Philosophy & Technology (2022) 35:81
https://doi.org/10.1007/s13347-022-00576-6

COMMENTARY

Algorithmic Political Bias Can Reduce Political Polarization

Uwe Peters[1,2]

Received: 26 July 2022 / Accepted: 11 August 2022

Abstract
Does algorithmic political bias contribute to an entrenchment and polarization of political positions? Franke (*Philosophy and Technology*, 35, 7, 2022) argues that it may do so because the bias involves classifications of people as liberals, conservatives, etc., and individuals often conform to the ways in which they are classified. I provide a novel example of this phenomenon in human–computer interactions and introduce a social psychological mechanism (what I shall call 'implied political labeling') that has been overlooked in this context but should be experimentally explored. Furthermore, while Franke proposes that algorithmic political classifications entrench political identities, I contend that they may often produce the opposite result. They can lead people to change in ways that disconfirm the classifications (thus causing 'looping effects'). Consequently and counterintuitively, algorithmic political bias can in fact *decrease* political entrenchment and polarization.

Keywords Algorithmic political bias · Political entrenchment · Looping effects · Political polarization

1 Introduction

Some AI systems that are currently being used for decision-making about people in hiring, clinical, or many other domains may display *algorithmic bias*, i.e. they may operate on data in ways that systematically deviate from a normative (moral, epistemic, etc.) standard such that some people are unfairly privileged over others based on their social identity. I recently argued that political orientation (being liberal, conservative, etc.) is one aspect of social identity and that algorithmic bias against it (*algorithmic political bias*) may pose significant and distinctive ethical and epistemic risks that have gone unnoticed in the AI literature (Peters, 2022).

✉ Uwe Peters
up228@cam.ac.uk

[1] Center for Science and Thought, University of Bonn, Bonn, Germany

[2] Leverhulme Centre for the Future of Intelligence, University of Cambridge, Cambridge, UK

Published online: 23 August 2022

In his interesting and sympathetic reply, Franke (2022) presents 'one more distinctive risk related to algorithmic political bias: the risk that such bias exacerbates political entrenchment to the detriment of the polity' (p. 1). To make his point, Franke draws (*inter alia*) on Hacking's (1999) work on the social construction of human kinds and argues that algorithmic political bias involves social classifications of people in terms of political categories that may solidify political identities by eliciting conformist responding.

Franke's approach of relating Hacking's research to algorithmic political bias is promising. I shall illustrate this by giving a new example of how even merely implied (not explicit) political classifications in human–computer interactions can reinforce political identities. However, unlike Franke, who emphasizes the identity-*entrenching* impact of political classifications, I think that the classifications involved in algorithmic political bias can in fact frequently *reduce* political entrenchment and polarization. I begin by briefly introducing Franke's argument.

2 The Construction of Political Identities and the Zigzagging of Politics

Franke connects algorithmic political bias to Hacking's (1999) work on human kinds. Hacking argued that when an individual is publicly classified as a refugee, criminal, obese, and so on, these classifications come with social expectations that can lead those classified to change (for better or worse) so as to conform to the classifications and gradually become the person that they are classified as. This is because people may wish to belong to a certain group, may want to avoid being sanctioned for violating expectations, or may just like to 'behave in ways that are expected of [them]' (Hacking, 1995, p. 21). Given their interactive effects, social classifications can thus help construct social identities (for details, see also Peters, 2020, pp. 8, 15–18; 2021, pp. 6–13).

Franke notes that algorithmic political bias, too, can belong to a 'matrix' of social classifications. For instance, in AI-assisted recruitment, based on their digital footprints, people may become algorithmically classified as 'liberal' or 'conservative' and potentially encounter preferential or negative treatment due to their political orientation. To respond adaptively (e.g., to 'get scholarships or employments, or to get academic manuscripts published'), Franke argues, "we may [thus], to paraphrase Hacking, 'learn what characteristics to establish [and] know how to live our lives' to fit the ideas of certain political positions" that benefit us (2022, p. 4). When anticipating certain algorithmic political labels and preferences in hiring and so on, we may, for instance, for prudential reasons more strongly signal them, and those of us still politically undecided or moderate may adopt more committal, stable positions, Franke suggests.

This way in which algorithmic political bias can make people more politically entrenched differs from another, more familiar one: On the Internet, many websites personalize online content to us based on (*inter alia*) our political attitudes. This process, too, can be viewed as a kind of algorithmic political bias as it involves algorithms favoring in their operations some contents or individuals over others based on their political identity (Peters, 2022, p. 5). These algorithms may solidify people's political orientations by presenting website users predominantly with

online contents that support their political views. The process Franke highlights is interestingly different: algorithmic political bias can promote political entrenchment *indirectly* through people's perception of, and reaction to, political labels.

Why is this entrenchment problematic?[1] Franke contends that no political position will always adequately include all of the values and goals one wants pursued in the political realm. Correspondingly, many people in the electorate (e.g. 'swing voters' in the US) usually change their voting over time rather than consistently choose the same party. Franke argues that in contributing to an entrenchment of political orientations, algorithmic political bias can impede this healthy political 'zigzagging' (Nozick, 2006, p. 286).

