men were, on average, more qualified than women in the particular field at issue in this case. If, on the other hand, we knew or had reason to believe that male and female candidates were, on average, equally qualified for such work, then we would have reason to suspect that the ranking was influenced by prejudice and to take precautionary measures to prevent biased rankings in the future. [4] As Jönsson

[3] For more details about post hoc interventions, see the introductory chapter to this book.

[4] The fact that men and women in general are equally qualified for a particular type of job does not, of course, mean that the men and women who have actually been ranked by the recruiter in question were equally competent because the job applicants in the ranking history might not be

and colleagues have shown, post hoc interventions can under certain conditions mitigate the influence of biases during recruitment.[5] Such mitigation may not only increase the chance that the best qualified candidate gets the job, but also assist employers in preventing discrimination. The next section discusses the specific forms of discrimination that post hoc interventions might prevent.

3. The Prohibition of Discrimination

3.1 Direct Discrimination

The Discrimination Act prohibits six types of discrimination, including direct and indirect discrimination (Discrimination Act, ch. 1 §4). The act classifies some other acts as discrimination, including harassment and instructions to discriminate, but none of these acts seems relevant to the problem that post hoc interventions aim to address, namely biased rankings of job candidates. Direct discrimination in the recruitment context occurs when an employer treats a candidate less favourably than another candidate in a comparable situation for reasons associated with sex, transgender identity or expression, ethnicity, religion or other belief, disability, sexual orientation, or age (ibid., ch. 1, §4(1)). Candidates who are roughly similarly qualified are considered to be in a comparable situation (Government bill 2007/08:95, p. 487). To determine whether two applicants have equivalent qualifications, the Labour Court looks at the criteria set by the employer; what kind of knowledge, skills, and personal qualities the employer is looking for; and how well the candidates meet these criteria. To constitute direct discrimination, the employer's behaviour must, of course, also be related in a certain way to one or more of the discrimination grounds listed above. The preparatory works speak about a "causal link" between the employer's behaviour and the job applicant's sex, ethnicity, disability, etc. (ibid., p. 488). The discrimination ground need not, however, be

representative of the population of which they belong. Still, I think it is reasonable to say that a skewed ranking history gives us reason to suspect that bias influenced the ranking.

[5] For a post hoc intervention to correctly identify and mitigate biases, a number of conditions have to hold. The history of rankings must, for example, be large enough for the analysis to generate statistically reliable results, the recruiter's bias has to be relatively stable, and the statistical analysis must group the candidates into more or less the same social groups as the recruiter. A full account of the conditions that must hold is provided by Jönsson (2022) and Jönsson and Bergman (2022).

the sole or decisive reason behind an employer's action. It is sufficient that the candidate's sex, ethnicity, disability, etc. contributed to a negative recruitment decision (ibid, p. 489). Such a link is obviously present in a situation in which a recruiter chooses to rank, for example, Arab candidates lower than Swedish ones because the recruiter dislikes Arabs or holds negative stereotypes about them. It is also present if the recruiter puts Arabs in a disadvantageous position because he or she or prefers to work with people from his or her own culture (ibid., p. 488). Social science research has shown that in-group favouritism – that is, people being more loyal and more benevolent towards people they consider to be like themselves (their in-group) than towards people they do not identify themselves with (the out-group) – may prompt such behaviour (Tajfel and Turner, 1979; Brewer, 1999; Wolgast and Wolgast, 2021, pp. 28–29).

From the above we can conclude that there is a significant overlap between the kind of biased rankings that GIIU seeks to address and the behaviour outlawed by the prohibition of direct discrimination. This suggests that GIIU could indeed help employers to prevent this kind of discrimination and hence facilitate better compliance with the Discrimination Act. The overlap between the biased rankings identified by GIIU and the legal prohibition of direct discrimination is, however, not total. The prohibition of direct discrimination covers many more acts than just those related to hiring decisions, and unlike GIIU the prohibition of discrimination is concerned only with biased behaviour connected to one or more of the discrimination grounds. These differences aside, the most difficult aspect to assess is how well the statistical analysis, which is a key part of GIIU, corresponds to the legal analysis of particular job applicants' competences, which forms the heart of discrimination analysis. If these two approaches to identifying biased and discriminatory behaviour tend to generate different outcomes, that would speak against the usefulness of GIIU in preventing discriminatory hiring decisions.

As described above, a legal assessment of whether a job applicant has been discriminated against involves a comparison of his or her qualifications and the qualifications of other candidates who made it further in the recruitment process. In such assessments, no attention is paid to the mean scores awarded by recruiters or data about competence distribution across groups. In a case concerning the recruitment of a production artist, the plaintiff, represented by the Equality Ombudsman, presented data showing that people of Swedish ethnic origin were in a clear majority in the workplace in question. To be relevant to discrimination analysis, the Labour Court held, such data had to be combined with data concerning the proportion of people in Sweden who are of another ethnic origin than Swedish or, perhaps better, with information about

the extent to which persons of an ethnic origin other than Swedish are represented within the specific branch under consideration in this particular case (Labour Court, 2009, no. 16, p. 26). It ought to be noted that this was not the main reason why the court rejected the Equality Ombudsman's claim. Still, the court's reasoning provides some pointers about what kind of statistical data the court might find relevant in future cases.

The fact that statistical data and analysis have played a limited role in individual cases concerning direct discrimination law does not necessarily mean that they should continue to do so. To be sure, even a clear pattern of a recruiter repeatedly giving lower scores to candidates from marginalised or subordinated social groups than to candidates from more privileged groups in the labour market does not provide conclusive evidence that these rankings are biased. Other explanations are possible. Even if we could establish that a particular recruiter's past rankings were biased, that would not necessarily mean that the recruiter continued to let his or her biases influence future rankings. For that reason, a careful investigation of the particular ranking decision at issue in a case is indispensable. Still, a history of skewed rankings suggests either that candidates from the social group that benefits from the higher rankings are indeed better qualified, or that the recruitment process does not provide all candidates with equal opportunities. These are empirical matters, which cannot be settled by stipulation, and determining the most plausible explanation in a given context will depend on what we know about the distribution of relevant competences across groups within the relevant sphere, in combination with our knowledge of how bias and prejudice may influence recruitment processes.

3.2 Indirect Discrimination

I proceed now to indirect discrimination and the question of whether post hoc interventions can assist employers in preventing such misconduct. Indirect discrimination involves the application of a criterion or procedure that appears to be neutral but that puts people of a certain sex, transgender identity or expression, ethnicity, religion or other belief, disability, sexual orientation, or age at a particular disadvantage, unless the criterion or procedure has a legitimate purpose and the means that are used are appropriate and necessary to achieve that purpose (Discrimination Act, ch. 1, §4(2)). In recruitment processes, examples of such superficially neutral criteria are language requirements and dress codes that may be more difficult for 'foreign' job

seekers to comply with. At first glance, post hoc interventions and the prohibition of indirect discrimination do not seem to target the same phenomenon. GIIU is not designed to identify, let alone question, job requirements *per se*. GIIU looks at rankings, and is designed to target prejudice and biases, attitudes that cannot be said to be neutral – at least not if they concern any group protected under the Discrimination Act. Nevertheless, what GIIU classifies as biases are repeated misrepresentations of job candidates' competences associated with their sex, ethnic origin, or similar factors that cannot be explained by real or assumed differences in competence between men and women, Swedes and foreigners, etc. As noted above, this tool does not investigate the reasons behind these misrepresentations. It does not make an independent assessment of how well the ranked candidates' competences match the job requirements for a specific position. Thus, although what GIIU identifies as a biased ranking may be the result of stereotypical thinking and/or explicit or implicit biases related to sex, ethnic origin, age, etc., it may also be a result of the application of a neutral criterion, such as a language criterion, that puts certain groups at a disadvantage. Unless such requirements correspond to real business needs, such as, for example, the need to communicate with customers in Swedish or some other language, they cannot be justified and are thus likely to violate the prohibition of indirect discrimination (Labour Court, 2002, no. 128, and 2005, no. 98).

To sum up, post hoc interventions seem to be designed to prevent direct discrimination in the form of biased ranking decisions that lead to discriminatory hiring decisions. Such interventions may, however, also capture instances of indirect discrimination. Given that GIIU does not evaluate possible explanations behind seemingly skewed ranking histories, except for explanations connected to the distribution of competences across groups, we cannot conclude that what GIIU classifies as a biased ranking will always result in unlawful discrimination unless the recruiter follows GIIU's recommendation and updates the ranking. It is possible that the prior rankings can be explained or justified by reasons that GIIU has not considered. The next section discusses the room for such justifications in discrimination law.

4. Justifications and the Burden of Proof

Cases concerning direct discrimination often revolve around questions of evidence. Has the plaintiff been treated less favourably than others in a similar situation? If so, is the negative treatment related to the plaintiff's sex, ethnicity,

age, or any of the other prohibited discrimination grounds? The plaintiff must demonstrate circumstances that give reason to presume that he or she has been discriminated against (Discrimination Act, ch. 6, §3). If he or she is successful in doing so, the employer must show that discrimination has not occurred, in other words that the plaintiff was not subjected to less favourable treatment or that such treatment was not related to his or her sex, ethnicity, disability, etc. In other words, employers do not have to show that they selected the best qualified candidate for the job, but they need to convince the court that prejudice or other illegitimate considerations related to one or more of the discrimination grounds did not contribute – at all – to any unfavourable treatment. It is not enough simply to point to some other factor that *also* contributed to the decision (Government bill 2007/08:95, pp. 488–489). Employers have, for example, been held liable for discrimination based on sex in situations in which a job candidate's pregnancy was one of the reasons why an employer decided not to offer her the job, even though the decision was also based on other (legitimate) reasons concerning doubts about her skills and enthusiasm for the job (Labour Court, 2011, no. 23, p. 12).

Some victims of discrimination have access to evidence revealing an employer's "real" or openly discriminatory intentions, such as a secretly recorded conversation or similar evidence. In many cases, however, such evidence is not available, which means that claims about discrimination often depend on inferences from facts about the plaintiff's competence in comparison to the competence of other candidates who differ from the plaintiff only with respect to their sex, ethnicity, or some other discrimination ground. To establish a presumption of discrimination based on, let us say, sex, a female candidate typically tries to establish that she has better, or at least equal, formal qualifications compared with one or more male candidates who were offered the job and/or invited for an interview. In a case concerning discrimination based on sex and age, the Labour Court found that it was sufficient to establish a presumption of discrimination based on sex, and thereby shift the burden of proof to the employer, for the plaintiff, a 62-year-old woman, who was not invited for an interview, to show that she had a stronger CV than some men who were invited for an interview (Labour Court, 2010, no. 91, p. 14). In addition, the fact that no woman over 50 was invited for an interview was sufficient to establish a presumption that the plaintiff was also discriminated against on the basis of age (ibid.). To defend its decision, the employer pointed to the fact that more women than men were interviewed, that the interviewees were of various ages, including a man in his 60s, and that two women were

eventually hired (ibid., p. 7). None of these circumstances was, however, sufficient to rebut the presumption of discrimination.

The Labour Court has, however, accepted other arguments as refuting a presumption of discrimination. In the context of discrimination based on ethnic origin, the court accepted the employer's argument that a highly competent candidate was overqualified for the job (Labour Court, 2009, no. 16). The case concerned the recruitment of a production artist. The candidate, a man of Bosnian origin, made it to the interview stage. The interviewers, however, got the impression that he had "moved on" to more qualified and creative work, and was therefore less interested in the rather standardised tasks performed by a production artist. This, in combination with their impression of the candidate as being an individualist rather than a team player, made him less suitable for the job than Swedish candidates with poorer formal qualifications but more fitting personal qualities (ibid., pp. 24–26). A recent study of professional recruiters shows that outgroup applicants may prompt recruiters to focus more on the applicant's values and social skills and to subject these to closer scrutiny (Wolgast, Björklund and Bäckström, 2018). However, this risk was not discussed in the court case, which was decided in 2006.

Moreover, in situations in which candidates are roughly equally qualified, the Labour Court has accepted minor differences between the candidates' qualifications as sufficient to rebut a presumption of discrimination. In a case concerning recruitment to a hospital unit responsible for moving patients from one ward to another, the employer defended the decision to select two Swedish applicants over a candidate of Kosovo Albanian origin with reference to the fact that one of the Swedish applicants had knowledge of the hospital's underground corridor network, and that the other applicant had a friend who worked at the hospital and had put in a good word for him (Labour Court, 2006, no. 60, p. 13). Although we have little reason to doubt that knowledge of the corridor network was relevant to the position, it was not a competence specified in the job advertisement. This case and the case concerning the overqualified production artist illustrate that the prohibition of non-discrimination does not oblige employers to choose the candidate with the best qualifications; rather, it prohibits employers from rejecting candidates for reasons connected to their sex, ethnicity, disability, etc. Employers' rather broad freedom to select employees dates back to an agreement from 1906 between the labour unions and employers, and has since been reaffirmed in the jurisprudence of the Labour Court (Labour Court, AD 1985:129, p. 797, and AD 1996:147, p. 1189).

The court's lenient approach to the arguments and explanations put forward by employers has nevertheless been criticised by legal scholars and practitioners (Fransson & Norberg, 2017, pp. 105–106; Schömer, 2016). The low success rate of discrimination cases, in particular cases involving discrimination based on ethnic origin, even prompted an official inquiry into whether the rule governing the burden of proof ought to be amended to enable the Discrimination Act to better achieve its aim of combating discrimination and promoting equal rights and opportunities (SOU 2016:87, ch. 15). However, the inquiry concluded that the difficulty of proving discrimination was related not to the design of the burden of proof rule but rather to its application in individual cases (ibid., 463).[6]

Even if Swedish law grants private employers considerable freedom in employment decisions, it is reasonable to assume that many employers would be interested in a tool that could assist them in ensuring that their decisions are based on rankings that accurately reflect the candidates' actual qualifications.[7] Post hoc interventions are one such tool. However, using this tool to adjust rankings is not without risk. As described in section two, the method relies on assumptions that may turn out to be incorrect in particular situations. The next section asks what happens if an employer relies on an incorrect recommendation provided by a post hoc intervention and, as a result, offers a job to a less competent candidate at the expense of a more qualified one.

5. Liability for Decisions Based on Bad Advice

For post hoc interventions such as GIIU to work properly and generate correct recommendations, a few conditions must hold. There is not enough space here for a detailed discussion of these conditions, but Jönsson and Bergman address this topic elsewhere (Jönsson and Bergman, 2022, and Jönsson, 2022). If one or more of these preconditions is not fulfilled in a situation in which GIIU has been applied, there is a risk that the tool will either fail to identify a set of biased rankings as biased, or suggest ways of correcting for bias that is not in

[6] A proposal was made to further clarify the normative content of the rule, but this proposal did not result in any amendments to the Discrimination Act.

[7] Specific rules apply to recruitment for jobs within the state administration. When making these recruitment decisions, only objective factors, such as the candidates' qualifications and competences ("förtjänst och skicklighet"), may be considered (Public Employment Act, 1994:260, §4; Instrument of Government, ch. 12, §5).

fact present. For GIIU to generate appropriate recommendations, the statistical analysis involved in the intervention must among other things group the candidates into (roughly) the same social groups as did the recruiter whose level of bias is being tested (Jönsson and Bergman, 2022, p. 17). If, for example, the statistical analysis is focused on prejudice against women, but the recruiter does not hold biases against women in general but only against old women or very feminine women, there is a risk that the analysis will miss these biases, because biases against subgroups of women might have a small impact on the mean scores of the entire group of ranked women. If, on the other hand, statistically significant differences in mean scores are found, GIIU will suggest compensating for prejudice in cases where there is none; it will suggest that all women in the ranking are compensated, even though the recruiter's biases affected only old women or those who come across as very feminine (Jönsson, 2022, section 3). A similar problem arises if a recruiter is biased against subgroups of men and women that are of roughly equal size.[8] If they are of equal size, an analysis that focuses on differences between men and women per se will not find any statistically significant differences. As Jönsson notes, the method struggles with intersectional prejudice, both in terms of identifying such prejudice and in terms of making accurate recommendations about how to compensate for it (ibid., p. 20).

Another precondition that might give rise to incorrect recommendations in particular cases is that GIIU presumes that a recruiter's prejudice is fairly stable between rankings. If in a particular case this is not true, and the recruiter's prejudice has increased compared to previous recruitments, GIIU will undercompensate. It will still make a recommendation that will mitigate the effect of prejudice on the ranking under review, but it will not fully compensate for the negative impact of that prejudice (Jönsson and Bergman, 2022, pp. 16–17). If, on the other hand, the recruiter's prejudice has decreased, the method will overcompensate. Following GIIU's recommendations will, in such cases, decrease the veracity of the ranking, making it less representative of the candidates' actual competences.

