

The question of bias in philosophy and computing

Yasemin J. Erden¹

Abstract

During a discussion about racism, a colleague once asked *why can't we stop talking about race?* More than a question it was an expression of frustration; the topic is old, the subject well covered, why aren't we *post-race*, in a utopia of *one human race*? This is not an unfamiliar question nor position, and it tends to occupy the same intellectual territory as remarks like, *I don't see colour*, or *that's not my experience*.² In this paper I begin from these ideas in order to explore topics like *implicit bias*, diversity, and why some areas (like philosophy and computing) struggle more than others in tackling the problems that arise as a result. To do this, the paper will explore particular instances of bias in philosophy and in technology, and offer reasons for such tendencies and outcomes. The paper offers that while there is a relation between language and bias, the correlation needs to be understood in the context of culture and practice, or in Wittgenstein's terms, the *language-games* in which words are meaningful. Given this, the paper concludes that any attempt to solve problems of bias via management of language will have only limited success, especially if the context within which such problems occur is not given sufficient attention.

1 BIAS AT THE HEART OF REASON

In 2013, a researcher sought to compare the rate at which the work of male philosophers and female philosophers are cited. His study looked at the “most-cited items from 1993 to mid-2013” in top philosophy journals. By *top* we can understand ideas like well-known, high circulation, high impact, good reputation, and so on. These journals were *Nous*, *Mind*, *Journal of Philosophy*, and *Philosophical Review*. Of the 520 top cited items in these journals during a 20 year period, only “19 are by women—3.6 percent of the total” from “15 women authors in total”. In comparison, “6.3 percent of the items are by David Lewis” [13]

In a similar vein, another study [4] measured the number of publications by black philosophers in the US (US BIPs), across 15 journals (including some of those publications noted above) between 2003-2012 [cf. 9; 16; 25]. The author found that publications by black authors were few. In fact,

there were 30 publications by US BIPs for all journals during this period. By contrast, there were 10659 publications overall during this period. This means that publications by US BIPs were 0.28% of the publications during this period. Of the 30 publications, 15 were research articles as opposed to book reviews. There were 7638 research articles overall in this

period, meaning that research articles by US BIPs were 0.19% of the research articles published.

While it is difficult to establish causation, there does seem to be some bias at play.³ The question of how this is possible strikes at the very heart of the nature of philosophy, a discipline which is apparently an ‘intellectual inquiry in which *anything* is open to critical challenge and scrutiny. This, at least, explains many of its salient features. Philosophy is *subversive*. Time and again, philosophers have shot at religions, political systems, public mores. They do this because *they are prepared to challenge things which everybody else takes for granted*, or whose rejection most people do not countenance’ [18, p. 202, italics added]

Thus we seem to have a contradiction, where the idea of a neutral, objective, reason-based approach in philosophy, meets the stark reality of at worst intentional negative bias or prejudice, and at worst (lazy?) habit. Because while it is not evident that the work of women or black philosophers in philosophy is overlooked *intentionally*, but that such work *is* overlooked, if it even gets as far as publication, remains. Where other attributes are taken into account, such as inclusion of those with disabilities, the numbers dwindle only further. Consider for instance, that in 2012/2013 disabled people comprised only ‘4% of full-time faculty in philosophy departments in the US’ and ‘less than 1% of full-time faculty in philosophy departments in Canada.’ [23] While *causes* may be difficult to establish, the disparity remains evident.

The ramifications of this supposedly self-critical, self-aware, *subversive* discipline failing to recognise its own subjective biases, and thus objective limitations, is one that is matched by the un-self-critical attempts toward a kind of value neutrality in computing and technology. It is to the outcomes of this sort of approach (whether intentional or otherwise) that we now turn.

