May 18th, 9:00 AM - May 21st, 5:00 PM

Commentary on "Eliminating Gender-, Racial- and Age-Biases in Medical Diagnostic Reasoning"

Steve Oswald
University of Fribourg

Commentary on "Eliminating Gender-, Racial- and Age-Biases in Medical Diagnostic Reasoning"

STEVE OSWALD
Department of English
University of Fribourg
Av. De l'Europe 20, 1700 Fribourg
Switzerland
steve.oswald@unifr.ch

1. Introduction

The notion of cognitive bias has attracted the attention of argumentation scholars in different ways over the years. From a theoretical perspective, a number of them (see e.g., Correia, 2011, 2014; Bardone, 2011; Jackson, 1996; Walton, 2010; Oswald & Hart, 2013) have discussed the role of cognitive biases in relation to the production and reception of argumentative fallacies, building on the assumption that argumentative fallacies can be taken to exploit errors and non-optimal outputs of biased cognitive processes. From an applied and pedagogical perspective, scholars interested in critical thinking teaching have reflected on the means instructors might have at their disposal in order to counter the effect of biases (see e.g., Zenker 2013; Larrick 2008). MacPherson's paper is directly relevant to this second strand of research, as it proposes a discussion of debiasing techniques.

In the context of medical diagnosis and more specifically regarding misdiagnosis errors that have been shown to originate in cognitive biases, MacPherson discusses the debiasing solution advocated by Croskerry (2003), namely cognitive forcing, reviews empirical counter-evidence to Croskerry's claims, explains why cognitive forcing might not work and argues that a better solution is to be found in adapting the shared mind approach to medical diagnosis. In this commentary I take the opportunity to further explore the theoretical ramifications of the notion of cognitive bias (section 2.1), the merits of the MacPherson's assessment of debiasing techniques (section 2.2) and the virtues of argumentation as a systematic tool to be implemented in the shared mind approach to debiasing (section 3).

2. Some critical comments

2.1 What counts as a cognitive bias?

The first point of discussion I want to raise has to do with the notion of bias MacPherson uses, and so this is rather a discussion of the work MacPherson draws on rather than a discussion of his work proper.[1]

Building on a body of literature on the role of cognitive biases in medical contexts, MacPherson identifies different types of biases that might negatively influence the outcome of a medical diagnosis: according to him (and to the literature he draws on), race, age, ethnicity and gender are typically the sort of personal features that physicians unconsciously take into account

[1] Even if the observations made here in principle percolate to MacPherson's account.

Bondy, P., & Benacquista, L. (Eds.). *Argumentation, Objectivity, and Bias: Proceedings of the 11th International Conference of the Ontario Society for the Study of Argumentation (OSSA), 18-21 May 2016*. Windsor, ON: OSSA, pp. 1-8.

as they provide a diagnosis, and this can, in some cases, result in a biased diagnosis, with the associated potential of being extremely detrimental (if not lethal) to the patient. From his review, MacPherson rightly concludes that "cognitive biases are a very real problem in medical diagnoses and that this problem therefore needs to be addressed if every patient regardless of race, ethnicity, age or gender is to have access to the best medical opinions free of *irrational* biases" [my italics]. Recalling a real example (see Croskerry et al., 2013), MacPherson reports that a physician who, based on the fact that a young female patient suffering from respiratory problems is referred to by a psychiatrist, does not think of pneumonia and attributes the symptom to anxiety, is responsible for making a biased decision: the physician is making a decision which *a priori* rules out (or does not even take into account) this condition based on the patient's recent history, age and gender—and this should not be the case (in the example given, the patient died following misdiagnosis).