From the perspective of discrimination law, both undercompensation and overcompensation are problematic, but for different reasons and to varying degrees. Undercompensation (failure to fully correct for prejudice) implies that the use of a post hoc intervention will not be sufficient to avoid responsibility

[8] One could, for example, imagine a recruiter who holds biases against very feminine women and very muscular men.

under the Discrimination Act: other measures will have to be implemented to ensure that no candidate is subjected to unfavourable treatment for reasons associated with sex, transgender identity or expression, ethnicity, etc. Such measures may involve, for example, criteria-based decision-making or the anonymisation of job applications. By contrast, overcompensation (correction for bias that is not there) is problematic because it entails a risk that the very use of a post hoc intervention will lead to a discriminatory decision. Take the example of a recruiter whose prejudice has decreased significantly since he or she compiled the rankings that were used to estimate his or her level of prejudice – perhaps thanks to some diversity or de-bias training.[9] In an attempt to minimise the impact of any prejudice or stereotypical beliefs related to, for example, sex, he or she now uses GIIU to modify a ranking. GIIU recommends that female candidates have their scores increased and, as a result, a male candidate is ranked below a female one, even though the male candidate is actually better qualified. As a result, the male candidate is not invited for an interview or offered the job. This course of events seems to match the criteria for direct discrimination on the basis of sex (Discrimination Act, ch. 1 §4(1)). The man was certainly treated less favourably than similarly qualified women, and this negative treatment was undeniably related to his sex. Had he been a woman, he would have benefited from the same score increase as the female candidates. To constitute direct discrimination, it is sufficient that the candidate's sex contributed to a negative recruitment decision; it does not have to be the sole or decisive reason behind that decision (see section 4, above).

Legally speaking, if a recruiter relies on incorrect recommendations from a post hoc intervention, it does not matter that the recruiter had no intention of treating candidates differently on the basis of a protected characteristic, nor does it matter that the recruiter was unaware that GIIU's recommendations were erroneous. As described in section three, the Discrimination Act does not attach much weight to the employer's intentions. Employers with benevolent intentions can also be held liable for discriminatory behaviour (Government bill 2007/09:95, p. 488). In a report on the use of automated decision-making in different areas covered by the Discrimination Act, the Equality Ombudsman argued that employers remain responsible for their decisions throughout the recruitment process regardless of which digital tools they use to make such decisions and regardless of whether they fully understand how such tools work (Equality Ombudsman, 2019, p. 16). If inspected by the Equality Ombudsman,

[9] However, we have reason not to be too optimistic about the impact of such training on hiring decisions (see e.g. Palluck et al., 2021, and FitzGerald et al., 2019).

an employer using a digital recruitment tool must, furthermore, be prepared to explain how it works and how it has been applied in particular recruitment cases. Given that it is up to the employer to design their recruitment process and determine what tools to use, and in view of the impact that hiring decisions have on people's career prospects and livelihoods, this rule seems reasonable.

6. Concluding Remarks

This chapter has discussed the implications of discrimination law for the use of post hoc interventions during recruitment, and has argued that post hoc interventions such as GIIU may serve as a form of decision-making support that helps recruiters to select the most qualified candidate for the job and thereby avoid discriminatory hiring decisions. This argument is based on the view that GIIU simply corrects for biases and prejudice. It does not provide any candidates with preferential treatment but merely corrects biased rankings so that they better reflect the candidates' actual competences. On this view, nothing in the Discrimination Act prevents an employer who has doubts about whether their recruitment procedures provide all candidates with equal opportunities from using a post hoc intervention as a form of decision-making support during recruitment.[10] Post hoc interventions may very well form a part of the employer's systematic work of preventing discrimination and promoting equal rights and opportunities during recruitment and promotion – work that Swedish employers are obliged to undertake (Discrimination Act, ch. 3 §§4 and 5(3)).

It is also possible, however, to view what GIIU does as a form of preferential treatment. Think back to the example in the introduction. Imagine that a candidate of a different ethnicity than yours is given a higher ranking because GIIU suggests that the recruiter is biased against people of that ethnicity. As a result, you lose your place as the top candidate, despite the fact that you and your competitor are equally qualified. Wouldn't that be preferential treatment on the basis of ethnicity? If so, it would not be lawful under the Discrimination Act, because it results in unfavourable treatment on the basis of ethnic origin and violates the prohibition of direct discrimination. We could perhaps avoid this problem if the employer merely used GIIU to indicate that bias might have

[10] The extent to which such interventions are compatible with data protection and privacy law, such as the General Data Protection Regulation, warrants further legal analysis.

influenced the process, and this indication triggered a second, careful consideration of the candidates' competences, which in turn led to an adjustment of the ranking. Still, there is a risk that a court would find that considerations of ethnicity contributed to the adjustment of the ranking.

Even if we view GIIU as engaging in some form of preferential treatment, it would still be lawful to use it to compensate for prejudice against persons with disabilities and persons with transgender identity or expression. This is simply because the Discrimination Act does not protect persons *without* disabilities against disability-based discrimination, and nor does it protect persons who are *not* transgender against differential treatment associated with this characteristic. Moreover, it would arguably be lawful to use GIIU as part of a systematic plan to achieve gender equality at a workplace in which one gender is underrepresented (Discrimination Act, ch. 2 §2(2)). It is, however, important that GIIU remains a form of decision-making support and that the employer makes an "objective assessment" of the candidates' qualifications before the hiring decision is made (*Hellmut Marschall v. Land Nordrhein Westfalen*, C-409/95, §33). According to EU law, affirmative action on the basis of sex must not entail an automatic preference for the candidate of the underrepresented gender.

Moreover, interventions such as the one discussed in this chapter must always be implemented with care. If incorrectly applied, they may decrease the veracity of rankings, and even contribute to discriminatory hiring decisions (see section 5, above). Thus, it is important that those using the tool understand how it works and are able to assess whether the preconditions for its proper functioning obtain. In my view, these constraints ought not to discourage employers interested in the technique. The tool builds on established statistical methods and is transparent about the rules that govern the outcome. If applied correctly and in the right circumstances, GIIU will increase the veracity of ranking decisions and mitigate the influence of bias and prejudice. It may not always produce perfect outcomes, but there is reason to believe that its results will often be better than those based on a recruiter's judgement alone (Jönsson and Bergman, 2022, pp. 22–26).

Acknowledgements

The author wishes to thank Senior Lecturer Per Norberg, Senior Lecturer and Associate Professor Leila Brännström, and Professor Jenny Julén Votinius for valuable comments on previous drafts of this text.

References

Agerström J. et al. (2012) Warm and competent Hassan = cold and incompetent Eric: A harsh equation of real-life hiring discrimination. *Basic and Applied Social Psychology*, *34*(4), 359–366. https://doi.org/10.1080/01973533.2012.693438

Agerström, J. and Rooth, D-O. (2007) Etnicitet och övervikt: implicita arbetsrelaterade fördomar i Sverige [Ethnicity and obesity: evidence of implicit work performance stereotypes in Sweden]. Uppsala*:* Institute for Evaluation of Labour Market and Education Policy.

Ahmed, A. M., Andersson, L. and Hammarstedt, M. (2013) Are gay men and lesbians discriminated against in the hiring process?. *Southern Economic Journal*, *79*(3), 565–585. https://doi.org/10.4284/0038-4038-2011.317

Boschini, A. (2017) Olika kön, olika lön – en ESO-rapport om diskriminering på arbetsmarknaden [Different sex, different salary – a study of discrimination at the labour market]. Stockholm: Wolter Kluwers.

Brewer, M. B. (1999) The psychology of prejudice: Ingroup love or outgroup hate?. *Journal of Social Issues*, *55*(3), 429–444. https://doi.org/10.1111/0022-4537.00126

Carlsson, M. and Eriksson, S. (2017) Påverkar arbetssökandes ålder och kön chansen att få svar på en jobbansökan? Resultat från ett fältexperiment [Does a job applicant's age and sex affect his or her chances to of getting a response on to a job application? Results from a field experiment]. Uppsala: Institute for Evaluation of Labour Market and Education Policy.

Equality Ombudsman (2019) Kunskapsöversikt om användningen och utvecklingen av automatiserad databehandling med algoritmer (artificiell intelligens) och stordata och diskriminering eller risker för diskriminering [Knowledge overview of the use and development of automated data processing with algorithms (artificial intelligence) and big data and discrimination or risks of discrimination]. Stockholm: Equality Ombudsman

Eriksson, S., Johansson, P. and Langenskiöld, S. (2012) Vad är rätt profil för att få ett jobb? En experimentell studie av rekryteringsprocessen [What is the right profile to get a job? An experimental study of the recruitment process]. Uppsala: Institute for Evaluation of Labour Market and Education Policy.

FitzGerald, C. et al., (2019) Interventions designed to reduce implicit prejudices and implicit stereotypes in real world contexts: a systematic review. *BMC Psychology*, 7(29). https://doi.org/10.1186/s40359-019-0299-7.

Fransson, S. & Norberg, P., (2017) Att förstå lagstiftning om diskriminering och mänskliga rättigheter [Understanding legislation about discrimination and human rights] Falun: Premiss förlag.

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

Jönsson, M. and Sjödahl, J. (2017) Increasing the veracity of implicitly biased rankings. *Episteme*, *14*(4), 499–517. https://doi.org/10.1017/epi.2016.34

Jönsson, M. (2022) On the prerequisites for improving prejudiced ranking(s) with individual and post hoc interventions. *Erkenntnis*. https://doi.org/10.1007/s10670-022-00566-2

Palluck, E. L. et al. (2021) Prejudice reduction: Progress and challenges. *Annual Review of Psychology*, *72*, 533–560. https://doi.org/10.1146/annurev-psych-071620-030619

Rooth, D-O. (2010) Automatic associations and discrimination in hiring: Real world evidence. *Labour Economics*, *17*(3), 523–534. https://doi.org/10.1016/j.labeco.2009.04.005

Rudman, L. A. and Glick, P. (2001) Prescriptive gender stereotypes and backlash toward agentic women. *Journal of Social Issues*, *57*(4), 743–762. https://doi.org/10.1111/0022-4537.00239

Schömer, E. (2016) Sweden, a society of covert racism: Equal from the outside: Everyday racism and ethnic discrimination in Swedish society. *Oñati Socio-legal Series*, *6*(3), 837–856.

Tajfel, H. and Turner, J. (1979) An integrative theory of intergroup conflict. In Austin, W. G. and Worchel, S. (ed.) *The social psychology of intergroup relations*. Monterey, California: Brooks/Cole, pp. 33–48.

Toneva, Y., Heilman, M. E. and Pierre, G. (2020) Choice or circumstance: When are women penalized for their success?. *Journal of Applied Social Psychology*, *50*(11), 651–659. https://doi.org/10.1111/jasp.12702

Wolgast, M. and Wolgast, S. N. (2021) Vita privilegier och diskriminering - Processer som vidmakthåller rasifierade ojämlikheter på arbetsmarknaden [White privilege and discrimination – Processes perpetuating racial inequalities at the labor market]. Stockholm County Administrative Board.

Wolgast, S., Björklund, F. and Bäckström, M. (2018) Applicant ethnicity affects which questions are asked in a job interview. *Journal of Personnel Psychology*, *17*(2), 66–74. https://doi.org/10.1027/1866-5888/a000197

6

Post Hoc Interventions and the General Data Protection Regulation

Martin L. Jönsson and Jonas Ledendal[1]

Abstract. Post hoc interventions rely on having access to certain personal data – such as the gender, age, ethnicity, and sexual orientation of the persons being evaluated – in order to detect and correct for prejudice. This brings these interventions into possible tension with pertinent data protection legislation, which might restrict the processing of said data. We discuss the compatibility of post hoc interventions, more specifically the Generalized Informed Interval Scale Update (GIIU), and the General Data Protection Regulation (GDPR). In particular, we investigate the legality of applying GIIU to datasets which haven't been collected with consent from the data subjects that their data is to be processed by GIIU. We conclude that many such applications are in compliance with the GDPR, but others, specifically those where the processing includes special categories of personal data that is considered sensitive, might not be.

1. Introduction

Post hoc interventions (Jönsson and Sjödahl 2017; Jönsson 2022; Jönsson and Bergman 2022; Bergman and Jönsson in preparation) embody the idea that prejudiced evaluations (competence scores, grades, performance reviews etc.) can sometimes be made more accurate after they have been produced. The most worked out such intervention, GIIU (Generalized Informed Interval Scale

[1] Martin L. Jönsson, Senior Lecturer in Theoretical Philosophy, Department of Philosophy, Lund University. Jonas Ledendal, Senior Lecturer in Business Law, School of Economics and Management, Lund University.

Update), relies on statistically identifying patterns of prejudiced (quantitative) evaluations in the history of evaluations of a particular evaluator, and then correcting for these patterns in future evaluations produced by the same evaluator.

Post hoc interventions rely on having access to certain personal data – such as the gender, age, ethnicity, and sexual orientation of the persons being evaluated – in order to detect and correct for prejudice. This brings these interventions into possible tension with pertinent data protection legislation, which might restrict the processing of said data. The following article is concerned with investigating this tension in the context of the European union, by investigating the compatibility of GIIU and the General Data Protection Regulation (GDPR).[2]

To illustrate the tension and to make the discussion below more vivid, the article will discuss the legislation in conjunction with two fictitious cases, corresponding to two types of situations that GIIU was designed to handle.

In the first case, imagine an upper secondary school math teacher – Matt – who consistently awards significantly lower grades to female students than what is to be expected from the national average for these students.[3] And imagine further that there is no reason to believe that the students in Matt's class are not representative of the populations to which they belong.

In the second case, consider a recruiter for a private care unit – Phyllis – who evaluates black applicants for positions as physicians at a significantly lower level than her fellow recruiters.[4] And imagine that there is no reason to believe that the applicants handled by Phyllis should stand out from the norm in the way they do.

---

[2] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ L 119, 4.5.2016, p. 1–88.

[3] For instance, we might be in a situation where we know that there is little difference between boys and girls on national pseudonymized math tests. Cf. the methodology used by the Swedish National Agency for Education, e.g. Skolverket (2019; 2020).

[4] GIIU essentially corrects for deviations from expectations based on population means. Since these are seldomly directly available they must be estimated, and this can be done in different ways. This is illustrated by our two cases. In the first one we use independently obtained information about the national averages for the students (see previous footnote), and in the second we use the evaluations of Phyllis' colleagues to estimate what non-prejudiced assessment looks like (cf. Jönsson 2022: fn. 12).

An advocate of GIIU might recommend that Matt's and Phyllis's future evaluations be modified in order to compensate for the detected incongruities. This can either be done automatically or by way of a recommendation of a decision support system. Either way, such a procedure requires that we know – in Matt's case – the gender of the students that Matt has evaluated in the past, and – in Phyllis's case – the skin color of the applicants that Phyllis has evaluated in the past. If we know this, we can calculate the average score members of the relevant social groups have received by Matt and Phyllis, and thus measure the size of the prejudice we are looking to correct. This presupposes, of course, that we can legally process the required data which is dependent on pertinent data protection legislation.

2. The General Data Protection Regulation

The General Data Protection Regulation is a regulation in EU law on data protection which was adopted in April 2016, and which has been directly applicable in all member states since May 25th 2018. The EU data protection framework does, however, also to a large degree rely on legal acts in the form of guidelines form the European Data Protection Board (EDPB), formerly the Article 29 Working Party. Such guidelines are adopted by the board under the GDPR to ensure consistent application and interpretation of the regulation.[5] Although, non-binding EDPB guidelines have a high de facto impact on how GDPR is applied by supervisory authorities and courts.[6]

The primary aim of the regulation is to protect the fundamental rights of individuals (data subjects) with regard to the processing of their personal data, mainly by making the processing more transparent and enhancing an individual's control over his or her personal data. The regulation also has a second aim of safeguarding the free movement of personal data within the union by ensuring that data protection legislation is uniform.[7] The GDPR lists

[5] See Article 70 of Regulation (EU) 2016/679 (GDPR). The consistent application is also ensured by the consistency mechanism (Article 63 of GDPR), which enables the EDPB to resolve disputes between national data protection supervisory authorities. The decision of the board is binding on the member states.

[6] Article 288 of the Treaty on the Functioning of the European Union (OJ C 326, 26.10.2012, p. 47–390). See also Craig & de Búrca 2020 on how the admixture of formal and informal law is a common feature of the legal order but can nonetheless give rise to problems.

[7] Article 1 of Regulation (EU) 2016/679 (GDPR).

detailed and fairly restrictive rules for how to process personal data. Although it is formally applicable only to a restricted region of the world, it has since its adoption become a model for similar legislation in many other parts of the world as well (Bradford 2021).

A noteworthy aspect of the regulation is its preventive nature. A person processing personal data is responsible in various ways (described below) for how the data is processed. It is, however, not enough that the person responsible implements safeguards to manage risks arising from its own processing but must also account for risks related to how the data can be used by others, such as potential malicious actors.

3. The Processing of Personal Data Required by Post Hoc Interventions

The GDPR lays down rules relating to the protection of natural persons (i.e., humans) with regard to the processing of personal data (Art. 1, GDPR). A first natural question then – to determine the applicability of the GDPR to post hoc interventions like GIIU – is to ask whether post hoc interventions involve (1) *the processing* of (2) *personal data*? These questions can in turn be fruitfully subdivided into five sub-questions, each relating to one of the following steps or processing operations required by GIIU:[8]

Collection – The step in which an evaluator passes an evaluative judgment concerning someone. For instance, the math teacher Matt, deciding to give one of his students, Molly, a particular grade in math.[9]

Recording – The step in which the evaluator records his judgment. For instance, Matt entering the grade he has decided on into a grade reporting system on his computer.

[8] Although the exact division of a process into steps, and the granularity of such a division, can be important from the perspective of the GDPR, we don't see a need for a more fine-grained division in the present context.

[9] The concept of data "collection" in data protection law is not limited to the act of obtaining data through literal collection, but also encompass passively receiving data and creating data. Hence, setting a grade for an assignment or passing a judgement is equated with collection.

