Virtue signaling and moral progress

Evan Westra, York University

Forthcoming in *Philosophy and Public Affairs*

Abstract: 'Virtue signaling' is the practice of using moral talk in order to enhance one's moral reputation. Many find this kind of behavior irritating. However, some philosophers have gone further, arguing that virtue signaling actively undermines the proper functioning of public moral discourse and impedes moral progress. Against this view, I argue that widespread virtue signaling is not a social ill, and that it can actually serve as an invaluable instrument for moral change, especially in cases where moral argument alone does not suffice. Specifically, virtue signaling can change the broader public's social expectations, which can in turn motivate the adoption of new, positive social norms. I also argue that the reputation-seeking motives underlying virtue signaling impose important constraints on virtue signalers' behavior, which serve to keep the worst excesses of virtue signaling in check.

1. Introduction

Virtue signaling is the act of engaging in public moral discourse in order to enhance or preserve one's moral reputation.[1] Typical examples of virtue signaling might include an individual making a social media post vehemently condemning some offensive action taken by a public figure, or a brand launching a marketing campaign that invokes themes of social justice. What makes the act in question an instance of virtue signaling is not the content of the moral expression itself, but rather the status-seeking desires of the person or corporate entity making it.

[1] James Bartholomew, "Easy Virtue," *The Spectator*, April 18, 2015; David Shariatmadari, "'Virtue Signalling' the Putdown That Has Passed Its Sell by Date," The Guardian, January 20, 2016.

One engages in virtue signaling in the hopes of seeing one's moral reputation improve in the eyes of one's peers (or potential customers); the desire to make a constructive, sincere contribution to public moral discourse is at best a secondary motivation.[2]

It is easy to see why 'virtue signaling' is a pejorative term, and why many find it annoying. But for some, the prospect of widespread virtue signaling represents something much darker. Recently, philosophers Justin Tosi and Brandon Warmke have argued that virtue signaling (or, as they call it, *moral grandstanding*[3]) is not just irritating, but actively harmful to the proper functioning of public moral discourse itself, such that it impairs our collective ability to improve our moral beliefs and promote positive moral changes in the world.[4] When contributions to public debate are driven by status-seeking goals, these authors argue, this leads to a series of increasingly inflated moral claims, exaggerated expressions of outrage, and aggressive piling on and public shaming. This causes onlookers to disengage from moral discourse, either because of increased cynicism about their would-be interlocutors' motives, increased political polarization, or sheer exhaustion. In short, when the sphere of public moral discourse ceases to be a forum for

[2] The internal, psychological nature of virtue signaling makes it difficult to determine whether a given act of moral speech is a case of virtue signaling. This makes the empirical study of virtue signaling and its real-world effects challenging. It also means that the term is prone to false accusations. Indeed, one might legitimately worry that the label 'virtue signaling' can be used to silence or undermine legitimate moral concerns, which would be a form of testimonial injustice; see Miranda Fricker, *Epistemic Injustice: Power and the Ethics of Knowing* (Oxford: Oxford University Press, 2007). For the purposes of this paper, I am setting these undeniably problematic cases aside, and am focusing solely upon instances of genuine, status-seeking virtue signaling.

[3] Tosi and Warmke use the technical term 'moral grandstanding' because they see the term 'virtue signaling' as politically charged and prone to certain misunderstandings; see Justin Tosi and Brandon Warmke, *Grandstanding: The Use and Abuse of Moral Talk* (New York: Oxford University Press, 2020), 37–40. However, their definition of moral grandstanding coincides with the commonsense understanding of 'virtue signaling,' and so it is reasonable to treat the two terms as synonymous, even if they have slightly different connotations. Since it is more widely understood, I have chosen to use 'virtue signaling' in this paper; see also Neil Levy, "Virtue Signalling Is Virtuous," *Synthese*, 2020, 1–18.