It is interesting to note in passing that McPherson takes biases to be "irrational". This, I believe, warrants some further discussion. Looking at the example above, one cannot help but wonder whether the same cognitive process, in case the patient's respiratory symptoms were indeed attributable to anxiety issues, could lead to a desirable outcome such as improving the patient's condition. In other words, it is perfectly reasonable to assume that the same cognitive process could lead to an optimal result—which would in fact be compatible with the idea that biases and heuristics are not always a bad thing (see Gigerenzer et al., 1999). The question we need to ask, therefore, is whether this is (ir)relevant to identifying the process as a bias or not: since what changes from Croskerry et al.'s example to my hypothetical case is the context and not the cognitive process itself, saying that in one case we witness biased decision-making and not in the other falls short of making an informative claim on the process—since in both cases the process is the same. Moreover, in this hypothetical example, there are grounds to consider that the decision is perfectly rational: given the situation and the means at the physician's disposal (factoring in time and efficiency constraints in the ER, for instance), the decision is indeed rational because it represents a desirable course of action in view of attaining the desired goal under the circumstances. The question is, again: is it a bias in this case? If it so happens that the patient is a member of the group targeted by the bias, then the diagnosis might just as well turn out to be correct. This would compel us to take a stance on whether we are prepared to identify successful diagnosis as the result of bias.[2]

In order to start answering the question, let us first focus on the operational definition MacPherson (2016) uses, and which he borrows from Haselton et al. (2005), namely that "a cognitive bias involves a departure from accepted standards of rationality morphing into an idiosyncratic standard." As seen above, the purported (ir)rationality of biases is directly relevant to critically examining the definition. Yet, there is more at stake here. One of the major features of this definition is that it does not discriminate between types of biases. The biases examined by MacPherson include gender, race, ethnicity, age: their influence in medical diagnoses is assessed in terms of the outputs these biases generate, which are problematic. I would argue that no clear-cut reason is offered in the paper to consider that these are instances of biases—rather, MacPherson assumes that they are—and not of something simpler, namely prejudices or mere opinions.

If we take a simple mechanistic perspective on how information is processed, we may consider that a medical diagnosis involves an inferential process which takes into account some

[2] I would probably go as far as claiming that the diagnosis is informed by heuristic processing. Whether this can be assimilated to a bias is a separate question.

input information (a patient's observable or measurable psycho-physical condition), processes it against background information (medical encyclopaedic knowledge, knowledge of probabilities based on history, etc.) and yields an output (the medical diagnosis) which then informs the decision to treat the patient in a particular way. Once we adopt this perspective, it becomes apparent that the biases MacPherson discusses have to do partly with the information that is immediately accessible (is the patient a female or a male, a young or old person, etc.?) but mostly with the content of background information, which consists for instance of information about specific groups being probabilistically more associated with certain conditions than others. In other words, these biases, per se, are due to the presence of a piece of background information (e.g., "young female patients are less likely to suffer from a heart condition than others", "respiratory problems are more likely to be caused by anxiety in psychiatric patients than by pneumonia") that is available to be used in the diagnostic inference. Deep down, this boils down to being prejudiced (and I am not using the term for its evaluative import) against certain groups, in the sense that the representations associated to that group, in the medical domain, are different from the representations associated to other groups. It thus seems that the sort of bias discussed by MacPherson is of a *conceptual* nature: these biases are triggered by a particular *content* playing a role in the inference. Moreover, looking at the inferential process, it appears that the only difference between gender, age, race or ethnicity biases is in the end the content of the piece of information that is being used in the diagnosis inference—the nature of the inference does not have to be different in all cases; what differs is the nature of the information you feed it.

To further illustrate this point, let us compare a gender bias with another overwhelmingly studied bias in cognitive psychological literature, namely the confirmation bias (see e.g., Oswald & Grosjean, 2004), which is usually defined as the mind's propensity to represent, select, store and look for information that confirms (or does not contradict) information that is already present in the cognitive system. From this description, it appears that the confirmation bias is a *procedural* bias, that is, it consistently applies regardless of the specific contents of the cognitive system and of the information it is being fed. Put differently, the nature of the confirmation bias is a constraint on the way information is processed, selected, and not a given piece of information, i.e., not a representation—which, to me, seems to be what the biases discussed by MacPherson amount to. Because its operation is independent from content, the confirmation bias is therefore *procedural*, while a so-called gender bias would be *conceptual*, since its presence is synonymous with the presence of a representation about some feature typically associated to a given gender.