Storing – The retention of the recorded judgment so that a history of evaluation (i.e. a set of such judgements which is big enough for statistical analysis) can be constructed. For instance, the retention of Molly's and her peers' grades on a server.

Analysis – The statistical analysis of differences in means between members of different social groups. For instance, the comparison of the means for the female students and the male students Matt has graded from the perspective of a particular assumption about how these means relate on the population level.

Modification – The potential modification of newly passed evaluative judgments in light of found incongruities in the history of evaluations. For instance, increasing the grades of Mikes newly evaluated female students in light of past female student having received biased grades.[10]

So in order to determine whether the GDPR is applicable to post hoc interventions we need to ask, for each of these steps, whether it requires (1) *the processing* of (2) *personal data*.

Processing is quite broadly defined as follows:

> 'processing' means any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction;
>
> (Article 4(2) of GDPR)

From this it is clear that each of the five steps involves processing, at least to the extent that they involve personal data: steps 1, 2 and 3 are all explicitly mentioned in the definition, step 4 involves retrieval and use, which is mentioned in the definition, and step 5 involves alteration which is also mentioned in the definition. In addition, it is clear from the language of the definition ("such as") that this list is non-exhaustive and intended to be illustrative.

[10] If GIIU is used as a decision support system, this step does not involve actually modifying data, but only suggesting to the evaluator that data should be modified. This is an important difference in the present context. See Section 6.

Personal data is also broadly defined in the following way:

> 'personal data' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person;
>
> (Article 4(1) of GDPR)

Hence, personal data is any data that both "relates to" a natural person and makes it possible to identify him or her.[11] In the context of post hoc interventions, data "relates to" humans in the sense that they are statements about humans (e.g., their school or work performance). However, since non-identifiable data is outside the scope of the GDPR, it is possible to anonymise personal data to make further processing steps compliant with the regulation.[12] This requires an appropriate anonymisation method which makes the risk of re-identification practically impossible or at least insignificant due to that it would require a disproportionate effort in terms of time, cost and man-power.[13]

The pieces of data that post hoc interventions process have the following relational form (illustrated by our first example):

> (M1) Matt Berry has given Molly Sinclair the grade 3.

[11] See further on the concept of personal data Article 29 Working Party, Opinion 4/2007 on the concept of personal data, WP136, adopted on 20 June 2007. See also Judgment of the Court of Justice of the European Union of 20 December 2017 in Case C-434/16, Nowak (ECLI:EU:C:2017:994).

[12] It is here worth noting that data pseudonymisation is not the same as anonymisation (see Article 4(5) of GDPR, which defines "pseudonymisation" as "means the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person"). Whereas anonymous data is non-personal data, pseudonymised data remains personal data and must comply with GDPR.

[13] Judgment of the Court of Justice of the European Union of 19 October 2016 in Case C-582/14, Breyer (ECLI:EU:C:2016:779), para. 46. See also Article 29 Working Party, Opinion 5/2014 on Anonymisation Techniques, WP216, adopted on 10 April 2014. The European Data Protection Board is developing new guidelines, which have not been made public at the time of writing.

M1 is clearly a piece of personal data on account of it featuring two names, which makes their bearers easily identifiable. Step 1 ('collecting') in particular thus seems inescapable to involve the processing of personal data since there is no way to anonymize at this point – particular students need to be assigned their grades – and is thus subject to the GDPR. The recording and storage of pieces of data like M1 typically involves the recording and storage of them in an unaltered state and this would mean that Steps 2 (Recording) and 3 (Storing) would also be subject to the GDPR. It should be noted though that this is not needed from the perspective of Step 4 (Analysis). In particular, what is needed from the perspective of this step (and thus what this step needs from steps 2 and 3) is instead something like the following:

> (M2) Matt Berry has given a female student the grade 3.

The application of post hoc interventions does not require us to retain any identifiers relating to the people being evaluated in the past. However, M2 would still count as personal data from the perspective of the previous definition since it features Matt's name. M1 featured identifiers for two data subjects and one still remains in M2. Moreover, since it is Matt's future evaluations that we are looking to update, it seems that we must retain his identifier. It seems highly unlikely that any anonymisation technique can be applied to the data that would break the link to the evaluator in a useful way that would make him or her unidentifiable in the manner required by the GDPR.[14] We can thus conclude that each of the aforementioned five steps involve the processing of at least some personal data, as defined by the GDPR.

4. Controllers, Processors and Territorial Scope

In order to determine the applicability of the GDPR to any particular processing of personal data, certain roles described by the GDPR must be identified, in particular a controller – "the natural or legal person, public authority, agency or other body which, alone or jointly with others,

[14] This is so because GDPR is not limited to data that is directly identifiable, i.e., data that is contained in the same dataset or otherwise held by the data controller. It is enough that the data subject is indirectly identifiable, e.g., by combining the data with other data which might or might not be held by the controller. See Judgment of the Court of Justice of the European Union of 19 October 2016 in Case C-582/14, Breyer (ECLI:EU:C:2016:779).

determines the purposes and means of the processing of personal data" (Article 4(7) of GDPR, our emphasis) – and one or more processors – "a natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller (Article 4(8) of GDPR).[15] The processing must also fall within the territorial scope of the regulation. Whether the GDPR is applicable to the relevant processing is hence a matter of whether it is carried out "…in the context of the activities of an establishment of a controller or a processor in the Union…". (Article 3(1) of GDPR) It is not significant whether that processing actually takes place in the union (Ibid.).[16]

Usually in cases like the first example, the controller is the public authority or similar entity in charge of the school, and in the second the employer. Although natural persons can also be controllers, the evaluators, i.e. Mike and Phyllis, who only access and process the personal data under the authority of their respective employers, are not considered controllers. Like other employees they are not directly responsible for the processing, but can themselves be data subjects, since their personal data is also processed during the intervention. The same is true for the post hoc intervener – the person administering GIIU – which is involved in the last two processing steps. Given that the school and the company are located within the European Union – which we will stipulate – the GDPR is applicable.

Another important role assignment in what follows is that of the data subject.[17] As was mentioned above, M1 featured identifiers for two different data subjects: the evaluator – which we will refer to using "data subject^{EV}" – and the evaluated – which we will refer to using "data subject^{ev}". This will become important in Section 5.

[15] See further European Data Protection Board, Guidelines 7/2020 on the concepts of controller and processor in the GDPR, adopted on 7 July 2021. Joint controllership is also possible when the purposes and means of the processing have been jointly determined by two or more controllers (Article 26 of GDPR).

[16] See further European Data Protection Board, Guidelines 3/2018 on the territorial scope of the GDPR (Article 3), adopted on 12 November 2019.

[17] The data subject is always a living human (natural person) and is protected regardless of nationality or residence. Deceased persons or legal persons are not protected (Recitals 14 and 27 of GDPR). See further on the concept of data subject Article 29 Working Party, Opinion 4/2007 on the concept of personal data, WP136, adopted on 20 June 2007.

5. Principles of Personal Data Processing

Now that we have determined that the GDPR is applicable to our two cases, we need to determine whether the corresponding post hoc interventions can be carried out in compliance with the principles of personal data processing stipulated by the GDPR. These are as follows:

> "Personal data shall be:
>
> a) processed lawfully, fairly and in a transparent manner in relation to the data subject ('lawfulness, fairness and transparency');
>
> b) collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes; … ('purpose limitation')
>
> c) adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed ('data minimisation');
>
> d) accurate and, where necessary, kept up to date; … ('accuracy');
>
> e) kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed; … ('storage limitation');
>
> f) processed in a manner that ensures appropriate security of the personal data, … ('integrity and confidentiality')."
>
> (Article 5(1) of GDPR)

Of these principles, the last two are of little importance when it comes to whether post hoc interventions can be in compliance with the GDPR and they will not be discussed further.[18]

5.1 Lawfulness and Purpose Limitation

The first point of tension between GIIU and the GDPR comes from the fact that post hoc interventions are fairly novel (first described by Jönsson and Sjödahl 2017). This means that there are few, if any, data sets ('histories of evaluations') that have been collected with the express purpose of applying post hoc interventions to them. The legality of applying GIIU to extant datasets

[18] The controller would have to ensure that these requirements are fulfilled, but in our opinion post hoc intervention does not pose any burden, legal uncertainties or complications that would go beyond what is required for any other processing of personal data.

without collecting consent is thus important to determine, because being able to do so could increase GIIU's scope of applicability. This depends on at least two different conditions.

First, it can be noted that the concept of lawfulness in Article 5(1) lit. a of the GDPR, also known as the requirement of legal basis, is expounded in a later article as follows:

> "Processing shall be lawful only if and to the extent that at least one of the following applies:
>
> a) the data subject has given consent to the processing of his or her personal data for one or more specific purposes;
>
> b) processing is necessary for the performance of a contract to which the data subject is party or in order to take steps at the request of the data subject prior to entering into a contract;
>
> c) processing is necessary for compliance with a legal obligation to which the controller is subject;
>
> d) processing is necessary in order to protect the vital interests of the data subject or of another natural person;
>
> e) processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller;
>
> f) processing is necessary for the purposes of the legitimate interests pursued by the controller or by a third party, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject which require protection of personal data, in particular where the data subject is a child."
>
> (Article 6(1) of GDPR)

This makes it clear that in cases where the data subject has not given its consent to the processing required by the last two steps of post hoc interventions (analysis and modification), lit. f (legitimate interest) can be used as the legal basis in the relevant processing context. Thus, the lawfulness of the analysis and modification steps depends on whether that processing is necessary "for the purposes of the legitimate interests pursued by the controller or by a third party, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject" (ibid.)

Second, lit. b of Article 5(1) stipulates that processing that goes beyond the purposes for which the data was originally collected must not be incompatible

with those purposes. It is worth noting that pursuant to lit. b such purposes must also have been explicitly specified by the controller at the point of collection.

Given the aforementioned, we must thus determine (1) whether post hoc analysis and modification is in the legitimate interests of the controller and not overridden by the interests of the data subject, as well as (2) whether this processing is compatible with the purposes for which the personal data was originally collected.

With respect to the first question, the answer seems to be affirmative for the two cases we are working with. The purpose for which the data was collected is likely something that can be paraphrased as "attempting to accurately measure a student's math proficiency" and "accurately determine an applicant's competence as a physician" respectively. It is difficult to see then how attempts to increase the accuracy of the relevant evaluations could be illegitimate (or in conflict with "interests or fundamental rights and freedoms of the data subject").

With respect to the second question, the answer also seems to be affirmative. Since the aim of the intervener is to increase accuracy, it appears to be an aim very much in line with the original purpose (as paraphrased in the preceding paragraph). Article in 6(4) of the GDPR seems to support this conclusion.

> "Where the processing for a purpose other than that for which the personal data have been collected is not based on the data subject's consent or on a Union or Member State law which constitutes a necessary and proportionate measure in a democratic society to safeguard the objectives referred to in Article 23(1), the controller shall, in order to ascertain whether processing for another purpose is compatible with the purpose for which the personal data are initially collected, take into account, inter alia:
>
> a) any link between the purposes for which the personal data have been collected and the purposes of the intended further processing;
>
> b) the context in which the personal data have been collected, in particular regarding the relationship between data subjects and the controller;
>
> c) the nature of the personal data, in particular whether special categories of personal data are processed, pursuant to Article 9, or whether personal data related to criminal convictions and offences are processed, pursuant to Article 10;

d) the possible consequences of the intended further processing for data subjects;

e) the existence of appropriate safeguards, which may include encryption or pseudonymisation."

Let's go through these in order. There is a clear link between the purposes for which the data was initially collected and the purpose of the further processing (cf. lit. a). The relevant context is one of processors attempting to give accurate measures on behalf of the controller (cf. lit. b). Hence, the further processing would meet the reasonable expectations of the data subject (Recital 50 of GDPR).[19]

The personal data that is being processed might belong to special categories and might relate to criminal convictions and offences but we will discuss such cases separately below so we will assume for now that it doesn't (cf. lit. c) as is the case in our first example.

The consequences for the data subject^{ev} in post hoc *analysis* is likely negligible, but they might be quite dramatic in post hoc *modification*. Still, the consequences are not more severe than they were in the original processing (one might get a poor grade in math, or be graded as a poor physician). Post hoc *analysis* might, however, have real consequences for the data subject^{EV} since it might reveal him or her to be biased with the possible effect of stigmatization if this becomes known. It should be noted though that such consequences are already possible for the data subject^{EV} if their evaluation is overtly biased. Still, careful analysis makes detection of bias more likely. It thus becomes important that the result of the analysis is safeguarded appropriately (e.g. with encryption). The consequence of post hoc *modification* for the data subject^{EV} might be either that they are informed that a bias has been detected and it is suggested to them they update, or that their evaluation is overridden (Jönsson forthcoming). Both might of course cause the data subject^{EV} concern, but since GIIU attempts to help the data subject^{EV} make more accurate decisions, this concern shouldn't lead us to think that post hoc

[19] At least for post hoc analysis, it seems plausible to say that the kind of statistical analysis carried out there falls within the reasonable expectations of an average data subject (both Mike and his students in the first example for instance). Post hoc modification is more questionable since the kind of modification being carried out there is likely not expected by most data subjects. However, if the modification is not automated but merely suggested to the evaluator, it seems more plausible that it falls within the subjects' reasonable expectations.

modification goes against the purpose for which the data was originally collected. (cf. lit. d)

For the purposes of post hoc analysis anonymization or pseudonymization is possible with respect to the data subject who is being evaluated (i.e. moving from M1 to M2), although this is not possible from the perspective of automatic post hoc modification (we have to know who should be updated). For both kinds of processing, we can use encryption in the processing. (cf. lit. e).

Jointly, these five considerations seem to point towards the further processing being compatible with the purposes of the original processing. Both the legitimacy and the compatibility of the processing can, however, depend on appropriate safeguards to ensure that the personal data is not misused for other purposes. For example, as have been mentioned above, that new knowledge about data subject^{EV} is not used by the employer to evaluate him or her, since such further processing would go beyond the original purpose and would require a separate analysis to determine whether it is lawful under the GDPR.

It thus seems to us that applying post hoc analysis and modification to the personal data like that in our two cases can be compatible with the GDPR even if the data has not been collected explicitly for that purpose.

5.2 The Risk of Inducing Error

According to the principle of accuracy, personal data shall be ”accurate and, where necessary, kept up to date” (Article 5(1) lit. d of GDPR). Although GIIU aims to improve accuracy, it is statistically possible, although unlikely, that it will instead decrease accuracy. This is another point of tension between GIIU and the GDPR.

The GDPR mandates in the present context, that the controller must ensure that appropriate safeguard measures are in place to minimize the risk that GIIU induce errors when personal data (e.g., grades) are modified. This is even more important if modifications are automated as the impact of batch processing might potentially affect a much larger number of data subjects (see also below Section 5.3). Safeguards could include mechanisms that detect deviations from the conditions under which GIIU will work as intended (cf. Jönsson and Bergman 2022) or other appropriate statistical measures. The transparency of the processing in relation to the data subject is also imperative, since such transparency makes it possible for him or her to review the accuracy of the processing and request that inaccurate results are rectified pursuant to Article

16 of the GDPR. The controller can, however, not solely rely on such data subject review, but must regularly perform its own audits to detect and rectify inaccurate data. Personal data is considered inaccurate when it is unfit for the purpose of the processing.[20] Hence, in relation to historical data, it must be considered whether later changes need to be reflected in the data set that is used for the post hoc intervention assessment. Provided that such safeguards are in place to detect and rectify inaccuracies both in the data that is used as the basis for the assessment and the resulting modifications, such interventions should be compatible with the GDPR.

5.3 Automated Decision-Making and Profiling

GIIU can be implemented in different ways that range from manual to fully automated processing of data. In practice, the processing is likely to use some automated processing, but, as was mentioned in the introduction, the final decision to correct the grade or evaluation could be left to the evaluator. GIIU would then only act as a decision support tool. It would, however, also be technically possible to fully automate the procedure in such a way that it would not involve any human intervention. Such processing would bring GIIU into tension with the GDPR in a further way.

This kind of processing would have to comply with Article 22 of GDPR, which contains rules on automated individual decision-making. The provision is applicable when a decision based solely on automated processing produces legal effects concerning an individual or similarly significantly affects him or her. The latter includes decisions affecting someone's employment opportunity.[21]

The interpretation of the rule is disputed, but it either generally prohibits automated decisions falling within its scope or at least gives the data subject a right to object to such processing (Drożdż 2020). It is also disputed whether the rule only covers profiling or any automated decision-making.[22] In this

[20] See Article 5(1) lit. d and Article 16 of GDPR, which both state that the accuracy of personal data should be determined having regard to the purposes for which they are processed.

[21] Article 29 Working Party, Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679, WP251, adopted 6 February 2018.

[22] See Article 4(4) of GDPR, which defines "profiling" as "any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural

context, it is relevant to note that GIIU does not intend to make any analysis or predictions about data subject[ev], but instead assess the potential bias of data subject[EV]. Hence, the person that is potentially being profiled is not the person affected by the automated decision (i.e., the post hoc modification). With regard to the legal uncertainties surrounding Article 22 of the GDPR, controllers should carefully assess whether a fully automated GIIU procedure is permitted and where required acquire consent from the relevant data subjects and implement appropriate safeguards (see also Section 6 on the need for Data Protection Risk Assessment).[23] Since this depends on the context the assessment has to be made on a case-by-case basis.