[4] Justin Tosi and Brandon Warmke, "Moral Grandstanding," *Philosophy and Public Affairs* 44, no. 3 (2016): 197–217; Tosi and Warmke, *Grandstanding: The Use and Abuse of Moral Talk*.

the sincere exchange of moral ideas and turns into an arena for competitive moralizing, it just stops working.[5]

In this paper, I offer a defense of virtue signaling and its role in public moral discourse. While it may not be particularly praiseworthy or noble, I will argue virtue signaling is also not something that should be actively discouraged, nor should we think of it a dangerous social ill. Rather, virtue signaling and the status-seeking motives behind it should be viewed as neutral, stable features of the social environment that function as vehicles for the diffusion of social norms. Far from standing in the way of moral progress, understanding virtue signaling this way actually reveals how it can serve as an invaluable tool for positive social change. I will also argue that certain reliable features of human psychology – our epistemic and moral vigilance – create social constraints against excessive virtue signaling, limiting the potential for vicious spirals of the sort that might lead to widespread hypocrisy and cynicism. In short, virtue signaling is no great threat to the integrity of public moral discourse, and if we correctly understand its relation to social norms, it can even serve as an instrument for positive moral change.

The structure of this paper is as follows: in section two, I discuss how we should understand the relationship between moral discourse and moral progress. In section three, I provide the conception of social norms that will underlie my key positive claims about virtue signaling. In section four, I discuss how, in light of this conception of social norms, one would need to go about changing them. In section five, I show how virtue signaling can play a role in this process. Sections six and seven further defend virtue signaling by addressing worries about its potential harms, and outline some of the practical and epistemic constraints on the contents of virtue

[5] Tosi and Warmke do not deny that virtue signaling might occasionally have positive moral and social consequences, but they view these positive effects as incidental and vastly outweighed by more negative ones; see Tosi and Warmke, "Moral Grandstanding"; Tosi and Warmke, *Grandstanding: The Use and Abuse of Moral Talk*.

signals. Section eight addresses concerns about the potential for virtue signaling to contribute to moral and political polarization. Section nine responds to the worry that the moral progress wrought by virtue signaling and norm change is too thin by arguing that widespread shifts in social norms can precipitate deeper changes at the level of a community's moral values.

2. Public moral discourse and moral progress

Going forward, my focus will be on the way that virtue signaling affects the proper functioning of public moral discourse. As such, I accept Tosi and Warmke's claim that public moral discourse is supposed 'to improve people's moral beliefs, or to spur moral improvement in the world.[6]' However, it is worth dwelling for a moment on the different ways that public moral discourse might promote these ends. One very natural way of thinking about this process is to understand it in epistemic, deliberative terms: moral talk exposes us to new moral arguments, which leads to improvements in people's moral beliefs, which in turn motivates moral action and good consequences. When deciding whether to take part in a protest or to support a particular policy, for example, agents can visit the public square, weigh the moral reasons being presented on all sides of the debate in question, and then act based on what they take to be the most compelling arguments. Because the arguments presented in the public square are sincere attempts by epistemically responsible agents to arrive at moral truth, this process provides us with an overall reliable procedure for achieving moral progress. This epistemic, deliberative conception of public moral discourse seems to be what Tosi and Warmke have in mind in their arguments against virtue signaling: because their moral claims are driven by a desire for status rather than a desire for truth, virtue signalers cause the contents of moral discourse to become

[6] Tosi and Warmke, "Moral Grandstanding," 209.

untethered from the reality, thereby undermining its epistemic integrity.[7] According to this view, public moral discourse is supposed to help us reason our way to a better world, while virtue signaling blows us off-course.

This epistemic conception of how moral talk contributes to moral progress is incomplete. As we shall see, the deliberative process envisioned by Tosi and Warmke often fails to spur moral progress, even when it succeeds in changing people's moral beliefs. This is because many of the social practices that stand in the way of moral progress are not motivated by moral beliefs at all, but rather by the strong desire to conform to local norms.[8] To get people to abandon these practices, it is neither necessary nor sufficient to change their moral beliefs: one must instead change the social beliefs that undergird their beliefs about what the people around them think and do. Public moral discourse can however function as a vehicle for conveying precisely this kind of social information, albeit not by the deliberative route that Tosi and Warmke envision. Instead, moral claims in the public square double as a source of evidence through which people are able to infer the norms of their local community. It is in this role that virtue signaling can serve as a means for moral progress: not as a source of evidence for moral beliefs, but as a vector for information about social norms.[9]

3. Social norms

[7] See for example Tosi and Warmke, *Grandstanding: The Use and Abuse of Moral Talk*, 74–74, 90; see also Tosi and Warmke's responses in C.A.J. Coady, "Philosophy & Public Affairs Discussion at PEA Soup: Justin Tosi and Brandon Warmke's 'Moral Grandstanding,'" PEA Soup, 2017.

