A Brief Introduction to Post Hoc Interventions

Martin L. Jönsson

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

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

DOI: https://doi.org/10.37852/oblu.184.c501

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

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.