6. The Processing of Special Categories of Personal Data

From the above discussion we can gather that post hoc processing of personal data can be in compliance with the GDPR even if the data subjects haven't given their consent to this processing. However, this doesn't take into account the possibility of the data containing 'special categories' of personal data (also known as "sensitive personal data"). Concerning such categories the GDPR maintains the following:

> Processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation shall be prohibited.
>
> (Article 9 (1) of GDPR)

What this means is that the processing of such sensitive personal data is prohibited unless the processing is subject to one of the exceptions stipulated

person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements".

[23] See Article 22(2) lit. c of GDPR, which stipulates that such consent must be explicit. Consent is required unless the automated decision-making is permitted by law or necessary to conclude or perform a contract between the controller or the data subject. Article 22 para. (2) and (3) also stipulates that suitable safeguards must be implemented. This includes but is not limited to the rights to require human intervention.

in point 2 of the same article. The first thing to note is that gender and age are not special categories (as described by Article 9(1)), and our first case thus remains unproblematic.[24] Our second case, however featured skin color which likely falls under "revealing racial or ethnic origin". An evaluation of the exceptions available in Article 9(2) of the GDPR shows that most are usually not relevant for our case, although some might be applicable in certain special cases. It is for example possible that some post hoc interventions could be viewed as being necessary for the purposes of carrying out the obligations and exercising specific rights of the controller or of the data subject in the field of employment law (cf. lit. b). For instance, in situations where pro-active work is required to avoid discrimination or improve diversity. In general, the controller would, however, have to rely on the explicit consent of the data subject (cf. lit a).

Does the post hoc processing involving sensitive personal data thus require explicit consent from the data subjects (both the evaluator and the evaluated)? To answer this we need to treat post hoc analysis and post hoc modification separately.

As we saw above, post hoc analysis only requires personal data of the following form (here illustrated with an example from our second case).

(P2) Phyllis Berry has given a black applicant the competence assessment 3.

Not the more inclusive P1

(P1) Phyliss Berry has given Donald Glover the competence assessment 3.

The name of the applicants (i.e. the evaluated) are not needed in order carry out this processing. But since skin color is a property of the data subject^{ev} (and not the data subject^{EV}, i.e. the evaluator), P2 does not fall under the requirements of Article 9. The reason for this is that although the sensitive data is part of the data set being processed it does not "relate to" the data subject, i.e., data subject^{EV} (cf. Article 4(1) of GDPR). The GDPR aims to protect the rights of the data subject, not the data as such. Hence, the data must be sensitive to the data subject to fall under Article 9 and not merely sensitive in nature, since it would otherwise not

[24] This can, however, depend on the context of the processing. See Judgment of the Court of Justice of the European Union of 1 August 2022 in Case C-184/20, Vyriausioji tarnybinės etikos komisija (ECLI:EU:C:2022:601). The Court found that name-specific data relating to someone's spouse, cohabitee or partner can reveal sexual orientation and fall under Article 9 of the GDPR even where that was not the intention of the processing.

be able to create the kind of special risks for the data subject that is the object of the prohibition in Article 9. This means that consent is still not required for post hoc analysis even if processes data that are derived from personal data relating to special categories. The exception to this would be cases where the black applicant in P2 could be identified indirectly, e.g. through being one of very few black applicants in the history of evaluations of Phyllis.

However, in the fully automated post hoc *modification* we need to know who we should update and we cannot anonymize (or even pseudonymize). There thus seem to be no escaping the need to ask for explicit consent pursuant to Article 9 of the GDPR from the evaluated persons for this step to be in compliance with the GDPR. However, if we consider the variant of GIIU which acts as an advisory decision support system, one can envision it only generating general recommendations concerning members of certain social categories to the evaluator, e.g. "It looks like your grades for female students are 1 point lower than what is to be expected". This would not require the processing of personal data concerning data subject^{ev} and thus not of sensitive personal data, and hence there is no need to ask for explicit consent.

In the above we have assumed that no sensitive data relating to data subject^{EV} is processed during the intervention. Since such data (e.g. grades) would constitute personal data (also relating to the evaluator) and cannot be anonymized, we also need to assess whether evaluative judgements can constitute sensitive data under Article 9(1) of the GDPR. At first glance, this does not seem to be the case. It should, however, be noted that the CJEU has held that Article 9 should be given a fairly wide interpretation.[25] It could for instance not be ruled out that processing of data concerning bias or discriminatory practices – given an extensive interpretation – could be considered personal data "revealing … political opinions". In our view this interpretation is too extensive since the processing merely creates an abstract risk of revealing someone's political opinions.[26] It would be another matter if this was the purpose of the processing, but that is not the aim of post hoc interventions.

[25] See Judgement of the Court of Justice of the European Union of 6 November 2003 in Case C-101/01, Lindqvist (ECLI:EU:C:2003:596), para. 50. See also the Judgement of 1 August 2022 in Case C-184/20, above footnote n.

[26] For a similar argument see Judgement of the German Administrative Court of Mainz of 20 February 2020 case no. 1 K 467/19.MZ, ECLI:DE:VGMAINZ:2020:0220.1K467.19.00. The court held that merely the abstract risk of the transmission of a (zoonotic) disease from an animal to a human did not mean that such data in general should be viewed as "data concerning health" relating to the animal's owner under Article 9(1) of the GDPR.

7. Risk Assessment and Data Protection Impact Assessment

The GDPR requires that controllers assess the risk for data subjects' fundamental rights prior to processing their personal data. When there is an indication that this processing is likely to result in high risks, pursuant to Article 35 the controller must conduct a formal Data Protection Impact Assessment (DPIA). This is particularly so when the processing involves new technologies. Post hoc interventions have a well-documented basis in the literature, but different implementations are still to be tested and applied in practical decision-making. Guidance on when processing is likely to result in a high risk can be found in the Article 29 Working Party's Guidelines on Data Protection Impact Assessments.[27] The guidelines have been endorsed by its successor the European Data Protection Board. Such guidance has also been adopted by the national data protection supervisory authorities. Since implementations of GIIU might involve new technologies and involve processing of both sensitive personal (e.g. ethnic origin) data and data concerning children (e.g. grades of schoolchildren), it is likely that such processing in the light of these guidelines falls within the scope of Article 35. This is so in particular when the processing consists of automated decision-making, including profiling (see above Section 5.3). Hence, organizations that intend to implement automated post hoc intervention procedures must perform a DPIA and when the assessment indicates that the processing would result in a high risk in the absence of measures taken by the controller to mitigate the risk also consult the competent supervisory authority (Article 36 of GDPR).

8. Conclusion

Even though restrictive on the face of it, the GDPR seems to be compatible with post hoc interventions being applied in cases like the first of our two examples (featuring Mike) even without the consent of the data subjects (either evaluators or the evaluated) specifically for this processing. Similar considerations apply in legally similar contexts such as government agencies.

[27] Article 29 Working Party, Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is "likely to result in a high risk" for the purposes of Regulation 2016/679, WP248, adopted on 4 October 2017.

Exceptions arise only if the personal data is of sensitive nature or if the processing is automated. So, for instance, if our second example features a version of GIIU where processing is fully automated, or features sensitive data, post hoc *modification* would require explicit consent, even though post hoc analysis would not.

Acknowledgments

The research was funded by a research grant from the Swedish Research Council (Dnr. 2017-02193) and research funding provided by the Pufendorf IAS in Lund. The text has benefitted from discussion with and/or careful reading by members of Post Hoc Interventions Pufendorf theme, as well as the participants at the conference, Post Hoc Interventions: Prospects and Problems, organized in Lund in October 2022.

References

Article 29 Working Party, Opinion 4/2007 on the concept of personal data, WP136, adopted on 20 June 2007.

Article 29 Working Party, Opinion 5/2014 on Anonymisation Techniques, WP216, adopted on 10 April 2014.

Article 29 Working Party, Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is "likely to result in a high risk" for the purposes of Regulation 2016/679, WP248, adopted on 4 October 2017.

Article 29 Working Party, Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679, WP251, adopted 6 February 2018.

Bergman, J. and Jönsson, M. L. (submitted) "Gender Bias in Grant Applications. Inquiry and the Potential for a Post Hoc Remedy". Manuscript.

Bradford, A. (2021) *The Brussels Effect: How the European Union Rules the World.* New York, NY.: Oxford University Press.

Craig P. and de Búrca, G. (2020) *EU Law: Text, Cases and Materials (7th ed.).* Oxford: Oxford University Press.

Drożdż, A (2020) *Protection of Natural Persons with Regard to Automated Individual Decision-Making in the GDPR.* Alphen aan den Rijn: Kluwer Law International.

European Data Protection Board, Guidelines 3/2018 on the territorial scope of the GDPR (Article 3), adopted on 12 November 2019.

European Data Protection Board, Guidelines 7/2020 on the concepts of controller and processor in the GDPR, adopted on 7 July 2021.

Jönsson, M. L. (2022) "On the Prerequisites for Improving Prejudiced Ranking(s) with Individual and Post Hoc Interventions" *Erkenntnis*.

Jönsson, M. L. and Sjödahl, J. (2017) "Increasing the veracity of implicitly biased rankings". *Episteme 14*(4), 499–517.

Jönsson, M. L. and Bergman, J. (2022) "Improving misrepresentations amid unwavering misrepresenters", *Synthese*, *200*.

Skolverket (2019) *Analyser av likvärdig betygssättning mellan elevgrupper och skolor.*

Skolverket (2020) *Analyser av likvärdig betygssättning i gymnasieskolan.*

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

8

Post Hoc Interventions and Machine Bias

Kasper Lippert-Rasmussen[1]

Abstract. In a US context, critics of court use of algorithmic risk prediction algorithms have argued that COMPAS involves unfair machine bias because it generates higher false positive rates of predicted recidivism for black offenders than white offenders. In response, some have argued that algorithmic fairness concerns calibration across groups – roughly, that a score assigned to different individuals by the algorithm involves the same probability of the individual having the target property across different groups of individuals – only. I argue that in standard non-algorithmic contexts, such as hirings, we do not think that lack of calibration entails unfair bias, and that it is difficult to see why algorithmic contexts, as it were, should differ fairness-wise from non-algorithmic ones. Hence, we should reject the view that calibration is necessary for fairness in an algorithmic context and be open in principle to post hoc interventions counteracting differential false positive rates.

1. Introduction

It is widely acknowledged that certain groups of people are disadvantaged across a wide range of contexts as the result of unfair biases working to their disadvantage. Traditionally, it has often been assumed that the biases are known to the bearers. However, much recent research focuses on "implicit biases" involving automatic dispositions of which, sometimes, the agent is

[1] Kasper Lippert-Rasmussen, Professor of Political Science, Department of Political Science, Aarhus University. Head of The Centre for the Experimental-Philosophical Study of Discrimination (CEPDISC).

unaware. Indeed, some implicitly biased agents will strongly disavow the biases their behavior manifests when questioned about them.[2]

While most people, at least at some point of their lives, will belong to at least one group with biases working against it, some people belong to many such groups all their lives. Biases are stronger against some groups than others. Some are active across a wider range of contexts than others, and sometimes biases are mutually enforcing. Because biases often result in undesirable, e.g., because unjust, outcomes, it is generally agreed that sometimes, at least, we ought to intervene to mitigate or prevent their effects.[3] Such interventions can be ante hoc or post hoc. *Ante hoc* interventions concern biases themselves or their manifestation in behavior such as decision-making. The aim is to make the biases less common or to reduce their influence on behavior. *Post hoc* interventions take biases and their manifestation in behavior as parametric and aim to reduce the degree to which these result in undesirable outcomes.[4]

Post hoc interventions, which form a big family, differ in various ways. First, they differ in terms of the means adopted to avoid the relevant undesirable outcomes, e.g., quotas, or an adjustment of qualification scores to counteract evaluators' known biases. Second, they differ in terms of the sort of undesirable outcome they seek to avoid. Thus, in a series of recent articles, Martin Jönsson (2022), and Martin Jönsson and co-authors (2017, 2022), have focused on post hoc interventions seeking to reduce the inaccuracy of rankings produced by biased evaluators. By contrast, traditional affirmative action interventions such as quotas are sometimes intended to reduce unjust inequality of opportunity (Lippert-Rasmussen 2020, 72-102). The undesirable outcome I shall focus on here – differential false positive rates – is neither of these, but it is one that has received considerable attention in recent discussions of algorithmic fairness.

I begin, in Section 2, by describing the well-known controversy over COMPAS. There, critics have argued that black offenders are victims of

[2] For philosophically informed overviews of the implicit bias literature, see Beeghly and Madva (2020); Brownstein 2019; Brownstein and Saul (2016).

[3] To agree with this is not to say that we would have no reason to counteract such biases if they did not result in undesirable outcomes.

[4] Ante hoc and post hoc interventions can supplement each other, and perhaps in many cases the aim informing either intervention can be achieved only by adopting both. However, doubts about how successful an ante hoc intervention is introduce doubts about what the correct post hoc intervention is (but see Jönsson and Bergman 2022, 12, 21). Nothing in what I say below hangs on whether the two interventions supplement each other.

machine bias in that the recidivism risk prediction algorithms burden them with a higher rate of false positives (roughly: inaccurate predictions that an offender will reoffend) than white offenders face.[5] An obvious post hoc intervention in which judges are instructed or advised to draw different conclusions from a given risk score depending on whether the offender is black or white could, potentially, mitigate that problem. Yet, many think such an intervention, resulting in a deviation from the guidance provided by a well-calibrated risk assessment would be unfair to white offenders.[6] Section 3 briefly explores the implications of a commonly held view about unfair bias on the job market in light of audit studies and the conceptual apparatus introduced in Section 2 in relation to COMPAS. The section explains that in a job market where, because of past sexist discrimination, men are more likely to be qualified for certain jobs, deeming an applicant to be qualified means different things across male and female applicants, since there is greater chance of being qualified for the former. Many, this author included, would see no fairness-based reason in this situation for a post hoc intervention to secure a well-calibrated hiring process.[7] Thus, Section 3 ends with a trilemma consisting of three claims: 1) Lack of calibration does not amount to unfair bias in job markets; 2) Job markets and sentencing do not differ as regards whether a lack of calibration amounts to unfair bias; 3) Lack of calibration amounts to unfair bias in sentencing. Plainly, we must reject at least one of these claims, so the following sections (4-6) go through each of them in turn, asking which should be abandoned. Section 7 concludes.

---

[5] False positive rates are defined as: False Positives (FP)/Actual Negatives=FP/True Negatives (TN) + FP. False negative rates are: False Negatives (FN)/True Positives (TP) + FN. See also Table 1 below.

[6] As will become clearer shortly, the post hoc intervention in question here might not be one that presumes people are psychologically biased and then seeks to mitigate the degree to which that bias translates into differential outcomes for different groups (cp. Jönsson and Sjödahl 2017, 500). Rather, it may seek to mitigate the extent to which differential recidivism base rates, through what in the literature is referred to as machine bias, are turned into differential unjust outcomes by seemingly – so the criticism goes – unfair algorithms. This shows that there can be a rationale for exploring post hoc interventions even in the absence of implicit psychological biases that are difficult, very costly, or even impossible to eliminate. In short, the justification for exploring post hoc interventions is robust regarding the manipulability of implicit psychological biases.

[7] This point relates to a point made in Thore Husfeldt's article (this volume) that if we give everyone equal odds, we will not get demographic parity unless we have equal base rates across different groups. However, Husfeldt is agnostic on the implications of this for concerns about fairness.

In a nutshell, I argue, *first*, that we should, as it were, bring what we think of algorithmic fairness into line with what we think about job market discrimination in an ordinary non-algorithmic setting. That result I am quite confident of. How we should do it, i.e., how we should resolve the trilemma, I am less clear about. However, I offer some reasons suggesting, *second*, that in certain cases involving differential base rates – in principle, at least – we should allow post hoc interventions to equalize false positive/negative rates even if that means violating calibration. These are the two main claims in this article.

2. COMPAS and Calibration

I start, then, with a thumbnail sketch of the COMPAS debate. COMPAS, which stands for Correctional Offender Management Profiling for Alternative Sanctions, uses information about, among other things, an offender's employment and housing status, personality traits and criminal record to arrive at a risk of recidivism score – basically, a number from 1 (least likely) to 10 (most likely) indicating how likely it is the offenders will recidivate relative to other offenders – which is used by the courts in sentencing. It does not use information about race. Higher scores, indicating a greater likelihood that the offender will reoffend, will generally lead the courts to sentence offenders to longer periods of incarceration than they would be given with lower scores.[8] Hence, a false positive is a bad thing for an offender and a false negative is a good thing.[9]

In a renowned article entitled "Machine Bias" in *ProPublica*, Angwin et. al. (2016) suggested that COMPAS is unfair because it is racially biased. Like other ways of assessing the risk of recidivism, e.g., simply relying on the judge's impression of the offender and a statement from a psychiatrist, COMPAS is far from perfectly accurate.[10] In some cases, it predicts it to be

[8] Some might object to this sentencing practice on the grounds that it involves sentencing offenders on bases other than the crime committed. I set aside the issues raised by this complaint, noticing though that in most jurisdictions assessments of an offender's dangerousness can play a lawful role in sentencing. In any case, COMPAS is also used for other purposes than sentencing, e.g., decisions about bail.

[9] For a useful and insightful description and analysis of the case, see Hellman (2020).