[8] For reviews of empirical work on this topic, see Cristina Bicchieri, *The Grammar of Society* (New York: Cambridge University Press, 2006); Cristina Bicchieri, *Norms in the Wild: How to Diagnose, Measure, and Change Social Norms* (New York: Oxford University Press, 2017).

[9] In his recent defense of virtue signaling, Neil Levy argues that virtue signaling can be defended on epistemic grounds as well, since virtue signaling can serve as higher-order evidence for our moral beliefs via expressions of confidence and social consensus. In section 7, I will give a different epistemic defense of virtue signaling, arguing that its reliability is shored up by the epistemic vigilance of the virtue signaler's audience. Levy, "Virtue Signalling Is Virtuous."

Following Cristina Bicchieri,[10] I define social norms as behavioral practices that individuals adopt conditional on certain beliefs about how other members of their community behave and think, or *social expectations*. Specifically,

> A *social norm* is a rule of behavior such that individuals prefer to conform to it on condition that they believe that (a) most people in their reference network conform to it (empirical expectation), and (b) that most people in their reference network believe they ought to conform to it (normative expectation).[11]

An individual's *reference network* refers to the group of people whose behaviors and attitudes they care about when making various decisions. This group can change depending on the context, and the people in it need not be physically proximate (especially in the digital age). A reference network can include members of one's local community, profession, political party, and so on.[12] As will become important later when I discuss the diffusion of norms between different groups, a person can belong to many different reference networks.

Social norm conformity depends upon two specific types of belief: beliefs about whether others in one's reference network are conforming to a social norm, or *empirical expectations*¸ and

[10] Bicchieri, *The Grammar of Society*; Bicchieri, *Norms in the Wild: How to Diagnose, Measure, and Change Social Norms*.
[11] Bicchieri, *Norms in the Wild: How to Diagnose, Measure, and Change Social Norms*, 35. This definition is meant as a *rational reconstruction* of how social norms work, rather than a literal description of the psychological mechanisms that underpin them (see also Bicchieri, *The Grammar of Society*, 3). However, several models of the mechanisms of social norm conformity do invoke social expectations similar to those that Bicchieri has in mind; see for example Matteo Colombo, "Two Neurocomputational Building Blocks of Social Norm Compliance," *Biology and Philosophy* 29, no. 1 (2014): 71–88; Jordan E. Theriault, Liane Young, and Lisa Feldman Barrett, "The Sense of Should: A Biologically-Based Framework for Modeling Social Pressure," *Physics of Life Reviews* 1 (2020): 1–37; Michael Tomasello, "The Moral Psychology of Obligation," *Behavioral and Brain Sciences* 43, no. e56 (2019): 1–58. For more on the cognitive foundations of social norms, see Kristin Andrews, "Naïve Normativity: The Social Foundation of Moral Cognition," *Journal of the American Philosophical Association* 6, no. 1 (2020): 36–56; Daniel Kelly and Taylor Davis, "Social Norms and Human Normative Psychology," *Social Philosophy and Policy* 35, no. 1 (2018): 54–76.
[12] Bicchieri, *Norms in the Wild: How to Diagnose, Measure, and Change Social Norms*, 14.

beliefs about whether others in one's reference network think that the behavior in question *ought* to be followed (where the 'ought' in question is prescriptive, rather than merely prudential or predictive), or *normative expectations*. Normative expectations are usually accompanied by the belief that non-conformity will result in social sanctions, which enforce norm compliance, thereby stabilizing empirical expectations.[13]

Bicchieri draws on a number of morally abhorrent examples of destructive social norms from around the world, including female genital mutilation and child marriage practices. But for our purposes, we can begin with a more familiar example: queuing. When one encounters a line of people waiting by a bus stop, one often ends up queuing up as well. What motivates queuing is not the belief that this is the only physically possible way to get on the bus: one could easily cut in front of everyone else once the bus arrives. Nor is queuing solely motivated by the fact that other people seem to be queuing (i.e. empirical expectations): if one thought that people just happened to have formed a line by chance, one would not see any reason to adopt that behavior. Likewise, one would not feel motivated to queue if one only had the normative expectation that everyone believed that one *ought* to queue ('If nobody else is following the rules, why should I?'). It is rather the combination of one's empirical and normative expectations that creates the motivation for queuing. Not only is everyone else queuing, but everyone else also thinks that you should queue too (and they might get angry if you don't).