2 BIAS IN PRACTICE

In 2015, Google image recognition software tagged the faces of two black people as *gorillas*. When this happened, the person who'd uploaded the images took to Twitter to voice their disgust. While this example is already rich as an example of bias, I am particularly interested in the reply by a Google representative soon after the initial Tweet. This response seems to have followed along the lines that they would “stop using the label ‘gorillas’ as a category. Later, they'll figure out how to make their algorithms more sensitive to language that could be offensive.” So the issue as it was seen in that first moment was not that the algorithm had failed to characterise a human *as a human*, but that *the language was offensive*, and by extension, the categorisation that followed. To fix this, at least in the short term, the researchers would simply *change the language* [17].

¹ Philosophy, St Mary's University, London

Email: yj.erden@stmarys.ac.uk

² Another such example is neatly described in one article by a young (white) student who wishes the conversations on race would stop. She implores, “When we stop pointing out the racial differences in people and decide instead to treat all races the same, in both positives and negatives, race will no longer have such a grand impact.” [11]

³ The alternative position (that these categories of researcher do not produce work of sufficient quality, or similar) will need to be made by a different author.

What this example suggests is that the broader issue, whereby a system fundamentally fails to recognise the variety and complexity of human faces, is ignored *at least in the short term*. The focus is on preventing future categorisation from being offensive, or in other words, to keep the system working. As a result, and as with the examples in philosophy noted above, the issues that are not being tackled are those that stem from a lack of both representation (in terms of images) and recognition, but also diversity more generally. Indeed, if the error is fixed, such issues may never again come to the fore where Google's programming is concerned. At least, not until another mistake is made.⁴ This does not mean the issues of representation and diversity are addressed.

The next issue concerns gender stereotypes, examples of which are not hard to come by. For our purposes we'll consider the computer game *Grand Theft Auto*. This game doesn't only perpetuate certain stereotypes; it makes a feature of them. For instance,

female characters are treated as sex objects rather than as individuals worthy of respect. GTA main male characters are always depicted as hyper-masculine, dominant, and aggressive men. In contrast, the female characters are portrayed as sexual objects—usually prostitutes or pole-dancers—who are peripheral to the game narrative and whose sole purpose is to entertain the main male characters. For example, after paying a prostitute for sex, players can kill her and get their money back. Rather than being punished for such behaviors, players are often rewarded (e.g., through points, extra health to their character, etc.). [12]

Translation tools have also shown such biases, as seen in statistical machine translation (SMT) systems which 'reflect existing gender stereotypes':

Translations to English from many gender-neutral languages such as Finnish, Estonian, Hungarian, Persian, and Turkish lead to gender-stereotyped sentences. For example, Google Translate converts these Turkish sentences with genderless pronouns: "O bir doktor. O bir hemsire." to these English sentences: "He is a doctor. She is a nurse." A test of the 50 occupation words used in the results presented in Figure 1 shows that the pronoun is translated to "he" in the majority of cases and "she" in about a quarter of cases; tellingly, we found that the gender association of the word vectors almost perfectly predicts which pronoun will appear in the translation.' [5, p. 11]

In another sphere, but resulting none the less in bias, data capture caused headaches for Amazon recently. The problem centred on Amazon Prime's Free Same-Day Delivery service, which was shown to pointedly include some zip codes while excluding others, primarily along the lines of historical racial divides. For instance, 'In New York City, same-day delivery is available throughout Manhattan, Staten Island, and Brooklyn, but not in the Bronx and some majority-black neighbourhoods in Queens. In some cities, Amazon same-day delivery extends many miles into the surrounding suburbs but isn't available in some ZIP codes within the city limits' [14]. Though the issue was acknowledged by Amazon, and stated to be one arising from data (e.g. how many users by area), the issue of perpetuating racial divisions cannot be avoided [see also 1; 2; 3; 6; 7; 22].

⁴ Another example that suggests a lack of diversity in representation are the HP cameras that assume an East Asian person must be blinking <https://www.flickr.com/photos/jozjozjoz/3529106844/>

These examples suggest there is a case to be made for further consideration, and also to seek causes. But as we will see in the next section, attributing causes, or establishing *cause and effect* is rarely simple. And even where reasons rather than causes may be found, this may not result in solutions.