From this brief discussion, two options emerge: either we restrict the notion of bias to the notion of *procedural* bias or we extend it to cover *content* as well. The first option would entail considering gender, race, age and ethnicity biases as mere prejudices that may or may not be present in a cognitive system—they are contents but not functions. The second option would include representations in the category of biases, alongside procedures, putting the former on a par with procedural biases. The first option would suggest, therefore, that there are no gender, age, race or ethnicity biases, but rather that representations which are gender-, age-, race- or ethnicity-relevant may play a role in a diagnosis inference and that they may contribute to yielding a problematic biased output. In this case, I would still argue that there is a bias, but I would attribute this to the overarching influence of the confirmation bias. Upon seeing a young female patient with respiratory problems sent in by a psychiatrist, a physician would look for a diagnostic that confirms the information that is available before their eyes, and this would

explain, provided the age prejudice is present in the physician's cognitive environment, why the diagnosis might exclude pneumonia and privilege anxiety issues.[3]

One reason to prefer this analysis has to do with its minimalistic advantages: once we adopt a *procedural* definition of bias, we prevent the list of biases from being ever-expanding. If, on the contrary, we allow for *conceptual* biases, there is virtually no limit as to what can constitute a bias, as illustrated by the following *reduction ad absurdum*: as stated by Haselton et al. (2005), biases become idiosyncratic, and there is no *a priori* argument to exclude unlikely—but possible—biases against green eyed people or biases against people whose breaths last thirteen point five seconds. In such a scenario, the theoretical model would simply collapse by allowing for a virtually infinite set of *ad hoc* biases.

What this discussion shows is that if we want to account for the influence of biases in inferences, a theoretical model needs to reflect on the very nature of bias. I have argued that a procedural construal is to be privileged. However, this is not directly a problem for MacPherson's account, since (i) his discussion can be adapted to fit this distinction by postulating the overarching operation of the confirmation bias – or any other suitable alternative – and a knowledge base which contains prejudiced representations, and (ii) the main point of his contribution is to discuss debiasing techniques in the context of medical examination. I now turn to discussing his assessment of existing debiasing techniques.

2.2 On debiasing

In his discussion of ways to avoid bias in medical diagnosis, MacPherson reports on Corskerry's (2003) techniques of 'cognitive forcing', which involve critical self-reflection. His main refutation of Corskerry's proposal rests on the premise that many biases are unconscious, which makes people who suffer from them unable to critically self-reflect on these biases. In so doing, MacPherson puts forward the idea that cognitive forcing techniques are only applicable to conscious biases—and as such that they problematically leave the debiasing of unconscious biases out of the picture.

Instead, MacPherson advocates a shared mind approach to prevent biased decisions from being made. Under this approach, medical decisions are envisioned as a collaborative process involving a number of medical practitioners and patients. Shared deliberation is claimed to be advantageous in that (i) it allows for truly informed consent, (ii) it produces better outcomes than individual processes, and (iii) it shares cognitive load between participants. These three features are taken to increase the likelihood of nonbiased decisions. To be more explicit, deliberation promotes critical testing, provides a platform for argumentation, fosters multiple perspectives and, crucially, provides the opportunity to identify the bias of one participant. Regarding this last point, MacPherson's idea is that in such a case, the biased diagnosis will end up being eliminated or reduced, which in turn will filter out its effects in the result of the decision making process. This type of approach makes intuitive sense and bears some resemblance with the encouragement of cognitive diversity which some philosophers advocate in order to weaken the impact of conspiratorial beliefs within a community (see Sunstein & Vermeule, 2009); the idea would be to make sure that potentially biased claims are publically confronted, with the presence of refuting evidence, so that community members are exposed to alternative viewpoints.

While this strategy seems effective to make sure that the end decision is unbiased, I believe we should be cautious in affirming that it represents a proper debiasing technique. What

[3] The same line of reasoning would hold for what MacPherson refers to as age, race and ethnicity biases.