Social norms thus pick out practices that are distinguished by the kinds of beliefs that move people to adopt them. Notably, these are not beliefs about the desirability or moral status of the

[13] Bicchieri distinguishes social norms from two other forms of social conformity: *descriptive norms*, which are motivated solely by empirical expectations (e.g. when one adopts a particular clothing style because others in one's reference network do so as well), and *customs*, which are motivated by non-social, practical considerations that cause people to behave in broadly similar ways (e.g. using umbrellas in the rain).

action itself. We conform to normative social practices because we are motivated by our beliefs about what *other people* do and think about those practices, not what we ourselves think about them. This means that conformity to a social norm is consistent with a person's viewing the prescribed behavior as distasteful, imprudent, or even morally objectionable. Indeed, it is even possible for a community to persist in following a social norm despite the fact that nobody actually thinks it is a good thing, provided that they all still hold the relevant social expectations about that behavior – a condition that Bicchieri calls *pluralistic ignorance.*[14]

4. Changing social norms

The psychological underpinnings of norm conformity make social norms resistant to change in the face of rational argument. If, for example, the reason that individuals in a community persist in marrying off their young daughters is because they fear the opprobrium or sideways glances of the other people in their reference network, then presenting those individuals with moral reasons to abandon the practice is unlikely to change their behavior. Indeed, a harmful social norm might persist even when the moral and rational arguments against it are well known and privately endorsed. In his book *The Honor Code: How Moral Revolutions Happen,* Appiah argues that this was the case with the practice of dueling in 18th- and 19th-century Britain.[15] The legal, religious, and moral arguments against dueling had been long acknowledged among the social circles that practiced it. Yet dueling continued to persist because it was intimately tied with one's reputation as a gentleman, a person of honor. Maintaining one's honor (and acknowledging the honor of another) meant being willing to respond to insults with lethal force, while also risking one's own life in the process. Gentlemen would participate in duels even when they knew that doing so was

[14] Bicchieri, *The Grammar of Society*, 15.
[15] Kwame Anthony Appiah, *The Honor Code: How Moral Revolutions Happen* (New York: WW Norton & Company, 2011).

foolish and immoral. To illustrate this, Appiah quotes from the will of a man written the night before he was killed in a duel: 'In the first place, I commit my soul to Almighty God, in hopes of his mercy and pardon for the irreligious step I now (in compliance with the unwarrantable customs of this wicked world) place myself under the necessity of taking.'[16]

What led to the eventual end of dueling, Appiah argues, was not the emergence of some newfound moral understanding among the populations in question: the relevant moral and rational considerations were always readily apparent. Instead, English gentlemen stopped dueling because it came to be viewed as vulgar and dishonorable, associated with a loss of respect among their peers. When dueling became ungentlemanly, the gentlemen stopped doing it. Interestingly, Appiah attributes the change in norms surrounding dueling in part to increasing democratization, literacy, and to newly available forums for public moral discourse:

> Newspaper comments and cartoons […] were of crucial significance in the changing response to the duel. The rise of the popular press and of working-class literacy made it increasingly clear – and, as democratic sentiment grew, increasingly unacceptable – that gentlemen were living outside the law. When dueling was an aristocratic practice known mostly only within the class of those who practiced it, there was no place for the attitudes of ordinary people to shape its honor world. The modern press brought all citizens of Britain into a single community of knowledge and evaluation.[17]

The story Appiah tells in *The Honor Code* reflects the fact that dueling was a social norm in Bicchieri's sense. The persistence of dueling did not depend upon individual beliefs about whether or not dueling was a good idea. It depended upon people's empirical and normative

[16] Appiah, 36.
[17] Ibid, 38.