3 THE PROBLEM WITH CAUSE & EFFECT

In the effort to face these issues head on, some psychologists and philosophers have sought to establish causes or reasons for bias. One such method is the attempt to uncover *implicit bias*. As Saul [19, p. 40] notes:

Psychological research over the last decades has shown that most people—even those who explicitly and sincerely avow egalitarian views—hold what have been described as implicit biases against such groups as blacks, women, gay people, and so on. This is true even of members of the 'targeted' group... So, for example, women as well as men are biased against women. These biases are manifested in, for example, association tasks asking subjects to pair positive and negative adjectives with black or white faces: Most are much speedier to match black faces with negative adjectives than with positive ones.

To this she adds, [19, p. 41]:

Academics are clearly affected by implicit bias, even if (as seems likely) explicit commitments to egalitarianism are widespread. First, take the case of journal submissions. Anonymous review is apparently only rarely practiced in ecology and evolution journals. But one such journal, Behavioural Ecology, recently decided to do it. They found that it led to a 33 percent increase in representation of female authors.

It's important to note that this account is not unproblematic, as noted by Singal [20] who argues that some of the tests to establish implicit bias lack reproducibility, among other flaws (though see also Schwenkler *et al* [21] for further discussion of this issue, plus Tversky and Kahneman [24] on cognitive bias).

Indeed, the issue is not only with attempting to find causes, but in our expectations of what we can do once (if) a cause is established. For instance, can a person simply ignore their bias? Caliskan-Islam *et al* [5] make the claim that, 'language itself contains recoverable and accurate imprints of our historic biases', which, as a result, means that any Natural Language Processing tool will 'share the same biases humans demonstrate in psychological studies'.

Interesting though these ideas are, the fatalism of such claims requires that language stands still in a way that is not borne out by evidence. While it is true that languages express biases, it is simplistic to conclude that such biases are necessarily encoded in that language. For instance, the term *hysterical* was originally applied to women (with the *womb* at its core as an indication of an apparent tendency towards exaggerated emotion). Yet in more recent times it has begun to lose some of this historical meaning. For instance, while it may be true to say that the original meaning often remains when used in relation to a person (and a man is rarely described as hysterical), it seems not to be so when applied to a situation (consider *that comedian was hysterical*, in instances where the comedian is male). This is unlike the term *bossy* which, though without an obviously gendered origin is nevertheless sometimes synonymous with women who take charge or are managers. That these associations ebb and flow

seems indisputable, even if in a particular time there may be inherent linguistic biases (e.g. *a woman is bossy while a man is assertive, and so on*).

The problem of context and meaning is not a new one, and it more obviously comes to the fore where insults or profanities are traded. The term *gorilla* is not offensive *in and of itself*, but when applied to a group of people, this term is embedded in a Wittgensteinian language game [27], and a particularly horrible one at that. More specifically this term as applied in this instance taps into a context of systemic racism, oppression, abuse, as well as a history of dehumanising certain groups of people for the purposes of slavery. As insults go, whether intentional or otherwise it is more than simply abhorrent; it is also very harmful.

In the terms Wittgenstein offers, we can see biases in language as reflections of social and political biases, embedded in cultures and enacted by societies. The semantics of a language can change just as quickly as a culture is willing or able to change its values and priorities. Such approaches include where terms are *reclaimed*, as by minority groups and communities, or in other instances are *appropriated*, with terms and meanings altered (it remains odd to some native Italian speakers that *ragu* is generally called Bolognese, as in *Spaghetti Bolognese*, in the UK).

Thus it is the context in which language is meaningful, and which imbues language with meaning. Wittgenstein's idea is that our use of language represents an *activity*, and these activities or *forms of life* are varied. Accordingly the biases within languages, especially those which lead to negative consequences, are not self-evident or necessary. Meanings, as well as rules for their use, are determined by the contexts or activities in which they are used (e.g. in giving or obeying orders, thanking, cursing, joking etc.). Our practices, traditions, and habits enforce these non-determinate, yet rule based language-games [27, §199 and §202], and it is to these that we should direct our attention. Beginning first with the practice of dialogue.