[10] According to ProPublica, COMPAS was only "somewhat more accurate than a coin flip". Whether it is more accurate than standard assessments of risk of recidivism is an important question given that such assessments, in some form or another, play a role in determining the level of punishment.

highly likely that an offender will reoffend and in fact they do not (false positives).[11] In other cases, it deems it highly unlikely that the offender will reoffend and in fact they do (false negatives). What is striking is that even though, overall, COMPAS is equally accurate in making correct predictions across black and white offenders, its false positive and false negative rates differ across white and black offenders.[12] COMPAS is more likely to misclassify a non-recidivating black offender (44.9%) than a non-recidivating white (23.5%) offender as dangerous, and it is more likely to misclassify a recidivating white offender (47.7%) than a recidivating black (28.0%) offender as not being dangerous. This seems unfair, because it seems that sentencing based on COMPAS treats black offenders (upon whom it imposes a greater risk of an unduly long incarcerations) worse than white offenders (whom it privileges with a greater prospect of an unduly short period of incarceration).[13] At any rate, this was the intuitively forceful complaint set out in the "Machine Bias" paper.

In response to this criticism, Northpointe – the company that sells COMPAS to US courts – conceded the factual basis of Angwin et. al.'s criticism. However, it replied that COMPAS is well calibrated across black and white offenders. Essentially, in the case at hand this means that, for any given risk score, the probability that the offender will recidivate is the same whether the offender is black or white. Or, to put this in more general terms, which will be helpful later in Section 3: for each possible score, the (expected) percentage of individuals assigned this score who are positive is the same for each relevant group.[14] Calibration across groups, Northpointe submitted, is necessary and sufficient for algorithmic fairness.

---

[11] Strictly speaking, COMPAS' risk scores are ordinal, not cardinal. A high-risk score simply indicates that the offender belongs to a percentile of offenders who are more likely to reoffend than offenders from most other percentiles, not that the offender is very likely to recidivate (though, as a matter of fact, they do).

[12] In fact, Angwin et. al. used a finer-grained taxonomy of racialized groups, but for present purposes this makes no difference.

[13] What, exactly, (un)fair treatment amounts to is complex. Here I shall simply assume that differential treatment of the sort involved here is unfair. I return to these issues in Section 7.

[14] Or to put this requirement differently: TP/Predicted Positives=TP/FP + TP is the same across different relevant groups (compare footnote 6). There is a further requirement often labelled a requirement of calibration, i.e., that, for each group, the risk score is equal to the percentage of individuals who are assigned this risk score and reoffend. Since my focus here is on fairness to individuals across different groups, this aspect plays no role in my argument.

Several theorists have offered at least partial support for this response. For instance, Brian Hedden (2021, 227) writes: "none of the statistical criteria considered in the literature are necessary conditions for algorithmic fairness, except Calibration Within Groups". Similarly, Robert Long (2020, 4, 17) submits that "when appropriate decision thresholds have been set, calibration is a necessary condition for procedural fairness … false positive [KLR: and negative] rate inequality is not, in itself, a measure of unfairness".

One interesting point emerging from the burgeoning literature on algorithmic fairness of recent years is that, other than in special circumstances,[15] when two groups differ in terms of their base rates – as they do in the present case, since the frequency of recidivism is, as it happens, higher for black American offenders than it is for white American offenders – it is mathematically impossible for a predictive algorithm to be *both* well-calibrated across groups *and* have equal false negative and false positive rates across groups.[16] This insight has given rise to a substantial debate, involving computer scientists, philosophers and others, over the right criteria of algorithmic fairness.

Another important point is the following. In effect, if we accept the criticism levelled by Angwin and colleagues, we are committed to the view that there is at least a pro tanto reason in favor of a post hoc intervention to prevent the "machine bias" of COMPAS from resulting in unfair, unequal positive rates across white and black offenders. For mathematical reasons, such an intervention would involve giving up on calibration by adjusting the way COMPAS risk scores are assigned such that, for a given high risk score, it takes more predictors of recidivism for a black offender than for a white offender to be assigned this risk score (and the reverse for low risk scores), the result being that a higher proportion of black than white offenders who are assigned a low risk score will recidivate, i.e., the reverse situation of what was the case in 2016.[17] To explore whether such a post hoc intervention involving violating of

[15] For example, those where the predictive algorithm is perfect.

[16] For an excellent overview of the debate, and of various impossibility results, that is accessible to mathematically less sophisticated readers, see Hedden (2021; see also Eva 2022).

[17] As many contributors to the literature emphasize, the unequal base rate claim is problematic in various ways. What is known is the rate at which offenders are charged or convicted, not the rate at which they reoffend, and biases boosting charging or conviction rates in the case of black offenders might explain why those offenders face a higher risk of being convicted of further offenses in the future even if recidivism base rates are identical across white and black offenders. To the extent that such biases shape the base rates of black and white offenders, the relevant post hoc intervention would still qualify as a post hoc intervention, albeit arguably not one that

calibration is desirable in the present case in principle at least, I want to consider one that is similar but raised in the different and, it would seem, well-examined non-algorithmic context of *discrimination in hiring*.

3. Post Hoc Interventions in the Job Market

There is a well-established literature on bias in hiring. In this, so-called audit studies[18] present survey experiments in which one independent variable, such as race or gender, is altered to reveal the effect of doing that. For instance, the experimenters might send out a large number of job applications with accompanying CVs. These will be identical except for the applicant's name, which in half of the applications strongly suggests the applicant is a man and in the other half strongly suggests the applicant is a woman. If the subsequent call-back rates vary, with, say, male-looking applicants getting more calls than female-looking ones, then, other things being equal, the audit study will conclude that female applicants, in the sector being examined, are subjected to (unfair) bias. If there is no difference in call-back rates, it will conclude that there is no (unfair) gender bias in the call-back phase of hiring (which, of course, is not to say that there might be no unfair gender bias in later phases. Whether there is can also be studied through audit studies).[19] What I now want to consider is:

> *Job Market*: There are 500 male and 500 female applicants for a certain position. As a result of past sexist discrimination preventing female applicants from acquiring the much-needed work experience, 180 of the male applicants

counteracts machine bias as opposed to (explicit or implicit) psychological biases exhibited by people (e.g., police officers who are more inclined to charge black people than white people).

[18] For some prominent examples, see Neumark (1996), Banerjee et. al. (2009), Widner and Chicoine (2011), Gaddis (2014), Pager and Quillian (2005).

[19] Or, more precisely, the audit study will conclude that there is no (unfair) *direct* bias in hiring. An audit study does not speak to the question of whether the requirements of the job are unfairly, indirectly discriminatory. Note also that the two inferences in question are not as straightforward as one might think, because the information provided in identical texts with differently gendered names might be different. For instance, in a sexist society information about a 9-month parental leave period will be interpreted differently depending on whether the applicant is male or female and thus differential responses might be informed by factors other than the mere gender of the applicant (see Hu forthcoming).

> are qualified, while only 20 female applicants are.[20] The hiring procedure is such that an audit study will conclude that it makes no difference whether the applicant is male or female and, thus, that there is no unfair gender bias in the hiring procedure – all other things being equal, for any hired and any non-hired applicant exactly the same outcome would have occurred had this applicant had a different gender. Hiring is conducted in a non-algorithmic way: I shall say more on this later, but briefly, it means that the members of the hiring committee look at the applications using their judgment and informal deliberation to form an opinion about who is, and who is not, qualified. As the audit study informs us, the hiring committee is unbiased, gender-wise, in its assessments. Finally, the hiring committee's assessments are quite accurate, but not perfect. If an applicant, whether male or female, is qualified, there is a 90% chance the committee will deem them to be qualified. If the applicant is unqualified, there is a 90% chance the committee will deem them unqualified.

Job Market, as described, gives:

Table 1: Confusion table

	In fact: qualified	**In fact: not-qualified**	
Prediction: qualified	162 (men)/18 (women) True Positives (TP)	32/48 False Positives (FP)	194/66 (260)
Prediction: not-qualified	18/2 False Negatives (FN)	288/432 True Negatives (TN)	306/434 (740)
	180/20 (200)	320/480 (800)	500/500

Since my aim is to compare fairness judgments in ordinary hiring contexts with fairness judgments in relation to machine bias, let me describe this situation in the language of COMPAS. Basically, it is a situation where the assessment of the applicants is not well-calibrated despite the fact that an audit study will conclude that the procedure involves no unfair bias. That is, the ascribing of the values "qualified" and "not-qualified" to the applicants does not, as it were,

[20] The assumption that the difference in base rates reflects past unjust discrimination is not essential to my argument, but it has certain presentational advantages, one being that, for some readers, it might make such a difference (see the discussion of compounding injustice below).

have the same meaning across gender.[21] If the hiring committee finds that a particular applicant is qualified, that implies that there is a greater chance that the applicant is qualified if the applicant is male (162/194) than there is if she is female (18/66).[22] However, the hiring procedure will involve equal false-positive and false-negative rates across gender, reflecting the fact that if an applicant is (un)qualified, then in 90% of those cases the committee will deem the applicant to be (un)qualified. Take, first, false positive rates. In the case of male applicants, 32 men are falsely predicted to be qualified (False Positives) relative to 320 who are unqualified (Actual Negatives). In the case of female applicants, 48 are falsely predicted to be qualified (False Positives) relative to 480 who are unqualified (Actual Negatives). So, the false positive rate is 10% for both male and female applicants.[23] Now take false negative rates. In the case of male applicants, 18 are falsely predicted to be unqualified (False Negatives) relative to the 180 who are qualified (Actual Positives). In the case of female applicants, 2 are falsely predicted to be qualified (False Negatives) and 20 are in fact qualified (Actual Positives). The false negative rate is therefore again 10% for both male and female applicants.

In the light of COMPAS, the interesting feature of Job Market is this. According to standard audit studies, there is no unfair bias in the Job Market hiring process.[24] Yet the hiring procedure is miscalibrated and involves equal false positive and false negative rates. On the face of it, a post hoc intervention to reduce miscalibration – e.g., by hiring a greater proportion of the men deemed qualified than of the female applicants deemed qualified – would not be a way of counteracting unfair bias. The message seems to be that in ordinary non-algorithmic hiring contexts with different base rates across different groups of applicants we should not worry about lack of calibration as

[21] The sense of "meaning" used here, and which is commonly used in the algorithmic fairness literature, is different and much more practically oriented than that involved when philosophers discuss the meaning of a term. In that sense, the fact that the same criteria are used across men and women to determine whether an individual applicant is (un)qualified implies that "(un)qualified" means the same whether it qualifies a male or a female candidate, e.g., "(un)qualified" applied to men and women has the same sense (in Frege's sense).

[22] In short: TP/FP + TP is higher for male and female applicants.

[23] In short: FP/TN + FP is the same for male and female applicants.

[24] According to Brian Hedden (2021, 225-226): "<Lack of calibration> seems to amount to treating individuals differently in virtue of their differing group membership". In Job Market, lack of calibration amounts to exactly the opposite, i.e., to not treating applicants based on their differing group membership; indeed, achieving calibration requires doing just that.

explained, but we should, possibly, worry about unequal false positive/negative ratios as such.

Assuming these claims reflect a correct assessment of the case at hand, this suggests that Northpointe's defense of COMPAS is mistaken, and that fairness might require a post hoc intervention of the sort entertained above. That is, there is no algorithmic fairness objection to white offenders with a risk score equal to that of black offenders having a lower risk of reoffending, because calibration is not a necessary condition of algorithmic fairness.

In light of reflections like these, the following claims seem plausible:

> (1) Lack of calibration does not amount to unfair bias in job markets (the *Standard View*).
>
> (2) Job markets and sentencing do not differ as regards whether lack of calibration amounts to unfair (direct) discrimination (the *Equivalence Claim*).
>
> (3) Lack of calibration amounts to unfair (direct) discrimination in sentencing (the *Northpointe View*).

Admittedly, though I say the Equivalence Claim is plausible, I have so far said nothing to justify it. I will do so shortly. What we can see already, however, is that *if* we embrace it, we are obliged to abandon one of the other two claims: we must *either* stop assuming – as audit study encourages us to do, and as I think that many people do, in effect, unreflectively – that lack of calibration reflecting differential base rate qualifications does not render ordinary hiring procedures unfairly biased *or* reject the Northpointe View that lack of calibration in sentencing amounts to unfair bias. This obligation arises from the fact that claims (1)–(3) are trilemmatic: from any pair of them we can derive the negation of the third. So the wider question is: Which of the three claims should be dropped? With this question in mind, I will assess the three claims in turn over the next three sections.

4. Rejecting the Standard View

Should we reject the Standard View of unfair bias? A response to this question that I have heard on several occasions is that audit studies, at any rate, appear to present no obstacle to doing so. The thinking here is that audit studies usually include a *ceteris paribus* clause implying that information about, say, gender or race has no probative value. However, in Job Market information

about gender does have such value, so the *ceteris paribus* clause would be unsatisfied in this case.

I have two thoughts about this response. First, we can simply stipulate that the employer in Job Market has no information about the relevant baseline differences, in their qualifications, between male and female applicants. This would mean that gender has no probative value, and that the *ceteris paribus* clause is satisfied. Yet our assessment of the case – no unfair bias – would remain, I submit, the same. Second, the fact that audit studies often apply an "other things being equal" clause favors the retention of the Standard View. The clause is meant to accommodate cases in which the employer believes that information about identity has probative value, not cases where such differences exist. Indeed, these clauses cover cases where, in fact, there are no base rate differences, in their qualifications, between, say, male and female applicants (same mean, same distribution etc.), but where the employers reasonably, but incorrectly, believe that such base rate differences obtain. In principle, once that is factored into an audit study, it might still conclude that there is no unfair discrimination despite lack of calibration (or, for that matter, lack of false positive rates).

What about the positive case for retaining the Standard View? One way to build that case is by pointing out that rejection of the view has implausible implications. Imagine that we tweak the hiring procedure in Job Market in favor of male applicants – e.g., applying the rubric "Give an extra five points for male gender" – so that in the case of equally qualified male and female applicants the male applicant is more likely to be deemed qualified. Even so, on the present view male applicants can have a complaint about unfair bias against them, because while the extra points mitigate miscalibration, they do not rule out the possibility that a male applicant deemed qualified is more likely to be qualified than a female applicant deemed qualified. However, it is quite unappealing to think that male applicants in these circumstances – circumstances, that is, involving a hiring procedure boosting their qualification score on grounds of their gender – can complain about unfair gender bias *against* them. If anything, intuitively, they benefit from unfair bias.

We might also ask: Who can have a fairness complaint about lack of calibration in Job Market?[25] Arguably, the answer to this question will depend

[25] I assume that only individuals have morally relevant complaints. This assumption is consistent with the view that individuals have complaints about how they are treated qua members of specific groups. It is also consistent with the view that, in a derivative sense, groups can have complaints, i.e., those deriving from the complaints of their members.

on what the alternative hiring procedure is. If the alternative is a procedure in which calibration is secured, then presumably those men who are presently deemed unqualified but would be deemed qualified with calibration might have a complaint.[26] How much moral weight this complaint would have will depend on how much weight we should attach to the fact that most of these men are not qualified. It may seem problematic to suppose that one is being subjected to unfair bias when one is not deemed qualified if, in fact, one is not qualified – especially, if one even enjoys a better chance of being deemed qualified than equally, or even better, qualified female applicants.[27] In any case, a complaint of this sort will have to be weighed against the complaint of those qualified women who, because of calibration, have a lower chance of being hired.[28] I recognize that these consideration are inconclusive, but in view of how we normally think of fairness – at least in the form of procedural justice – in cases of the kind I have been looking at, I fail to see that the complaints of the men in question are decisive.

5. Rejecting the Equivalence Claim

Are the COMPAS and Job Market cases different in that in the former the consideration of fairness gives us reason to be concerned about whether calibration is satisfied, whereas in the latter that same consideration gives us no reason to be concerned about lack of calibration? I take it the burden of proof here is on those who think the cases differ.[29] Hence, in defending the

[26] For simplicity, let us assume that who is deemed qualified does not change in surprising ways – e.g., a female applicant who is deemed unqualified with lack of calibration is deemed qualified in the presence of calibration.

[27] One option here is to reject the meritocratic view that fairness requires people to be hired on the basis of their qualifications. There is a real debate here (Lippert-Rasmussen 2020, 230-252). But for present purposes it is not especially interesting, because rejecting it would seem to undermine the case not only for equal false positives/negatives ratios but also (qualification-based) calibration.

[28] If only a subset of the applicants is deemed qualified, the male and female applicants who are deemed qualified and are so also have a complaint against calibration, since calibration will reduce their risk of not being hired as a result of the greater number of unqualified males being deemed qualified.

[29] Unlike jobs, the number of years of incarceration one is being sentenced to is not a positional good. Positional goods are special in the sense that if one gets the good, others are excluded from it and have a lower chance of enjoying a good of this kind. Plausibly, fairness

Equivalence Claim I shall merely rebut some suggestions as to why they are different. This will amount, I realize, to an inconclusive argument in favor of the equivalence claim. Still, if my sense of where the burden of proof lies is correct, we will be entitled for the time being to continue to affirm the Equivalence Claim.

One obvious difference between the two cases is that whereas in Job Market hiring decisions are not made algorithmically, in court cases relying on COMPAS the verdicts are partly so made.[30] It could be argued, then, that what is crucial is whether a decision is made algorithmically, or at least in an algorithmically assisted way, thereby introducing the risk of machine bias.