expectations. Accordingly, the end of dueling first required a change in normative expectations, not a change in beliefs surrounding the moral status of dueling. What made this change in normative expectations possible was the emergence of a new venue for public discourse, where gentlemanly duelists were rightly shamed and ridiculed. Appiah goes on to argue that analogous shifts in normative expectations or 'honor codes' precipitated moral revolutions surrounding footbinding in Ming Dynasty China and the abolition of slavery in the British Empire. The overarching lesson aligns with Bicchieri's central prescriptive recommendation: to change social norms, moral arguments are not enough: one must target the social expectations that undergird them.[18]

5. Virtue signaling communicates social expectations

To see how this understanding of social norms is related to virtue signaling, we must consider the kind of information that virtue signalers communicate to their audience. Normally, we think of virtue signaling as an attempt by a speaker to convey information about their character and thereby improve their reputation. According to this common understanding, the primary intention of any act of virtue signaling is to convey information about the speaker. However, this information is not usually conveyed directly, for obvious reasons. Straightforwardly asserting one's superior moral character is commonly viewed as evidence of one's immodesty and would thus prove self-defeating as a way of proving to others that one is virtuous.[19] Instead, the virtue signaler must attempt to convey information about their moral character indirectly, via information about how they think people *ought* to think and act. When a virtue signaler condemns a politician in harsh moral terms for taking donations from the oil industry, or

[18] Bicchieri, *Norms in the Wild: How to Diagnose, Measure, and Change Social Norms*.
[19] G. F. Schueler, "Why Is Modesty a Virtue?," *Ethics* 109, no. 4 (1999): 835–41.

demands that a company fire an executive who made a racist comment, or abstains from consuming meat and dairy in the name of animal rights, they are publicly endorsing a normative standard about the kinds of behaviors that are acceptable or unacceptable. The virtue signaler wants their audience to use this endorsement as evidence for their character. But in the process, they have also told their audience about what they think should be viewed as permissible and impermissible, honorable and shameful. From the perspective of an audience member learning about the social norms of their ingroup, this is valuable information in and of itself. Indeed, when many individuals chime in and make similar public moral statements (*piling on*, as Tosi and Warmke put it), the audience member receives evidence about what *many* people believe ought to be the case. If all these virtue signalers belong to one's reference network, then this starts to look like pretty good evidence for one's normative expectations – i.e. how one's peers think people ought to behave. And if this widespread virtue signaling rises to the level of public shaming or ostracism of the sort that creates real incentives for people to conform to the new norm, this will create evidence for one's empirical expectations about how people will actually behave. And so, whatever else it does for the virtue signalers themselves, *virtue signaling provides evidence about social norms.*

So understood, the potential utility of virtue signaling as a tool for positive norm change becomes clear. If a group of influential virtue signalers can be convinced that publicly committing to some new normative standard will increase their moral reputations, then they stand to play a valuable role in spreading that new norm throughout the broader population. This amounts to a three-step process: first, a group of sincere advocates for change seed a new, positive normative standard into the public discourse; second, virtue signalers eager to appear 'on the side of the angels' broadcast this new standard to a broader audience through a mix of

positive avowals and public shaming; third, a much larger population treats the behavior of these virtue signalers as evidence that they should change their social expectations, and become motivated to conform to the new norm.[20]

Recent changes in attitudes about the environmental impacts of flying and reductions in commercial and private aviation provide us with a case study for how this might work in practice. Long known to be significant source of global carbon emissions,[21] frequent air-travel remains an indicator of social status and a cosmopolitan lifestyle.[22] Recently, climate change activists have attempted to subvert these attitudes by using both traditional and social media to 'flight-shame' frequent flyers and private jet owners, while also encouraging people to display their 'train-pride' when adopting environmentally friendly alternatives.[23] This new moral construal of air-travel has surged into the mainstream in recent years, generating numerous articles in the popular press about flight-shaming,[24] high-profile flight-shaming incidents,[25] and a spike in worldwide Google Search traffic over the past five years.[26] In a number of countries, this movement has also coincided with noticeable decreases in air-travel and corresponding changes in attitudes,[27] and become a source of considerable anxiety for the aviation industry.[28]