4 WHY CAN'T WE JUST...?

"If it isn't my experience, it couldn't possibly be yours."

So why can't we just move on from dialogues on racism, gender, disability, and so on? Well, part of the issue concerns the *we* that is encompassed in the above statement. Though it seems to offer an inclusive grouping of humans generally, in fact it includes only those in a position to move on from such discourse. In other words, *if you don't see the need to talk about it, chances are you're not the one it's affecting*. Philosophy (as with computing, and a great many other disciplines and fields of research, particularly in the sciences, but not only) has a problem with inclusivity. Though it is beyond the scope of this paper to know *why*, I can offer the following observations.

First, philosophy and computing have at their hearts principles of *reason* and of *logic*, and in philosophy the aim is often towards *truth* in objective terms (or as objective as we might hope we can be). As Dotson notes [], 'A great deal of what passes for philosophical engagement is some form of criticism. It is fairly common to witness in those newly awakened to the world of academic philosophy the attitude that "being philosophical" amounts to "being critical" of others and their beliefs' [8, p. 404]. Such criticism-heavy approaches rely on technical skills, unimpassioned debate, and a somewhat adversarial method.

Yet philosophical debate, research, and engagement cannot, and never will be *value neutral*; we bring our biases, preferences, prejudices, opinions, hopes, and wishes with us. And this comment applies equally well to the sciences and computing. As Feng notes, with regard to design and technology [10, p. 212],

the determination of the "best" design results from a complex set of factors including cultural assumptions, economic interests, political considerations, organizational constraints, and so on. In a very deep sense, the design of technology has to be seen as a value-laden activity.

Similarly, Wittgenstein disputed the idea that academics (or anyone else for that matter) can somehow view things *sub specie aeterni*. In a note written to Sraffa in 1935 he describes the thinking of 'Cambridge people':

Here are people who try to speak in a queer way 'impartially' about things, they pretend to be able to slip out of their own skins and they speak as though they could understand everybody's feelings, wishes, tendencies etc. [15, p.235 n.7]

It is partly out of these kinds of concerns that he develops the motif of the *Language-game*, which evokes a sense in which, understanding language requires involvement in that language. Just as we understand a game—how it is played and the rules—through playing/watching the game, we understand a language by using it. We can no more view our language from an objective perspective than we can *slip out of our skin*. This has a very definite impact on academic practice, because it shows why, at the core, some disciplines may be full of biases. The argument goes that if diversity is limited in these fields, so too will be the sorts of voices heard, as well as representations of these kinds of voices. And in these respects, few of the details above regarding Google, Amazon, or philosophy departments are particularly surprising, at least not to those of us who are marginalised.

The second observation concerns what Saul refers to as *stereotype threat*, and it is sometimes grouped in with something called a *self-fulfilling prophecy*. This highlights the somewhat uncomfortable link between the stereotype that someone wants to avoid, yet in so doing they find themselves in the firing line. As Saul notes, this occurs when, for instance, "Victims of stereotype threat underperform on the relevant tasks because they are unconsciously preoccupied by fears of confirming the stereotypes about their group." In other words, and to clarify what is meant by underperformance, "some people face barriers that impede their performance. These people underperform in the sense that their performance is negatively affected by these barriers" [19, p. 41]. In other words, if the expectations are that a field like philosophy or computing offers *objective* arguments, with results that are somehow *neutral, logical, reasonable*, then this can perpetuate the view that the problems exist elsewhere. In the case of publications, for instance, if the belief is that these processes are fair, with the outcomes that fewer minority groups are included, then this can perpetuate the belief that minorities are not good at philosophy. Such experiences among marginalised groups can perpetuate the belief that publishing in such journals is for certain groups only.