I do not think this suggestion works. Let us distinguish between algorithms in a narrow and in a broad sense. In a narrow sense, an algorithm involves a precise mathematical formula that is applied – either by software or in manual calculations – to a certain dataset. In a broad sense, an algorithm is a process or procedure that "extracts patterns from data" (Lee and Floridi 2021, 170). If in the present context "algorithm" is intended in the broad sense, both cases – COMPAS and Job Market – involve algorithmic decisions and thus there is no difference of the proposed kind between the two cases. Members of the hiring committee in Job Market are not self-consciously applying a mathematical formula to process the information they receive. However, they do apply a procedure involving the extraction of "patterns from data", and it may even be that unselfconsciously their brains are operating along the lines of articulable mathematical formulae.

In the narrow sense of "algorithm", things are different: the hiring case does not involve an algorithmic decision in this sense. However, the problem with appealing to this narrow notion is that, with it in place, it is unclear why it should make any difference, from the point of fairness, whether one makes an algorithmic decision or not. Suppose there are two different openings. The first is filled by the hiring committee. The second is filled using a computer running a particular algorithm to determine which applicants are qualified and which are not. Suppose the same applicants apply for the two positions, and that, for every applicant, the hiring committee and the algorithm reaches the same

considerations have greater weight when it comes to positional goods than when it comes to non-positional goods. Hence, if calibration is a fairness concern, one would expect calibration to be even more important in the hiring case and this means that there is a particularly heavy burden of proof on those who think we should only be concerned with calibration in the sentencing case.

[30] "Partly" because judges are free to disregard COMPAS's predictions.

verdict. It seems incredible to suppose that some applicants can complain about unfair bias in one of these cases, but not in the other. Where fairness is concerned, the machine bias is surely no worse than the hiring committee's "non-machine" bias.

A second suggestion is that punishment involves harming whereas hiring involves benefiting, and that this difference somehow explains why concern about fairness has rather different implications in the two cases. The simple response to this is to note that the good in the penal context could be described as the benefit of avoiding long incarceration, in which case the two cases would no longer differ in the respect appealed to. But even granting the harm/benefit asymmetry, I fail to see how it would justify different fairness-based concerns about calibration. Suppose the job in question turns out to be a bad job. The successful applicant would have been better off with a different job. (Or suppose the punishment turns out to be better than the alternative.) To my mind, these suppositions would not oblige us to revise what we think matters, from the point of view of fairness, in each of the two cases.[31]

Third, it might be suggested that the two cases differ because in Job Market people are (primarily, at least) assessed on the basis of individualized evidence freely offered by the applicant, whereas in the COMPAS scenario the merits of different offenders are assessed using non-individualized evidence that is not freely offered by the offender and was obtained from criminal registers available to the court, etc. Setting aside the question whether this allegedly factual difference is as stark as this suggestion would require in order to go through, I think the difference fails to do the necessary explanatory work.[32] Suppose, in a job-market, that applicants simply indicate an interest in their preferred position, and that the employer then assesses their qualifications by collecting information about the applicants using statistical data on various reference groups to which the applicants belong. Similarly, suppose that offenders can decide, voluntarily, to have their risk of recidivism assessed by

[31] It might be suggested instead that the two cases differ morally because the harm of unjustifiably long incarceration imposed by an uncalibrated risk prediction instrument are morally wrong, while the harm of not being hired imposed by an uncalibrated hiring procedure are not. However, whether correct or not this suggestion is unhelpful in the present context, which in effect involves searching for an, and not just begging the answer to the question of what makes harms in the COMPAS case morally wrong (because unfair) and does not make harms in the ordinary hiring case morally wrong (because unfair).

[32] Depending on what, exactly, is meant by individualized evidence, COMPAS does in part use individualized information, e.g., information about prior convictions. In part, it also uses information offered – though perhaps not freely so – by offenders.

COMPAS (rather than a psychiatrist), and that the algorithm is adjusted in such a way that it is fed only individualized information. My conjecture is that, again, this would not result in our caring about lack of calibration in Job Market, or our ceasing to care about calibration in COMPAS.

A final suggestion: the key difference between the COMPAS scenario and Job Market is that in the former it is the state that makes the decisions (through the courts), whereas in the latter the decisions are made by a private employer. This could be held to be significant for various reasons. Thus, it might be said that it makes a difference because, arguably, the state cannot say "Nothing to do with me" in response to the different recidivism base rates across black and white offenders. Arguably, this difference reflects, in part at least, unjust political policies. By contrast, a private company will often be able to disclaim responsibility for the fact that fewer women have the necessary job experience than men. Again, I do not think these differences are significant in the way that is being imagined. Suppose all black offenders in the US are recent immigrants whose criminal dispositions are the result of injustices in their country of origin. My guess is that people who care about calibration would still care about it across white and black offenders in this scenario. Also, in the hiring context it makes no difference whether the employer is the state or a private employer.

At this point I shall move on. I have not demonstrated that the Equivalence Claim is true, but I have, I hope, shown that we have good reason to be skeptical about several (and in my view, the most obvious) suggestions as to why it is best regarded as false.

6. Rejecting the Northpointe View

Perhaps in the light of the above we should reject the Northpointe View – and this is indeed what I propose to do now. I shall propose a somewhat roundabout argument for this option that starts from Long's no preference argument against equal false positive/negative rates being necessary for algorithmic fairness:

> (4) *No preference:* When there is group-wise inequality of false positive rate, a higher false positive rate does not give members of a group reason to prefer that they had belonged to a group with a lower false positive rate.

> (5) *No preference, no complaint:* If inequality of some metric Y does not give members of some group a reason to prefer that they belonged to another group, then members of this group do not have a procedural fairness complaint grounded in the inequality of metric Y.
>
> (6) *No complaint, no unfairness:* If no member of a group has a procedural fairness complaint grounded in the inequality of metric Y, then group-wise inequality of metric Y is not sufficient for procedural unfairness towards members of this group.
>
> (7) *Conclusion*: Group-wise inequality of false positive rate is not sufficient for group-wise procedural unfairness.

I think this argument is forceful. For argument's sake, let us grant (5) and (6) and focus on (4). In defense of this premise, Long offers an analysis of the following complaint from a black offender whose conviction was based in part on input from COMPAS:

> I am a black defendant who was not rearrested, but I was detained. False positive rate inequality shows that I was unfairly more at risk of this false classification than a non-rearrested white defendant. After all, a greater share of non-rearrested blacks are false positives. (Long 2020, 13)

According to Long, this complaint involves a fallacy. The complaint goes subtly wrong because it incorrectly links "'risk of error' to the false positive rate. While miscalibration or inappropriately differential thresholds *are* evidence of systematically unequal risk of error, false positive rate inequality is not" (Long 2020, 13). To see this, suppose that the black defendant in a COMPAS setting is white instead, and that all other things are equal.[33] Here

[33] Why is this the relevant counterfactual to consider, one might ask? This question is particularly relevant because, in the US context, race is causally tied to many of the other properties that are used as data input in COMPAS. In the closest possible world in which the black defendant is white, plausibly, the defendant would also have been better educated, lived in an area with lower crime-rates, had a better job situation, and so on. So why is the question to ask (for purposes of assessing premise (1)) not: Would the black offender have received a high-risk score if all those things, and not just the offender's race, had been different? I take it that at this point Long could plausibly respond that the no preference argument pertains to procedural fairness complaints – see (5) – and not, say, some broader notion of social justice. For the former and narrow purpose, i.e., Long's own purpose, the indicated narrow counterfactual is relevant (both in the case of COMPAS and audit studies). That, of course, is not to deny that, in a broader social justice assessment, other counterfactuals may (also) be relevant. ("Also" because on many views social justice in a broad sense would include procedural fairness.) I thank Jenny Magnusson for pressing me on this issue.

COMPAS would have generated the same prediction, and accordingly the defendant would have faced the very same risk of ending up being a false positive, since the same information would have been feed into the algorithm. Hence, *No preference* applies in this case.

Suppose we accept this argument. It seems we can then construct a similar argument against calibration. Consider the following complaint – one mirroring that of Long's black defendant in the COMPAS setting – from an unqualified male applicant over the female-friendly calibration of the hiring procedure in Job Market:

> I am an unqualified man, who was not deemed qualified. Unequal calibration shows that I was unfairly denied a greater chance of this false classification than a non-qualified female applicant. After all, a greater share of women deemed qualified are false positives.

This complaint against lack of calibration involves a misunderstanding analogous to the one involved in Long's black defendant's complaint. Suppose the unqualified man had instead been an unqualified woman. By stipulation, this person's prospect of being falsely deemed qualified would be the same as it is in the actual scenario where he is a man: 10%. Given this, we can replace (4) in Long's argument with a similar premise regarding calibration (4*) and tweak Long's argument so that it targets the view that lack of calibration is sufficient for unfairness:

> (4*) *No preference:* When there is base rate-based lack of calibration, the lack of calibration does not give (unqualified) members of a group reason to prefer that they had belonged to a group where the (expected) percentage of individuals assigned this score ("qualified") who are qualified is lower.
>
> (5) *No preference, no complaint:* If inequality of some metric Y does not give members of some group a reason to prefer that they belonged to another group, then members of this group do not have a procedural fairness complaint grounded in the inequality of metric Y.
>
> (6) *No complaint, no unfairness:* If no member of a group has a procedural fairness complaint grounded in the inequality of metric Y, then group-wise inequality of metric Y is not sufficient for procedural unfairness towards members of this group.
>
> (7*) *Conclusion*: When there is base rate-based lack of calibration, lack of calibration is not sufficient for group-wise procedural unfairness.

In the light of this, and given the strengths of the arguments I presented above in support of the two other horns of the trilemma, a possible lesson to draw is that we should replace the third horn in the trilemma – that is, (3) the Northpointe View – with:

> (3*) Lack of calibration amounts to unfair (direct) discrimination in a sentencing context unless it reflects differential base rate (the *Northpointe* View*).[34]

(1), (2), and (3*) do not form an inconsistent triad, and all three claims seem to be compatible with the arguments I have presented. Specifically, the assertion of (1), (2), and (3*) is compatible with the way in which (I have argued) Long's argument against the idea that unequal false positive rates are sufficient for unfair bias generalizes to calibration. Neither equal false positives, nor calibration, is necessary for fairness. Perhaps, on reflection, this is unsurprising on the assumption that fairness is about the chances facing each individual of harms and benefits and given that algorithmic parity requirements such as equal false positive rates and calibration are about group probabilities. Note, finally, that (1) and (3*) are also consistent with the notion that lack of calibration and differential positive rates are indicators of unfair bias. In a version of Job Market where, on average, male and female applicants are equally qualified, lack of calibration might strongly suggest a gender-biased assessment of the applicants' qualifications. Similarly, in a US court the setting

[34] If, alternatively, we insist that COMPAS and Job Market are different, we can replace the first horn of the trilemma with (1*) "Lack of calibration does not amount to unfair bias in a job market when it reflects differential base rates resulting from injustices against the group favored by calibration", and the third horn with (3**) "Differential false positive/negative ratios amount to unfair (direct) discrimination in sentencing unless they reflect differential base rates across the two groups resulting from injustices against the group favored by the differential false positive/negative ratios". The rationale for the latter view would be that COMPAS and Job Market are different, since in COMPAS the differential false positive/negative ratios favor a privileged group, whereas in Job Market the lack of calibration favors a group subjected to unfair treatment. One take on this is that in the former case calibration compounds injustice against women, whereas in the latter calibration compounds injustice against blacks – that is why the two cases differ. For reasons I have no space to explain here, I am skeptical about the idea that there is a non-derivative reason not to compound injustice, so I mention this possibility simply to flag it, not to signal my acceptance of it. I have, however, suggested an alternative way of capturing the intuition pertaining to compounding injustice that may be relevant here (Lippert-Rasmussen 2022). Note, finally, that if the form of fairness that we are concerned with here is procedural, it is less clear what the relevance of compounding injustice is, since procedural fairness can, on some occasions, stand in the way of social justice.

of white-offender friendly lack of calibration – something COMPAS avoids – might well, in part at least, be an indicator of a racially biased legal procedure.

7. Conclusion

In this article, I have shown why the Northpointe View of COMPAS introduces a way of thinking about unfair bias that diverges from the way we think about unfair bias in the job market, especially in the context of audit studies. This way of thinking, I have argued, lands us in a trilemma to which we should respond by rejecting the view that calibration is necessary for algorithmic unfairness. My arguments suggest that post hoc interventions to prevent bias in relation to false positives and false negatives might be commendable, fairness-wise, even if they clash with calibration across groups.

I should conclude by noting that this article does not argue that such post hoc interventions are justified. I am not arguing, for example, that judges in the US context should update risk assessments of white and black offenders in a way that generates miscalibration but equivalent false positive rates. Avoiding unfair bias – assuming for the moment that unequal false positive/negative rates manifest unfair machine bias – is one concern. But there are others, such as the concern to prevent crime and concern for political legitimacy, and nothing in this article has shown that post hoc interventions to eliminate differential false positive and false negative rates in the legal context are justified all things considered. However, given our views on job market discrimination, and given also the difficulty of explaining why the job-market and punishment contexts should be assessed differently, it is difficult to see how such interventions could fail to serve fairness well – in principle, at least.

Acknowledgements

Previous versions of this paper were presented at the European Workshop on Algorithmic Fairness (EWAF), June 9, 2022, University of Zürich, a workshop on statistical discrimination at University of Mainz, August 18, University of Mainz, online at the Philosophy Department at Wuhan University, September 26, 2022, the Post Hoc Interventions conference at Lund University, October 6, 2022; and at the Nordic Network for Political Theory workshop, November 3, 2022, UiT-The Arctic University of Norway. I thank the audiences – especially my assigned commentator on the Lund event, Jenny Magnusson,

and Mattias Gunnemyr, Martin Jönsson, and Frej Klem Thomsen for helpful written comments and Lund University's Pufendorf Theme on post hoc interventions and the Danish National Research Foundation (DNRF144) for funding in relation to this article.

References

Angwin, Julia, Jeff Larson, Surya Mattu and Laure Kirchner (2016). Machine Bias. *ProPublica* May 26. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Banerjee, Abhijit & Marianne Bertrand, Saugato Datta, Sendhil Mullainathan (2009). Labor Market Discrimination in Delhi. *Journal of Comparative Politics, 37.1*, 14-27.

Beeghly, Erin & Alex Madva, (2020). *An Introduction to Implicit Bias: Knowledge, Justice, and the Social Mind*. New York: Routledge.

Brownstein, Michael (2019). Implicit bias. Stanford Encyclopedia of Philosophy: https://plato.stanford.edu/entries/implicit-bias/.

Brownstein, Michael & Jennifer Saul (eds.) (2016). *Implicit Bias & Philosophy vol. 1&2.* Oxford: Oxford University Press.

Eva, Benjamin (2022). Algorithmic Fairness and Base Rate Tracking. *Philosophy & Public Affairs, 50*(2), 239-266.

Gaddis, S. Michael (2015). Discrimination in the Credential Society. *Social Forces, 93.4*, 1451-1479.

Hedden, Brian (2021). On Statistical Criteria of Algorithmic Fairness. *Philosophy & Public Affairs, 49*(2), 209-231.

Hellman, Deborah (2020). Measuring Algorithmic Fairness. *Virginia Law Review, 106*(4), 811-866.

Hu, Lily (forthcoming). Interventionism in Theory and in Practice in the Social World. (On file with author).

Husfeldt, Thore (2023). Six Ways of Fairness. *This volume.*

Jönsson, Martin (2022). On the Prerequisites for Improving Prejudiced Ranking(s) with Individual and Post Hoc Interventions. *Erkenntnis*.

Jönsson, Martin, & Bergman, Jakob (2022). Improving Misrepresentations Amid Unwavering Misrepresenters. *Synthese*, *200*.

Jönsson, Martin and Sjödahl, Julia (2017). Increasing the veracity of implicitly biased rankings, *Episteme 14*(4), 499–517.

Lippert-Rasmussen, Kasper (2022). Is there a Duty not to Compound Injustice?. *Law and Philosophy*, online first: https://link.springer.com/article/10.1007/s10982-022-09460-y.

Lippert-Rasmussen, Kasper (2020). *Making Sense of Affirmative Action*. Oxford: Oxford University Press.

Long, Robert (2020). Fairness in Machine Learning. https://arxiv.org/pdf/2007.02890.pdf.

Neumark, David (1996). Sex Discrimination in Restaurant Hiring. *Quarterly Journal of Economics, 111.3*, 915-941.

Pager, Devah and Quillian, Lincoln (2005). Walking the Talk?. *American Sociological Review, 70.3*, 355-380.

Widner, Daniel and Chicoine, Stephen (2011). "It's All in the Name", *Sociological Forum, 26.4,* 806-822.

9

Six Ways of Fairness

Thore Husfeldt[1]

Abstract. Fairness interventions at any stage of a decision process, including post hoc, necessarily reify a moral intuition about which out- comes are viewed as "fair." Different moral intuitions formally contradict each other, and many suggestions for algorithmic, automated, or trans- parent fairness interventions are necessarily formal. I give a very simple, but complete, overview of such formal fairness notions and observe and solidify some basic contradictions between widely-held intuitions. The presentation aims to be interesting, accessible, minimal, precise, and dispassionate.

1. Introduction

This presentation aims to be an introduction to formalisations of fairness notions that is interesting, accessible, minimal, precise, and dispassionate. In particular:

Interesting. I want to explain some of the core insights, in particular about trade-offs and conflicts between widely-held fairness intuitions.