MacPherson refers to as an unbiased outcome is, actually, just that. i.e., an unbiased outcome, and we should refrain from assuming that it is indicative of genuine debiasing. By this I mean that the contribution of the medical practitioner who holds a biased representation (about gender, race, age, ethnicity) is certainly neutralised in the decision-making process, but that this is quite different from the situation in which the individual is properly debiased (which would correspond to the situation in which that same individual no longer holds the biased representation to be true).

Filtering out should consequently not be taken to be the same thing as debiasing, even if the outcomes of both might fully overlap. The difference between them is that the first is focused on making sure that the *result* of the inference is bias-free, which amounts to reducing the influence of bias through a collaborative process. Yet, this is not equivalent to making sure that the bias disappears: it is about preventing it from playing a role in the decision making process. Conversely, the second process, debiasing, targets the individual's cognitive processing in order to, ideally, get rid of the bias altogether. If debiasing is successful, then the inference itself would be bias-free from that point on. Taken as a "technique used by external agents to modify the decision environment" (Larrick, 2008, p. 317), a shared mind approach to medical diagnosis would be effective only under one crucial condition: the bias needs to be identified at some point of the discussion. In this respect, it must be acknowledged that this approach may fail to work for a number of reasons, among which include the following two.

First, imagine the (again) hypothetical but possible case in which all participants to the deliberative process are prejudiced in the same way. Imagine, for instance, that all of the agents share the assumption that young females are unlikely to suffer from heart disease, and moreover that none of them thinks of challenging that representation during the exchange. In such a scenario, and independently of the nature of the discussion, it is possible that the output diagnosis will be biased. This thought experiment suggests that even deliberation might fail to yield an unbiased outcome.

Second, imagine that the physician who holds a biased representation about young females not being likely to suffer from heart disease is an extremely skilled arguer, and that their co-participants are not. The deliberation might still turn in their favour, even in light of criticism, thus still allowing a biased diagnosis to emerge.

I would venture that the shared mind approach to debiasing does have merits on its own, but that it should be carefully accompanied by some auxiliary measures in order to lead to genuine debiasing. In other words, I believe that in addition to attempting to neutralise the effects of the bias, the shared mind approach should be supported by clear argumentative measures in order to achieve genuine debiasing.

3. Genuine debiasing: How (else) to reach it?

I believe that MacPherson's proposal may achieve genuine debiasing provided it is supported by careful argumentative measures (see also the discussion on countering the polarization effect in Zenker, 2013; Oswald, 2013).[4] I furthermore believe that this can be justified in light of both the cognitive and social ins and outs of argumentation, to which I now turn.

[4] I refer the reader to Oswald (2013) in particular for a more extensive account on the relationship between reasoning and debiasing. The shared mind techniques advocated by MacPherson are, in my view, likely to be successful for roughly the same reasons I identified in that paper.

In their argumentative theory of reasoning, Mercier & Sperber (2009; 2011) have shown that reasoning performs at its best in its 'natural habitat', namely in argumentative discussions. According to the theory, people provide and evaluate arguments in optimally efficient ways when they are motivated to do so, i.e., when their claims are explicitly challenged in conversation or when they feel the need to challenge the claims of others because they disagree with them. In other terms, people are good reasoners when they enter challenging argumentative exchanges. One of the main reasons for this is that an argumentative context provides both the motivation and the affordances required to induce self-reflection and multiple perspective taking. On their own, i.e., outside a conversational context, people are likely to trust their own cognitive processes, even if these are biased (see Sperber et al. 1995); however, in argumentative discussions they are likely to face contradicting claims. Accordingly, in such contexts they are likelier to need to find arguments supporting their claims and arguments contradicting their opponents' claims. This is actually the first necessary step towards debiasing: forcing one to find arguments for one's claim is exposing it and amounts to exposing its potential flaws. Due to the social pressure of the discussion, one will then have to find better arguments. If one is unable to come up with such arguments, chances are that one will recognise the weakness of one's position.[5] The advantage of argumentation over individual self-reflection is its ability to trigger a first-hand experience of being wrong but also the fact that it allows us to understand the causes of our errors. Summing up, encouraging people to reason might be the right way to go to achieve debiasing goals.