Finally, and in concrete terms, representation requires presence. In short, to be an academic requires the time and effort to get to that position, as well as the resources, and even then it requires that people believe in, invest in, and support you. There is a wealth of evidence to suggest that marginalised groups often lack these support systems, whether inside or outside the

academy, and as a result representation will necessarily be lacking. This impacts on research, on development, on what is included, what excluded. The *qualia* of what it is to be marginalised has not (so far) been represented (or even translated) in the academy, which has resulted in certain monocultural areas. Philosophy and computing are some examples, and until this changes it is unlikely the tendency towards certain kinds of bias or prejudice will diminish. As Caliskan-Islam *et al* [5, p. 10] note,

if AI is to exploit via our language the vast knowledge that culture has compiled, it will inevitably inherit human-like prejudices. In other words, if AI learns enough about the properties of language to be able to understand and produce it, it also acquires cultural associations that can be offensive, objectionable, or harmful. These are much broader concerns than intentional discrimination, and possibly harder to address.'

These biases are perpetuated by those who would seek to silence voices, or to downplay the experiences of those who are *other* to them. If this conversation is avoided in academia, we cannot assume it will somehow happen elsewhere. As Winner notes, 'scientific knowledge, technological invention, and corporate profit reinforce each other in deeply entrenched patterns that bear the unmistakable stamp of political and economic power' [26, p. 126], and we cannot rely on those with power to simply share it.

5 CONCLUSION

This paper has sought to contribute to a wider debate on the issue of bias, primarily in terms of negative bias (whether in intention or in outcome) that can be found in both philosophy and technology, and to relate this to the issue of a lack of diversity. I sincerely hope that the question *why can't we just stop talking about race?* has been answered to some extent. That said, as this is the first incarnation of this paper, I expect I've only begun to scratch the surface of an answer to such a big question. If it is somewhat lacking and in need of further development, I hope that the participants of AISB will see an argument worth engaging with, and an invitation to contribute to the debate.

I'll close with two quotations from Winner [26], both of which offer prudent advice with regard to the issues noted above. First concerns where we locate the source (or even cause) of problems like bias, and second covers the reasons for which we might expect harm to occur. On the first issue he who offers stern advice to those who would attribute political qualities to technical artefacts in and of themselves:

What matters is not technology itself, but the social or economic system in which it is embedded. This maxim, which in a number of variations is the central premise of a theory that can be called the social determination of technology, has an obvious wisdom. It serves as a needed corrective to those who focus uncritically on such things as "the computer and its social impacts" but who fail to look behind technical things to notice the social circumstances of their development, deployment, and use. This view provides an antidote to naive technological determinism—the idea that technology develops as the sole result of an internal dynamic, and then, unmediated by any other influence, moulds society to fit its patterns [26, p. 122]

It is at least partly as a result of this sort of view that my paper suggests we neither locate the problem in the technology, nor even in the language, as though the issues somehow lie *outside* of our own practices and customs. His second piece of advice is that we should note that

many of the most important examples of technologies that have political consequences are those that transcend the simple categories of "intended" and "unintended" altogether. These are instances in which the very process of technical development is so thoroughly biased in a particular direction that it regularly produces results counted as wonderful breakthroughs by some social interests and crushing setbacks by others. In such cases it is neither correct nor insightful to say, "Someone intended to do somebody else harm." Rather, one must say that the technological deck has been stacked long in advance to favour certain social interests, and that some people were bound to receive a better hand than others. [26, pp. 125-6]

More than 35 years later, this point remains as pertinent as it ever was. Or perhaps more so, given we've had so much more time to reflect on what we might need to do differently. The question is what we're prepared to accept in the short term, and when we will say that that short term measures are no longer acceptable. After all, there comes a point where whatever we keep doing becomes entrenched as *what we do*. Disciplines with reason at their core would do well to avoid such dogmatic slumbers.