Accessible. I try to not rely on prior exposure to the technical parts, including machine learning terminology, probability theory, and causality. As best as I can, I either avoid such concepts or define them from first principles.

[1] Professor of Computer Science, Department of Computer Science, Lund University, Sweden, and IT University of Copenhagen, Denmark.

Minimal. I want to introduce as *few* concepts as possible, while still being able to present the phenomena that I find interesting, puzzling, or appealing.

Precise. All concepts are meticulously defined and arguments presented carefully and in their entirety. To the extent that it makes sense, concepts and arguments are supported by diagrams. I've spent some time worrying about appealing notation, favouring an imagined reader that is new to this area and holds no established preferences. There are no incomplete proofs, either of the form 'it can be seen by standard arguments that' or 'the proof follows from chapter 7 in Feller (1950).'

Dispassionate. I aim to be agnostic about political ideals and assume that you and I share no ideological intuitions. Rich examples are important scaffolds for navigating abstraction, but I try to stick to a running *toy* example that aims to avoid triggering our tribal instincts.

This text is not, not does it want to be

Novel. Nothing here is new, except maybe the framework and some work on identifying necessary conditions in our definitions for various relationships to hold. This entire text is an attempt to explain existing notions and findings to myself in a way that I would have liked them explained to me.

Comprehensive. I know many more fairness notions than six, but the whole idea of this text is to be minimal rather than comprehensive. Much more complete presentations can be found in Verma and Rubin (2018) and Barocas et al. (2019) and the references therein.

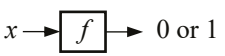
Reflective. Precise definitions, rigorous analysis, and contextual decoupling are *in themselves* epistemologically nontrivial choices, as is my focus on the trade- offs and tensions between various socially adaptive ideas. A lot can be said about this, and I don't. An accessible introduction to this discussion can be found in Friedler et al. (2021).

I also meticulously avoid *resolving* the dilemmas that result from observing the contradictions between various fairness notions. For an example of how such dilemmas may be approached, see Lippert-Rasmussen (2023).

2. Setup

2.1 Selection

Think of S as any decision-making procedure, such as an algorithm, a method, or a law. It takes an individual x and produces an outcome, either 0 or 1:



When $S(x) = 1$ we will say that "x has been selected." You can think of selection as "gets their loan application approved," "goes to jail," "is admitted to university," "gets the job," "is shown the ad," "receives the medical treatment," etc. Note that being selected can be beneficial or detrimental for x, depending on the context; mnemonically, I suggest thinking of x being selected for a *scholarship* or a *security check* when $S(x) = 1$.

Formally, S is a mapping

$$S : x \to \{0, 1\}$$

but we will often just write $S = 1$ instead of $S(x) = 1$. We draw the population classified as $S = 1$ using a thick black outline. This may encompass some, or even all of the population.

2.2 Target

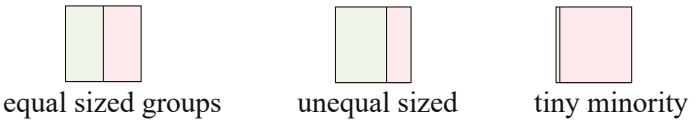
The *target* is the quality we try to select for. Think of $T = 1$ as "repays their loan," "commits another crime," "is highly intelligent," "is a pleasant and competent colleague," "will buy the advertised product," "benefits from the treatment," etc. The target value can be a desirable or undesirable quality of x; mnemonically, think of $T = 1$ as talent or *terrorist*. Both are compatible with the corresponding mnemonic for selection: We may want to select talents for the scholarship, and to select terrorists for the security check. In some contexts, you can think of T as *truth*.

In pictures, the population where $T = 1$ (the *target population*) is drawn in a more opaque colour. The mnemonic is *tinted*. We can draw examples where the target population makes up various fractions of the total population:

2.3 Groups

The population is partitioned into two groups 0 and 1. If x belongs to group 1 then $G = 1$, otherwise $G = 0$. This grouping may be by sex, gender, religion, ethnicity, caste, age, etc.[2] Think of $G = 1$ as *Greeks* in a (fictional) ancient population consisting entirely of Greeks and Romans. If the two subpopulations for which $G = 1$ or $G = 0$ have roughly equal size, you may want to think of G as gender.

We will draw group 1 in green, and the other group using not-green (in fact, red).[3] If you like the Graeco–Roman example, Greeks are green and Romans are red.



Depending on context, the group membership of x may be called "sensitive" or "protected," leading to implicit or explicit legal, social, or cultural ambitions for the interplay between S, T and G.

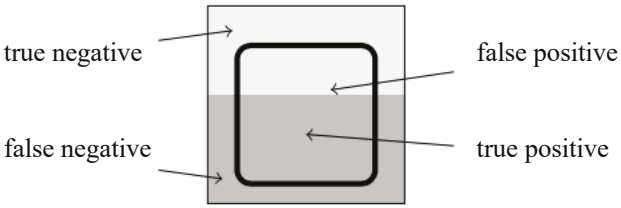
[2] Class is another plausible grouping. We avoid the word "class" here so as to avoid confusion with the classification provided S, which is often called a classifier.

[3] If you cannot distinguish the two colours, green will be on the left (*gauche* in French), and red on the right.

2.4 What we want to study

The three values S, T, and G provide a framework for studying the result of f, and we can draw them using schematic representations like this:

Example 1 (*S and T: accuracy, correctness, utility*). The interplay between the selection S and target T models notions like accuracy or correctness. We can express some standard terminology: The true positives are targeted individuals ($T = 1$) that are (correctly) selected ($S = 1$). Similarly, true negatives have $T = 0$ and are (correctly) de-selected ($S = 0$). False positives have $T = 0$ yet are selected ($S = 1$), and the false negatives have $T = 1$ and are de-selected ($S = 0$). Graphically:

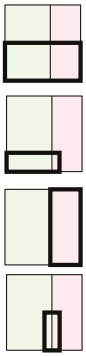


The closer S and T are, the more accurate or correct is the classifier. When $S = T$ for all inputs then the classifier is perfect, in the sense of making no mistakes:

perfect classification, $S = T$.

We say that a classifier has high utility if $S = T$. Note that "perfection," "utility," "accuracy," and "correctness" are value-laden words. Note also that it is not clear for whom a perfect classifier has high utility; the outcomes for the individual, the group, or society are often at variance with each other. For instance, failing to select terrorist x for security screening is the desired outcome for x, but catastrophic for others.

Example 2 *(S and G: representation and bias*). The classification given by *S* can relate to the grouping given by *G* in various ways. If one group is selected with more than their fraction of the population, that group is called overrepresented among the selected individuals, which is sometimes called bias.



half of population, no bias, more G

small selection, biased towards G

perfectly partisan selection, S = 1 − G

equal numbers from each group, G underrepresented

Our usual caveat applies: bias is a value-laden word that has negative connotations.

Example 3 (*G and T: diversity*). Finally, group membership *G* may relate with the target value *T*. The target quality may be rare or ubiquitous, and it may be equally or unequally represented in the two groups. Whether targeted individuals belong to either group depends on the relative group sizes.

target equally rare in both groups

target rarer among G

the majority of targeted individuals belong to G.

In the last example, note that all individuals in the red group are in the target population. Be aware that in many contexts the mere idea that target values are not equally distributed among groups is outrageous.

Example 4 *(Homerian Poetry School).* I will stick with the Graeco–Roman setting as a sufficiently silly and culturally remote toy example. The selection is a scholarship to the Athenian School for Homerian Poetry and the targeted value is talent (for Homerian poetry). The Greek population is tiny compared to the vast Roman empire; yet talent for Homerian poetry is much more widespread among the Greeks. (This may have entirely cultural reasons; Homerian poetry is written in Greek, not Latin, and highly valued in the Greek elite.)

If you're a formalist and happily navigate *S, T,* and *G* as mathematical abstractions, you can ignore my attempts at building intuition.

3. Fairness as Independence

We were able to express a few things using equality, such as $S = T$, but for the fairness notions we need *independence* from probability theory.

This will allow us to write expressions like "$S \perp G$" for "S is independent of G". The intended meaning is that S, which determines whether x is selected, is "independent of" (in the sense of "is not affected by" or "is indifferent to" or "contains no information about") G, the group that x belongs to.

If you want, you can largely ignore the fact that $\perp$ is a shorthand for a very rigorous and simple definition of "is independent of" and skip the next subsection. You can do the same if you do not need or want to be reminded of basic probability theory.

3.1 Event, Condition, Independence, Random Variable

For our purposes, an *event* E is a subset of the set Ω, with an associated *probability* $\Pr(E)$ satisfying $0 \leq \Pr(E) \leq 1$, $\Pr(\Omega) = 1$, and $\Pr(E \cup F) = \Pr(E) + \Pr(F)$ when E and F are disjoint.

Example 5. If you want, you can view Ω as 'the population,' so that events are subsets of the population, such as 'incompetent Romans falsely given a scholarship.' Then the population is finite and you can understand the probability function as $\Pr(E) = |E|/|\Omega|$.

The *conditional probability of E given F* written $E \perp F$, is the probability that E occurred given that F has occurred, and defined as

$$\Pr(E \mid F) = \frac{\Pr(E \cap F)}{\Pr(F)} \qquad \text{if } \Pr(F) > 0$$

(If $\Pr(F) = 0$ then $\Pr(E \mid F)$ is not defined.) Note that

$$\Pr(E \cap F) = \Pr(E \mid F) \Pr(F)$$

always holds, even if $\Pr(F) = 0$ (in which case both sides are 0).

Intuitively, an event E is independent of another event F, written $E \perp F$, if the fact that E happened includes no information about whether F happened. Formally, two events are *independent* if $\Pr(E \cap F) = \Pr(E) \Pr(F)$.

Proposition 1. The following are equivalent to $E \perp F$:

1. $F \perp E$.
2. $E \perp \overline{F}$.
3. $\Pr(E \mid F) = \Pr(E)$ if $0 < \Pr(F)$.
4. $\Pr(E \mid F) = \Pr(E \mid \overline{F})$ if $0 < \Pr(F) < 1$.

Proof. Set intersection and multiplication are both symmetric. For 3, we have

$$\Pr(E \mid F) = \frac{\Pr(E \cap F)}{\Pr(F)} = \frac{\Pr(E) \Pr(F)}{\Pr(F)} = \Pr(E)$$

For 2, observe

$$\Pr(E) \Pr(\overline{F}) = \Pr(E) \Pr(1 - \Pr(F)) = \Pr(E) - \Pr(E) \Pr(F) = \Pr(E) - \Pr(E \cap F) = \Pr(E \cap \overline{F}).$$ □

Whereas equality and independence are symmetric concepts, it is not true in general that $\Pr(E \mid F)$ is the same as $\Pr(F \mid E)$ (even if $E \perp F$).

Example 6. Alice and Bob each have their own (biased) coin. A is the event that Alice's coin comes up "heads," with probability $\Pr(A) = \frac{1}{10}$, Bob's with $\Pr(B) = \frac{1}{4}$. By tedious enumeration of the 400 different outcomes, we see that $\Pr(A \cap B) = \frac{10}{400} = \frac{1}{40}$. Then $A \perp B$.

Example 7. Claire has two coins. Coin 1 comes up "heads" with probability $\frac{1}{10}$, coin 2 with probability $\frac{1}{4}$. Claire picks one of her coins uniformly at random and lets both Alice and Bob toss it. The probability that Alice comes up "heads", is

$$\Pr(A) = \frac{1}{10}\ \frac{1}{2} + \frac{1}{4}\ \frac{1}{2} = \frac{7}{40}$$

Bob tosses the same coin, so $\Pr(B) = \frac{7}{40}$. However,

$$\Pr(A \cap B) = (\tfrac{1}{10})^2\ \tfrac{1}{2} + (\tfrac{1}{4})^2\ \tfrac{1}{2} = \tfrac{29}{800}$$

Thus, A and B are not independent.

To build some intuition, $\Pr(B \mid A) = \frac{29}{140}$, so having observed Alice's heads coin toss, Bob has a roughly 20% chance (much better than $\Pr(B) = \frac{7}{40}$) of a heads outcome. Intuitively, this is because it is quite likely that Alice received the 2nd coin from Claire.

Proposition 2 (Total probability)**.** Let $E_1, \ldots, E_n$ be a disjoint partition of Ω. Then for any event F, we have

$$\Pr(F) = \Pr(F \mid E_1) \Pr(E_1) + \cdots + \Pr(F \mid E_n) \Pr(E_n). \qquad (1)$$

Proof. Since the E_i for a partition of Ω and $F \subseteq \Omega$, we can write F as a disjoint union $F = (F \cap E_1) \cup \cdots \cup (F \cap E_n)$, which implies $\Pr(F) = \Pr(F \cap E_1) + \cdots + \Pr(F \cap E_n)$. By definition, $\Pr(F \cap E_i) = \Pr(F \mid E_i) \Pr(E_n)$. □

In particular, for $n = 2$, we have

$$\Pr(F) = \Pr(F \mid E) \Pr(E) + \Pr(F \mid \overline{E}) \Pr(\overline{E}),$$

which is the only version we need.

A *random indicator variable* A is a function $A : \Omega \to \{0, 1\}$. For a value a we write the (formally meaningless and abusive, but intuitively useful) expression "$A = a$" to denote the event $\{ x \in \Omega \mid A(x) = a \}$. Two random variables A and B are *independent* if for all *a, b*

$$\Pr(A = a \cap B = b) = \Pr(A = a) \cdot \Pr(B = b)$$

We extend the notation from events to random variables and write $A \perp B$. Note that if $A \perp B$ and $\Pr(B = b) \neq 0$ then

$$\Pr(A = a \mid B = b) = \frac{\Pr(A = a \cap B = b)}{\Pr(B = b)} = \Pr(A = a)$$

Two random variables are equal, $A = B$, if $A(x) = B(x)$ for all $x \in \Omega$. We write $A \not\perp B$ and $A \neq B$ do indicate that $A \perp B$ and $A = B$ fail to hold, respectively.

> **Proposition 3.** Assume $0 < \Pr(A = a) < 1$ or $0 < \Pr(B = b) < 1$ *for some a or b*. If $A = B$ then $A \not\perp B$. If $A \perp B$ then $A \neq B$.
>
> *Proof.* Assume $\Pr(B) < 1$; the other case is symmetric. If $A = B$ then $\Pr(A = a \cap B = a) = \Pr(A = a \cap A = a) = \Pr(A = a) \neq \Pr(A = a) \Pr(B = b)$, so A and B are not independent. Now assume $A \perp B$. If also $A = B$ then in particular $(\Pr(B = b))^2 = \Pr(A = b \cap B = b) = \Pr(B = b \cap B = b) = \Pr(B)$. But this can only hold if $\Pr(B = b) \in \{0, 1\}$, violating the assumption. We conclude $A \neq B$.

To avoid a misunderstanding: $A \neq B$ does not imply $A \perp B$, and $A \not\perp B$ does not imply $A \neq B$. Independence and equality are both very restrictive notions, and the relationship between A and B can fail to satisfy either.

> **Example 8.** Alice flips a fair coin, and Bob flips the same coin if it comes up 1 (else he accepts Alice 's outcome as his own). Then $\Pr(A = 1) = \frac{1}{2}$ and $\Pr(B = 1) = \Pr(A = 1) + \Pr(A = 0)\frac{1}{2} = \frac{3}{4}$. Clearly, $A \neq B$. (In fact, $\Pr(A = 1) \neq \Pr(B = 1)$.) Also, A and B are clearly not independent, and we can verify $\Pr(A = 1 \cap B = 0) = 0 \neq \frac{1}{2} \cdot \frac{1}{4} = \Pr(A = 1) \cdot \Pr(B = 0)$.

To avoid other misunderstandings: $A \perp B$ does not imply $\Pr(A = 1) \neq \Pr(B = 1)$. (Consider two independent coin flips.) $A \neq B$ does not imply $\Pr(A = 1) \neq \Pr(B = 1)$. (Let A be a random coin flip and define $B = 1 - A$.)

3.2 Demographic Parity

The relationship

$$S \perp G \tag{2}$$

means that selection is independent of group membership. In particular, group membership does not affect the classifier's selection outcome.

This is a very well-studied fairness notion and goes by many names: demographic parity, group fairness, statistical parity, equal outcomes, absence of disparate impact, Darlington's 4th criterion, equity, or just independence.

> **Example 9** (*Proportional representation*). Under (2), we have $\Pr(G = g \mid S = s) = \Pr(G = g)$. For instance, for $g = s = 1$, this means that $\Pr(G = 1 \mid S = 1)$ (the proportion of Greeks among those selected for a scholarship) equals $\Pr(G = 1)$ (the proportion of Greeks in the entire population). In words, Greeks (as well as non-Greeks) are represented among the selected (as well as among the de-selected) in proportion to their population size.

Perhaps misleadingly, the notion is often understood as the selected group *representing* the whole population. (This is misleading because selected in individuals may have very little else in common with their group.)

3.3 Target indifference

The relationship

$$S \perp T$$

means that the classifier selects individuals independently of their target value. Thus, $S \perp T$ means that selection is *indifferent* to the target value. In contrast, $S = T$ means that exactly the target value is selected for. The latter is sometimes called maximal *utility* and is often a desirable property of selection.

The choice between $S \perp T$ and $S = T$ reflects the importance of accurate selection.

> **Example 10** (*Sortition*)**.** This corresponds to flipping a coin (or some other random process) for each individual, weighted by the desired size of the selected set.