Thus, the shared mind approach defended by MacPherson has prospects of being effective as a genuine debiasing technique not on the grounds that it might filter out biased representations in the deliberative process,[6] but on the condition that the participants to the deliberation take their argumentative roles seriously. A shared mind approach might therefore be effective because (i) it provides a platform for disagreement, which is a necessary condition to identify and make multiple opinions public, (ii) it provides a platform for the production and evaluation of arguments, and, as such, (iii) it provides a social platform which is likely to activate the cognitive mechanisms involved in reasoning.

The constraints that I have previously evoked in the success of the shared mind approach are thus of an argumentative nature. In what follows I list some of them in order to paint a schematic picture of the sort of procedure that might be envisaged. In order to foster genuine debiasing, the requirement of the shared mind approach to medical diagnosis would first include making errors public as the result of critical testing, this taking the shape of an explicitly argumentative procedure. In other words, identifying the error should be the result of an argumentative discussion in which the party that is responsible for the error has been cornered into not being able to justify their claims. Only then can the debiasing process hope to really start. Such a framework should ideally specify the usual obligations and prerogatives linked to a critically reasonable argumentative practice. Participants should be aware that putting claims forward is committing, that they need to observe the obligations linked to the burden of proof, that they are required to provide arguments for their claims and that these arguments are to be evaluated according to some agreed-upon standard. In case they fail to justify a given diagnosis on its merits (and rejection of a diagnosis should also be inter-subjectively agreed upon), not

[5] We could argue that this is all the more likely in medical diagnosis deliberations, given the high stakes associated to their outcome (in a life and death situation, personal feelings such as pride might be put aside in favour of diagnosis accuracy).

[6] As seen above, this runs the risk of getting rid of the output of the bias but not of the bias itself.

only should they withdraw their claim, but they should also commit to reflecting on why they were defending the claim in the first place, which should allow them to identify the prejudiced representation—and this is something that will be made accessible by the argumentative discussion. Perhaps an additional constraint would be to explicitly instruct participants to be charitable with other participants' opinions, so as to make sure that evaluative procedures target the merits of the arguments being exchanged instead of being based on one's own position.

The main idea behind such a 'systematised shared mind approach' would be to make sure that deliberation in medical diagnoses systematically takes the form of an argumentative exchange, since it is likely that the way to debiasing goes through the triggering of reasoning processes, the latter being in full effect in argumentative contexts.

4. Conclusion

In this commentary I have first taken the opportunity granted by MacPherson's focus on cognitive biases in medical diagnoses to discuss the very notion of bias, arguing in favour of the narrower construal of *procedural bias*. In relation to this distinction, I have furthermore critically discussed the main argument given by MacPherson against Croskerry's debiasing technique of cognitive forcing, which then led me to assess his proposal of a shared mind approach to debiasing. I critically assessed his proposal, arguing that (i) it might prevent the result of bias from being influential, but not the bias itself, thereby failing to secure genuine debiasing and (ii) it might fail in some cases.

In an attempt to further develop the model presented, I suggested that a systematic enforcement of argumentative standards might enhance the model by making sure that the techniques advocated by MacPherson take place in a truly argumentative context. Therefore, I suggested that the input of argumentation theory could be extremely profitable to the implementation of debiasing techniques based on the shared mind approach. All in all, the idea is to make sure that participants to the deliberative process in medical diagnoses use their reasoning abilities; the way to do that, I believe, is to get them to argue.

A final set of caveats needs to be recognised. This discussion makes sense from a theoretical perspective, but we should not forget that the ideal speech situations depicted herein rarely occur in real life: in the ER, physicians oftentimes need to make decisions very quickly; medical staff might not always be available when needed; humans may fail, for whatever reason, to be charitable and to consider other people's opinions seriously, etc. Those are but three complications that would prevent the possibility of conducting diagnosis deliberations as outlined here, and this highlights that discussions should remain attuned to the specificities of the contexts under examination.