REFERENCES

- [1] L. Alexander Is an algorithm any less racist than a human? *Guardian* (2016). <https://www.theguardian.com/technology/2016/aug/03/algorithm-racist-human-employers-work>
- [2] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica* (2016).
- [3] S. Barocas and A. D. Selbst Big data's disparate impact. *California Law Review*, 104 (2014).
- [4] L. K. Bright. Publications By Black Authors in Leiter Top 15 Journals 2003-2012 *The Splintered Mind*, (2016). <http://schwitsplinters.blogspot.co.uk/2016/01/publications-by-black-authors-in-leiter.html>
- [5] A. Caliskan-Islam, J. J. Bryson, and A. Narayanan. Semantics derived automatically from language corpora necessarily contain human biases. arXiv preprint arXiv:1608.07187 (2016).
- [6] B. X. Chen. HP Investigates Claims of 'Racist' Computers, *Wired* (2009). <https://www.wired.com/2009/12/hp-notebooks-racist/>
- [7] K. Crawford. Artificial intelligence's white guy problem. *The New York Times* (2016).
- [8] K. Dotson. Concrete Flowers: Contemplating the Profession of Philosophy'. *Hypatia*, 26(2), pp. 403-409 (2011).
- [9] B Evans and G. Yancy The Perils of Being a Black Philosopher. *The New York Times* (2016).
- [10] P. Feng. Rethinking technology, revitalizing ethics: overcoming barriers to ethical design. *Science and Engineering Ethics*. April 2000. 6: 2. pp. 207-220 (2000).
- [11] A. Frankel. Why We Should Stop Talking About Race *The Odyssey* (2016). <https://www.theodysseyonline.com/stop-talking-race>
- [12] A. Gabbiani, P. Riva, L. Andrighetto, C. Volpato, B. J. Bushman. Acting like a Tough Guy: Violent-Sexist Video Games, Identification with Game Characters, Masculine Beliefs, & Empathy for Female Violence Victims. *Plos One* 11(4) (2016).
- [13] K. Healy. Citation networks in philosophy (2013). <https://kieranhealy.org/blog/archives/2013/06/24/citation-networks-in-philosophy-some-followup/>

- [14] D. Ingold and S. Soper. Amazon Doesn't Consider the Race of Its Customers. Should It? *Bloomberg* (2016). <https://www.bloomberg.com/graphics/2016-amazon-same-day/>
- [15] B. McGuinness What Wittgenstein Owed to Sraffa. In G. Chiodi & L. Ditta (Eds.), *Sraffa or an Alternative Economics*. Basingstoke: Palgrave Macmillan (2007).
- [16] W. Morgan Why is my professor still not black? *Times Higher Education* (2016).
- [17] M. Mulshine. A major flaw in Google's algorithm allegedly tagged two black people's faces with the word 'gorillas' *Business Insider Australia* (2015).
- [18] G. Priest. What is philosophy?. *Philosophy*, 81(02), pp.189-207 (2006).
- [19] J. Saul. Implicit bias, stereotype threat, and women in philosophy. *Women in philosophy: What needs to change*: pp. 39-60 (2013).
- [20] J. Singal. Psychology's Favorite Tool for Measuring Racism Isn't Up to the Job. *New York Magazine* (2017). <http://nymag.com/scienceofus/2017/01/psychologys-racism-measuring-tool-isnt-up-to-the-job.html>
- [21] J. Schwenkler *et al* What can we learn from the Implicit Association Test? A Brains Blog Roundtable (2017). <http://philosophyofbrains.com/2017/01/17/how-can-we-measure-implicit-bias-a-brains-blog-roundtable.aspx>
- [22] L. Sweeney. Discrimination in online ad delivery. *Queue*, 11(3):10:10–10:29 (2013).
- [23] S. Tremain. Disabling philosophy *The Philosopher's Magazine* (2014).
- [24] A. Tversky. and D. Kahneman. Rational choice and the framing of decisions. *Journal of business*, pp.S251-S278 (1986).
- [25] J. Weinberg. Blacks in Philosophy in the US *Daily Nous* (2014). <http://dailynous.com/2014/08/28/blacks-in-philosophy-in-the-us/>
- [26] L. Winner. Do artifacts have politics?. *Daedalus*, 109(1) pp.121-136 (1980).
- [27] L. Wittgenstein. *Philosophical Investigations*, Oxford: Blackwell (2001).