I agree that the prospect of facing algorithmic political bias in hiring and other areas can motivate people to learn what characteristics to establish to fit certain political positions and publicly signal commitment to them. However, the signaling of political identity (to avoid or exploit political bias) is perhaps primarily only needed in domains where people can identify us. Election voting isn't such a domain. It happens anonymously. People thus aren't under pressure to show their political identity and so may not become entrenched in their voting preferences. Why should we assume that the prospect of algorithmic political classification and bias also contributes to genuine changes in, for instance, our self-perception such that it may affect us even in domains where we aren't identifiable? Franke doesn't elaborate, leaving it somewhat mysterious how political classifications involved in algorithmic political bias may make people truly more liberal, conservative, and so on. To remedy this, I shall now introduce an example and a mechanism that help explain how such political entrenchment may arise.

3 The Effects of Implied Political Labeling

Return to website personalization, which may involve an algorithmic political bias when algorithms classify us according to our political orientation before selectively presenting us with online contents tailored to it. While this kind of algorithmic bias can contribute to a political entrenchment directly (by causing excessive exposure to viewpoint-consistent contents), it can also have important, so far largely unexplored indirect labeling-related entrenchment effects.

Consider first a study on algorithmic labeling that does not involve political categories such as 'liberal', 'conservative', etc. Summers et al. (2016) found that after exposure to an online ad for a sophisticated watch brand, participants evaluated themselves as more sophisticated and intent on buying the watch when they thought the ad had been targeted to them based on their previous browsing than when they thought it was just based on their age, gender, or no algorithmic personalization at all. Summers et al. argue that participants viewed the ad as implying a social label, as indicating a personality feature, namely sophisticated taste, that personalization algorithms had ascribed to them based on their browsing. This led participants to adjust their self-perception so as to align it with the implied algorithmic label,

[1] An entrenchment of political views can have significant benefits in making the realization of desired political states more likely; see Peters (2020).

resulting in them viewing themselves as more sophisticated. Importantly, this kind of effect occurred with other implied personality ascriptions too: After receiving a personalized ad for an environmentally friendly product, participants rated themselves as more 'green' and were subsequently more willing to buy the advertised product and donate to a pro-environmental charity (ibid).

Summers et al. didn't investigate whether political labels ('liberal', etc.) implied by algorithmic recommendations of certain articles or products can have similar effects. But there is little reason to doubt it. This is because viewing oneself as more 'green' in response to online ads of green products that one believes to be personalized is already closely related to a political self-identification (see also 'green politics'). Call this indirect process of inducing alterations in people's political self-perception 'implied political labeling'. If implied political labels can have reinforcement effects (a proposition that should be experimentally explored) then we have an example and mechanism of how political labeling tied to algorithmic political bias in website personalization may increase political entrenchment: implied political labels (in website personalization) might lead individuals to view themselves as more liberal, conservative, and so on.

Indeed, such intriguing effects are likely to be especially powerful when the political labels come from *algorithms* rather than from humans. This is because labeling effects depend on the authority of the labeler: We 'tend to behave in ways that are expected of us, *especially by authority figures* – doctors, for example [emphasis added]' (Hacking, 1995, p. 21). The higher a labeler's (perceived) authority is, the greater the chance that the target believes in the label's accuracy and conforms to it. Now, many people allocate especially high epistemic authority to algorithms and display 'automation bias', i.e. they prefer suggestions from automated decision-making systems and tend to disregard contradictory information made without automation even when it is accurate (Goddard et al., 2012; Logg et al., 2019). Given this, many individuals should also be particularly prone to endorsing algorithmic (e.g. implied) social labels. This should make conformist effects that reinforce political identities especially likely outcomes of (implied or explicit) political classifications in website personalization.

4 The Looping Effects of Algorithmic Political Bias

The preceding point supports Franke's argument. However, Hacking (1995, p. 21) also emphasizes 'looping effects': People classified in a certain way often *reject* their social classification and change themselves to disconfirm it, forcing revisions of it. While the process outlined in the previous section tends to stabilize social classifications, looping effects (as I shall here narrowly construe them) tend to *de*-stabilize them. I think that the political classifications involved in algorithmic political bias, too, may often trigger looping effects and not (pace Franke) political entrenchment.

Return to political bias in, for instance, hiring contexts. While Franke suggests that the increased prospect of being algorithmically politically labeled and subject to political bias may increase political entrenchment, often the opposite seems likely. Wary of the increasing risk of being algorithmically labeled 'liberal', 'conservative' and so on in hiring and other high-stakes domains, people may make more efforts than previously to avoid the adoption and expression of clearly identifiable liberal, conservative,

etc. positions. Indeed, the now reduced chances of success in hiding one's political orientation from algorithms (e.g. see 'political' face recognition AI; Peters, 2022, p. 14), the corresponding higher vulnerability to potential bias-related harms, and people's increasing fatigue about political polarization and partisanship (Klar et al., 2018) should make a personal political de-polarization increasingly more attractive to many. For political orientations are perhaps more easily detected (e.g. via algorithms tracking social media contents) and more disfavored the more engrained and extreme they are, making it increasingly more adaptive to avoid holding fixed orientations that clearly distinguish oneself from one's political opponents. Signaling instead a lack of partisanship in one's digital footprint can undercut algorithms' ability to correlate oneself with a particular political orientation, reducing one's risk of being harmed by political bias.