Many real-world societies have used or still use a random process for civic obligations such as jury duty in the United States. The very fact that the process achieves demographic parity (across all thinkable groups, not only G) is felt to outweigh the potential absence of any legal expertise or ethical schooling among jury members. In our notation, the benefits of achieving demographic parity or representativeness (in particular, $S \perp G$) are felt to outweigh the negative consequences of not selecting for competence ($S \perp T$). See Sec. 4.4.

In a raffle or lottery, individuals are selected based on a random process. For instance, if we want to select $k = |S|$ many individuals from the population $\mathscr{P}$, we can hand out numbered tickets numbered $1, \ldots, |\mathscr{P}|$ randomly and select those individuals receiving a number at most k. Lotteries can be used for entertainment (and the perceived fairness of the process is important for attracting customers that want to buy a lottery ticket), for selecting school children for an exciting activity such as a school trip, or for unpleasant activities such as latrine duty. See Stone (2009) for an introduction to random selection.

3.4 We're all equal

Consider

$$G \perp T.$$

In words, the target variable is independent of group membership. Sometimes called "equal base rates" or "the world is just." It represents a model of reality underlying many social theories, religions, and scholarly disciplines. When $G \perp T$ holds, a perfect classifier with $S = T$ satisfies demographic equality $G \perp S$. Most of the phenomena that make fairness definitions interesting simply vanish under this assumption.

3.5 Relationships

We have arrived at three different fairness notions,

$$G \perp S, \qquad G \perp T, \qquad S \perp T,$$

making up half of our "sixpack" of fairness. To set the stage for next two sections, we want to understand the interaction of these notions.

Luckily, there isn't much to understand. For instance, all three notions can hold simultaneously. Imagine for instance a situation in which we're all equal (so $G \perp T$ holds by assumption about the target distribution, such as letting T be the last digit of an individual's number of nose hairs in binary) and let S be the outcome of a fair coin flip (so $S \perp T$ and $S \perp G$). Then all three fairness notions are simultaneously satisfied.

We can also imagine $G = T$ (so the target is "membership in G") and still use a fair coin flip for S, so that $S \perp T$ and $S \perp G$. Now two notions are simultaneously satisfied and the third is maximally unsatisfied. The most attractive of these settings is where $S = T$ (perfect prediction), yet $S \perp G$ (equal outcomes) and $T \perp G$ (we're all equal.)

It is also thinkable that only one of the fairness conditions is satisfied. However, the two others cannot both be maximally unsatisfied: if both $S = T$ and $G = T$ then we cannot have $S \not\perp G$ (in fact, we do have $S = G$.) Finally, all three conditions can of course fail to hold. (In fact, in reality they presumably *do* fail to hold. The entire framework is a simplification that tries to conceptualise desirable properties.)

Even though I try to be almost comically agnostic and symmetric about the different notions, you may want to view the conditions Target Indifference $S \perp T$ and We're All Equal $G \perp T$ as *trivialising* conditions, at least on first reading, because of the following two examples.

> **Example 11** (*Sortition*). Let S be the result of a random process, such as a lottery. Then $S \perp T$ and $S \perp G$ hold. Thus, target indifference and demographic parity are very easy to achieve.

> **Example 12** (*Perfect world*)**.** Assume We're All Equal $T \perp G$. Now assume that the selection mechanism achieves perfect utility $S = T$. Then Demographic Parity $S \perp G$ holds. In other words, Demographic Parity is achieved by merely maximising the utility of the selection mechanism, so that the concepts of utility and fairness are identical. No conflicting goals arise, and the meritocratic intuition is well-aligned with the equity intuition, and the fairness perspective has added nothing new.

In other words, even though $S \perp T$ and $G \perp T$ may be very attractive fairness notions, keep in mind that our explorations become interesting mainly in settings where they fail. They are in some sense perfect, trivial, irrelevant, unrealistic, degenerate, boring, utopian, or even dystopian.

4. Fairness as Conditional Independence

Our central tool for modelling causality and fairness is the notion of conditional independence (Dawid, 1979; Pearl, 2009).

4.1 Conditional Independence

Let A, B, C be random variables. We say that A and B are *conditionally independent given C*, written

$$A \underset{C}{\perp} B \qquad \text{or} \qquad A \perp B \mid C \qquad \text{or} \qquad (A \perp B) \mid C,$$

if for all $a, b, c \in \{0, 1\}$,

$$\Pr(A = a \cap B = b \mid C = c) = \Pr(A = a \mid C = c) \cdot \Pr(B = b \mid C = c).$$

By $E \cap F \mid G$ we mean $(E \cap F) \mid G$.

Three conditional independence notions can be expressed:

$$S \underset{G}{\perp} T, \qquad G \underset{S}{\perp} T, \qquad \text{and} \qquad G \underset{T}{\perp} S.$$

Because of symmetry, these are *all* the ways in which our three variables can be conditionally independent.

4.2 Equal Odds

The relation

$$G \underset{T}{\perp} S$$

is called equal odds, equal treatment, conditional procedure accuracy equality, or separation.

It is easily understood in terms of *errors*: If you insist that no group is 'treated worse' (or better) than the other, then you are for equal odds. In particular, by 'treated equally' you mean the false positive rate should be the same for both groups, and the false negative rate should be the same for both groups. (By implication, the true positive rates and true negative rates are also the same.)

For instance, you achieve equal odds if you admit half of the targeted population in each group (say, half the talented Romans and half the talented Greeks), and one quarter of the untargeted population (say, a quarter of the untalented Romans and a quarter of the untalented Greeks.) From the perspective of a talented Roman, her odds of being selected are the same as a talented Greek (namely, $\frac{1}{2}$).

In other words, if we restrict our attention to only the individuals with the same T, we achieve demographic parity $G \perp S$. Yet another way of saying this is that any dependence between group membership G and selection S (i.e., deviation from demographic parity) is 'explained away' by the target distribution T.

4.3 Equally Good Prediction

The relationship

$$G \underset{S}{\perp} T$$

is also known as the Cleary model (absence of differential prediction), *sufficiency*, predictive rate parity, conditional use accuracy equality, or well-calibration within groups.

This notion is easier to understand by first looking at the relaxed version, for $S = 1$. The idea is that $G \perp T$ ("we're all equal"), when we restrict the population to the selected individuals. The selection may be heavily skewed towards one group or the other, and talent may be very unequally distributed in the population. The corresponding requirement for $S = 0$ is that the classifier *deselects* individuals *outside* of the target group with equal probability. If $G \perp T$ holds conditioned on both $S = 1$ and $S = 0$ then we have $G \perp T \mid S$.

> **Example 13.** The Homerian Poetry School of Athens admits students on a very harsh entrance exam. Greeks are much better at poetry (in particular in Greek!), so the cohort of freshmen is dominated by them. However, the (pitifully few) Romans who make it into the Homerian are every bit as talented as their classmates from across the Aegis. Students on campus can detect no group differences in performance. In fact, long-time teachers at the school, who seldom wander off-campus and only ever interact with the selected group, have the (false) impression that Greeks and Romans in general are equally good at Homerian poetry, and will lecture their worldlier friends at length about this.

Graduates from the Homerian are in high demand in the booming poetry economy, no matter their group membership.

This is an example of $G \perp T \mid S = 1$, sometimes called positive prediction parity or just predictive parity. The story says nothing about $S = 0$. For instance, in the story it is still possible that rejected Roman applicants on average are much better poets than rejected Greek applicants.

4.4 Stratified indifference

The cleanest example is sex-based draft lottery, used in many countries that select a random subset of the male population for military service, and none from the female. Also, the ancient Greeks used stratified sortition by implementing aleatoric democracy yet restricting it to males.

5. Relationships, Implications, and Trade-Offs

Three of the six fairness notions, stand out as being popular, intuitively appealing, politically viable, consistent with correctness, and achievable by manipulating the classifier:

$$\begin{array}{lll} G \perp S & G \perp T & S \perp T \\ G \underset{T}{\perp} S & G \underset{S}{\perp} T & S \underset{G}{\perp} T \end{array} \tag{3}$$

Mimicking our easy observations from 3.5, we will investigate the formal relationship between these notions. We saw that from the top row, it was possible to satisfy 1, 2, or 3 of the notions. The gist of this section is that this is not true of the second row.

This insight turns out to be a relatively pedestrian observation about the properties of conditionally independent random variables – it has nothing to do with which three fairness notions we picked. Thus, we will give a general and very simple treatment in terms of A, B and C, and spell out the implications to the popular case after each result.

5.1 Independence versus Conditional Independence

First we convince ourselves that the second row is indeed different from the first, in that the independence notion $A \perp B$ does not imply, nor is implied by, its conditional counterpart $A \perp B \mid C$. This is entirely pedestrian but seems to be psychologically counter-intuitive, so here are some counterexamples for $A = S$, $B = G$, $C = T$:

1. Demographic parity holds: Both groups are proportionally represented in the selection, roughly with 1/3 of their populations. Equal odds does not hold: targeted Greeks are certain to be selected, Romans aren't.

$S \perp_T G$ yet $S \not\perp G$

2. Equal odds holds: In each group, roughly half of the targeted individuals are selected, and none of the untargeted. But demographic parity fails: Romans are overrepresented in the selection – almost 1/3 of the Roman population is selected, but less than 1/6 of the Greek.

$S \not\perp_T G$ yet $S \perp G$;

3. If $T \perp G$ (we're all equal) then both can hold.

both $S \perp_T G$ and $S \perp G$;

In fact, both would hold if we set $S = T$, simultaneously achieving perfect prediction, equal odds and demographic parity.

4. If $S \perp T$ (target indifference), then both can hold as well:

both $S \perp_T G$ and $S \perp G$;

Since the two notions do not imply the other, we need assume both, requiring both demographic parity and equal odds, i.e., $G \perp S$ and $G \perp S$ mod T . This combination of two fairness notions is very close to the moral intuition of many

people and will be felt as a desirable requirement in the selection process. We already saw two simple examples above (3 and 4) for how this could be achieved (namely, assuming we're all equal $S \perp G$ or target indifference $S \perp T$.)

We now show that those two are the *only* possibilities.

Proposition 4. Let A, B, C denote random variables with $C \in \{0, 1\}$. If $A \perp B$ and $A \perp B \mid C$ then $A \perp C$ or $B \perp C$.

Proof. Write a for the event $A = a$, and similarly for b and c. By $\Pr(\bar{c})$ we mean $\Pr(\overline{C = c}) = \Pr(C = 1 - c) = 1 - \Pr(c)$.

By total probability we have

$$\begin{aligned}&\Pr(a) = \Pr(a \mid c)\ \Pr(c) + \Pr(a \mid \bar{c})\ \Pr(\bar{c}) = \\ &\Pr(a \mid c)\ \Pr(c) + \Pr(a \mid \bar{c})\ [1 - \Pr(c)] = \\ &[\Pr(a \mid c) - \Pr(a \mid \bar{c})]\ \Pr(c) + \Pr(a \mid \bar{c}) = \\ &q \cdot \Pr(c) + \Pr(a \mid \bar{c}),\end{aligned}$$

where

$$q = \Pr(a \mid c) - \Pr(a \mid \bar{c}).$$

From our assumptions, we can also write

$$\begin{aligned}&\Pr(a) = \Pr(a \mid b) = \\ &\Pr(a \mid b,c)\ \Pr(c \mid b) + \Pr(a \mid b,\bar{c})\ \Pr(\bar{c} \mid b) = \\ &\Pr(a \mid c)\ \Pr(c \mid b) + \Pr(a \mid \bar{c})\ \Pr(\bar{c} \mid b)] = \\ &\Pr(a \mid c)\ \Pr(c \mid b) + \Pr(a \mid \bar{c})[1 - \Pr(c \mid b)] = \\ &q \cdot \Pr(c \mid b) + \Pr(a \mid \bar{c}).\end{aligned}$$

(Note that $\Pr(a,b \mid \bar{c}) = \Pr(a \mid \bar{c})\ \Pr(b \mid \bar{c})$ holds because $\overline{\{C = c\}} = \{C = 1 - c\}$ is an event.) Combining these two expressions, we arrive at

$$q \cdot \Pr(c) = q \cdot \Pr(c \mid b).$$

For this to be true, either $q = 0$, i.e.,

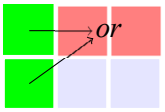
$$\Pr(a \mid c) = \Pr(a \mid \bar{c}),$$

which means $A \perp C$, or

$$\Pr(c) = \Pr(c \mid b),$$

which means $C \perp B$. □

In particular, assuming any column in (3) implies one of the other unconditional independence notions. Graphically,



and this is true for every column in (3), not only the first. Still, our most important conclusion is, in prose

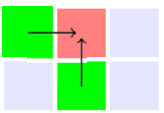
> **Trade-off 1:** if equal odds and equal outcomes both holds, then selection is target indifferent or we're all equal.

Equivalently, unless groups have equal target base rates, or selection is indifferent, the ideals of equal odds and equal outcomes are incompatible.

5.2 Triangulating

> **Proposition 5.** If $B \perp C$ and $A \perp B \mid C$ then $A \perp B$.
>
> *Proof.* Using again the shorthand a for the event $\{A = a\}$, etc., we have
>
> $$\begin{aligned}
> &\Pr(a,b) = \Pr(a,b \mid c)\Pr(c) + \Pr(a,b \mid \bar{c})\Pr(\bar{c}) = \\
> &\Pr(a \mid c)\Pr(b \mid c)\Pr(c) + \Pr(a \mid \bar{c})\Pr(b \mid \bar{c})\Pr(\bar{c}) = \\
> &\Pr(a)\Pr(b \mid c)\Pr(c) + \Pr(a)\Pr(b \mid \bar{c})\Pr(\bar{c}) = \\
> &\Pr(a)\left[\Pr(b \mid c)\Pr(c) + \Pr(b \mid \bar{c})\Pr(\bar{c})\right] = \\
> &\Pr(a)\Pr(b).
> \end{aligned}$$
>
> □

In particular, assuming an entry in each row in (3) from different columns implies the unconditional notion at the top row. Graphically,



The result is true no matter which two boxes we pick, as long as they are in different columns. But the most important conclusion, in prose, is that

Trade-off 2: if demographic parity and predictive parity both hold, then we're all equal.

Equivalently, unless we're all equal, a classifier cannot achieve both demographic and predictive parity.

Proposition 6. Assume $A \perp B \mid C$ and $A \perp C \mid B$. Then $A \perp B$ and $B \perp C$, unless the involved probabilities are zero.

Proof. To be precise, we will show that under the assumptions,

1. $A \perp B$ if for all b and c, we have $\Pr(B = b, C = c) > 0$, and
2. $A \perp C$ if for all a and c, we have $\Pr(A = a, C = c) > 0$.

We show the first statement; the other is similar. Write again a for $\{A = a\}$, etc. Using conditional probability and the assumptions, we have

$$\Pr(a,b,c) = \Pr(a\ b \mid c)\ \Pr(c) = \\ \Pr(a \mid c)\ \Pr(b \mid c)\ \Pr(c) = \Pr(a \mid c)\ \Pr(b,c).$$

and

$$\Pr(a,b,c) = \Pr(a,c \mid b)\ \Pr(b) = \\ \Pr(a \mid b)\ \Pr(c \mid b)\ \Pr(b) = \Pr(a \mid b)\ \Pr(c,b)$$

Since $\Pr(b,c) = \Pr(c,b) \neq 0$, we deduce $\Pr(a \mid c) = \Pr(a \mid b)$ for all a, b, c. We can use this (for c and $\bar{c}$) in the following derivation,

$$\Pr(a)\ \Pr(b) = [\Pr(a \mid c)\ \Pr(c) + \Pr(a \mid \bar{c})\ \Pr(\bar{c})]\ \Pr(b) = \\ [\Pr(a \mid b)\ \Pr(c)\ \Pr(b) + \Pr(a \mid b)\ \Pr(\bar{c})]\ \Pr(b) = \\ \Pr(a \mid b)\ \Pr(b)\ \Pr(c) + \Pr(a \mid b)\ \Pr(b)\ \Pr(\bar{c}) = \\ \Pr(a,b)\ [\Pr(c) + \Pr(\bar{c})] = \Pr(a,b),$$

which establishes $A \perp B$. □

In particular, assuming two of the fairness notions from the bottom row implies their unconditional counterparts. Graphically,

Most importantly:

> **Trade-off 3:** if equal odds and predictive parity both hold then we have demographic parity and we're all equal.

Equivalently, unless we're all equal and the classifier achieves demographic parity, then the classifier cannot both guarantee equal odds and predictive parity.

References

Barocas S., Hardt M., and Narayanan, A. (2019) *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org. http://www.fairmlbook.org.

Dawid, A.P. (1979) Conditional independence in statistical theory. *Journal of the Royal Statistical Society, Series B.*, 41(1):1–31.

Friedler S.A., Scheidegger C., and Venkatasubramanian S. (2021) The (im)possibility of fairness: Different value systems require different mechanisms for fair decision making. *Communications of the ACM*, 64(4):136–143.

Lippert-Rasmussen, K (2023). Post hoc interventions and machine bias. *This volume*.

Pearl, J. (2009) *Causality*. Cambridge University Press.

Stone P. (2009) Logic of random selection. *Political theory*, 37(3), 2009.

Verma S. and Rubin J. (2018) Fairness definitions explained. In *Proc. of 2018 ACM/IEEE International Workshop on Software Fairness, FairWare'18, May 29, 2018, Gothenburg, Sweden*.