References

Bardone, E. (2011). *Seeking Chances: From Biased Rationality to Distributed Cognition*. Berlin/ Heidelberg: Springer.

Correia, V. (2011). Biases and fallacies: The role of motivated irrationality in fallacious reasoning. *Cogency* 3 (1), 107-126.

Croskerry, P. (2003). Cognitive forcing strategies in clinical decision making. *Annals of Emergency Medicine* 41,110-120.

Croskerry, P., Singhal, G., & Mamede, S. (2013). Cognitive debiasing 1: Origins of bias and theory of debiasing. *BMJ Quality & Safety* 22, 58-64.

Gigerenzer, G., Todd, P. M., & the ABC Research Group (1999). *Simple Heuristics that Make Us Smart*. NY; Oxford: Oxford University Press.

Haselton, M. G., Nettle, D., & Andrews, P. W. (2015). The Evolution of cognitive bias. In D. M. Buss (Ed.), *The handbook of evolutionary psychology* (pp. 724–746). Hoboken, NJ: John Wiley & Sons, Inc.

Jackson, S. (1996). Fallacies and heuristics. In J. van Benthem, F. H. van Eemeren, R. Grootendorst & F. Veltman (Eds.), *Logic and argumentation*. Amsterdam: Royal Netherlands Academy of Arts and Sciences.

Larrick, R. P. (2008) Debiasing. In D. J. Koehler, & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 316–337). Malden, MA: Blackwell Publishing Ltd.

Mercier, H., & Sperber, D. (2009). Intuitive and reflective inferences. In J. St. B. T. Evans & K. Frankish (Eds.), *Two Minds: Dual Processes and Beyond* (pp. 149-170). Oxford: Oxford University Press.

Mercier, H. & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioural and Brain Sciences* 34, 57–111.

Oswald, S. (2013). Commentary on Frank Zenker's "Know thy biases! Bringing argumentative virtues to the classroom". In D. Mohammed & M. Lewiński (Eds.), *Virtues of Argumentation: Proceedings of the 10th International Conference of the Ontario Society for the Study of Argumentation (OSSA), 22-26 May 2013,* (pp. 1-7). Windsor, ON: OSSA.

Oswald, M., & Grosjean, S. (2004). Confirmation bias. In R. Pohl (Ed.), *Cognitive Illusions. A Handbook on Fallacies and Biases in Thinking, Judgement and Memory* (pp. 79-96). Hove & New York: Psychology Press.

Oswald, S., & Hart, C. (2013). Trust based on bias: Cognitive constraints on source-related fallacies. In D. Mohammed & M. Lewiński (Eds.), *Virtues of Argumentation: Proceedings of the 10th International Conference of the Ontario Society for the Study of Argumentation (OSSA), 22-26 May 2013*, (pp. 1-13). Windsor, ON: OSSA.

Pohl, R. (Ed). (2004). Introduction: Cognitive illusions. In R. Pohl, (Ed.), *Cognitive Illusions. A Handbook on Fallacies and Biases in Thinking, Judgement and Memory* (pp.1-20). Hove & New York: Psychology Press.

Sperber, D., Cara, F., & Girotto, V. (1995). Relevance theory explains the selection task. *Cognition* 57, 31-95.

Sunstein, C. R., & Vermeule, A. (2009). Conspiracy theories: Causes and cures. *Journal of Political Philosophy* 17 (2): 202–227.

Walton, D. (2010). Why fallacies appear to be better arguments than they are. *Informal Logic* 30 (2), 159-184

Zenker, F. (2013). Know thy biases! Bringing argumentative virtues to the classroom. In D. Mohammed & M. Lewiński (Eds.), *Virtues of Argumentation: Proceedings of the 10th International Conference of the Ontario Society for the Study of Argumentation (OSSA), 22-26 May 2013,* (pp. 1-11). Windsor, ON: OSSA.