Granted, more political neutrality may also involve a higher risk of losing out on *benefits* in domains where there is a bias towards a particular political orientation. However, to the extent that one can't generally be certain about such a specific bias in all domains in which one may need to get past algorithms that are able to detect political orientations, it likely remains overall more risky to signal a rigid political identity. Since risk aversion is pervasive (Zhang et al., 2014) and convincingly signaling certain orientations (incl. neutrality) depends on actually holding the underlying beliefs (Peters, 2020), when people become more aware of potential algorithmic political biases in AI systems, this can prompt them to de-polarize politically. Hence, while an increasing, perceived[2] presence of algorithmic (incl. implied) political labeling may make some people 'stay committed to their [political] positions come what may' (Franke, 2022, p. 5), and while algorithmic political bias clearly has significant ethical and epistemic costs (see Peters, 2022), it can also cause many individuals to become more politically *flexible*. It remains to be seen, however, whether the overall effects of algorithmic political bias related to political entrenchment and polarization will in the end mitigate them more than deteriorate them.[3]

In any case, since we can arguably change our political orientation more easily than our gender or racial identity, there is, then, a sense in which algorithmic political bias can significantly differ from algorithmic gender or race biases: The bias may indirectly (via individuals' perception) contribute to a reduction in people's possession of the very feature of social identity that the bias targets (i.e. a fixed political orientation). By triggering looping effects, algorithmic political bias can thus both de-polarize and de-politicize people.

5 Conclusion

Franke (2022) contends that algorithmic political bias can exacerbate political entrenchment because it increases the social space where people are politically classified, and those socially classified often conform to the classifications. I gave a new example of how this can happen through merely implied classifications during website

[2] The argument only requires that people believe there to be algorithmic political labeling and bias in hiring etc. The de-polarizing effects outlined can occur even if such labeling or bias is in fact absent.

[3] My argument here assumes that people often act prudently and in risk-averse ways. To what extent this is the case and whether algorithmic bias has predominantly de-polarizing or polarizing effects remain interesting and important open empirical questions. The key point here is that the bias can have both kinds of effects.

personalization affected by algorithmic political bias. However, I argued that individuals' prospect of encountering algorithmic political labeling and bias may in fact often also attenuate political entrenchment: In being able to detect people's political orientations in previously impossible ways, many algorithms (e.g. website personalization or face recognition AI systems) make people increasingly more vulnerable to becoming targets of political bias in more domains than before. This can provide a strong incentive for risk averse, i.e. most individuals to refrain from holding fixed political orientations that are clearly distinct from those of their opponents. Raising awareness of potential algorithmic political bias in AI systems may thus help tackle a key social problem that facilitates the emergence of this bias: political polarization.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Ethical Approval Not applicable.

Consent to Participate Not applicable.

Consent to Publish Yes.

Competing Interests The author declare no competing interests.

References

Franke, U. (2022). Algorithmic political bias – an entrenchment concern. *Philosophy and Technology, 35*, 7. https://doi.org/10.1007/s13347-022-00562-y

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association: JAMIA, 19*(1), 121–127.

Hacking, I. (1995). The looping effects of human kinds. In D. Sperber, D. Premack, & A. J. Premack (Eds.), *Causal cognition: A multidisciplinary debate* (pp. 351–394). OUP.

Hacking, I. (1999). *The social construction of what?* Harvard University Press.

Klar, S., Krupnikov, Y., & Ryan, J. B. (2018). Affective polarization or partisan disdain? Untangling a dislike for the opposing party from a dislike of partisanship. *Public Opinion Quarterly, 82*(2), 379–390.

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes, 151*(151), 90–103.

Nozick, R. (2006). *The examined life: Philosophical meditations*. Simon and Schuster.

Peters, U. (2022). Algorithmic political bias in artificial intelligence systems. *Philosophy and Technology, 35*(2), 25. https://doi.org/10.1007/s13347-022-00512-8

Peters, U. (2020). What is the function of confirmation bias? *Erkenntnis*. https://philarchive.org/archive/PETWIT-6. Accessed 14 Nov 2021.

Peters, U. (2021). Science communication and the problematic impact of descriptive norms. *British Journal for the Philosophy of Science*. https://philpapers.org/archive/PETSCA-5.pdf. Accessed 28 July 2021

Summers, C. A., Smith, R. W., & Reczek, R. W. (2016). An audience of one: Behaviorally targeted ads as implied social labels. *Journal of Consumer Research, 43*, 156–178.

Zhang, R., Brennan, T. J., & Lo, A. W. (2014). The origin of risk aversion. *Proceedings of the National Academy of Sciences of the United States of America, 111*(50), 17777–17782.