---

[20] Notably, the same individuals can play multiple roles in this process: at times, sincere advocates for change might behave more like virtue signalers, once they become aware of the reputational benefits of their actions; meanwhile, virtue signalers might sometimes act as audience members for one another.
[21] Joyce E Penner et al., "IPCC Special Report on Aviation and the Global Atmosphere," *Intergovernmental Panel on Climate Change* (Geneva, 1999).
[22] Lisa Oswald and Andreas Ernst, "Flying in the Face of Climate Change: Quantitative Psychological Approach Examining the Social Drivers of Individual Air Travel," *Journal of Sustainable Tourism* 29, no. 1 (2020): 68–86.
[23] Umair Irfan, "Air Travel Is a Huge Contributor to Climate Change. A New Global Movement Wants You to Be Ashamed to Fly," Vox, 2019.
[24] Ingrid K. Williams, "A Dispatch From the Land of Flight Shaming," *The New York Times*, 2019.
[25] Katie Nicholl, "Harry and Meghan Under Fire After Yet Another Private Jet Flight: 'Frankly It Is Hypocritical,'" Vanity Fair, August 2019.
[26] "Topic: Flight Shame," Google Trends, 2021.
[27] Stefan Gössling, Andreas Humpe, and Thomas Bausch, "Does 'Flight Shame' Affect Social Norms? Changing Perspectives on the Desirability of Air Travel in Germany," *Journal of Cleaner Production* 266 (2020); Elena Berton, "Flight Shaming Hits Air Travel as 'Greta Effect' Takes Off," Reuters, 2019.
[28] Ahmed Hagagy, "Aviation Industry to Counter Flight Shaming Movement: IATA Chief," Reuters, 2019.

Alongside its apparent success as an environmentalist campaign, the 'flight-shaming' movement seems to have shaped the public moral discourse in a very particular way. Before, a jet-setting lifestyle might have been viewed as normal or even glamorous; now, these same habits have become a reputational liability. Meanwhile, broadcasting one's own abstention from aviation – before, a somewhat puzzling form of ascetic self-denial – has become a way for people to enhance their moral reputations in the eyes of their peers. And most importantly, publicly flight-shaming others suddenly became a way for people to burnish their environmentalist bona fides. In effect, the flight-shaming movement seems tailor-made for virtue signaling.

Not all flight-shaming is virtue signaling, of course. Many of its proponents and earliest instigators are no doubt motivated by sincere moral conviction. However, as many philosophers have argued, genuine virtue of this sort is probably rare in the population at large: most ordinary folk are of decidedly uneven moral character, motivated by a mix of egoistic and moral reasons and rarely aiming for more than moral mediocrity.[29] It is likely that a number of the people who subsequently engaged in flight-shaming were motivated at least in part by reputational considerations, which exert subtle, unconscious influence over such acts of moralistic punishment.[30] But for the purposes of changing social norms, this is a good thing. To instill new empirical and normative expectations within a community, a substantial number of people need to publicly commit to the new normative standard and start enforcing it upon others, not just the most virtuous individuals. Given this end, it makes sense for sincere actors to welcome the efforts of virtue signalers in spreading the new norm. By appealing to human beings' deeply

[29] Christian B Miller, *Moral Character: An Empirical Theory* (Oxford: Oxford University Press, 2013); Eric Schwitzgebel, "Aiming for Moral Mediocrity," *Res Philosophica* 96, no. 3 (July 2, 2019): 347–68.
[30] Jillian J. Jordan and David G Rand, "Signaling When No One Is Watching: A Reputation Heuristics Account of Outrage and Punishment In One-Shot Anonymous Interactions," *Journal of Personality and Social Psychology* 118, no. 1 (2020): 57–88.

engrained disposition to enhance their moral reputations,[31] virtue signaling exploits a much more effective channel for transmitting social information throughout the broader population than appeals to pure virtue ever could.

For the virtue signalers themselves, reputational considerations can also offset the potential costs of adopting and committing to new social norms. Adopting new social norms can mean making effortful changes to one's lifestyle, or else giving up something one normally enjoys. And even taking a public stance on an issue can be costly: it risks alienating members of one's social network, and creates the potential for backlash and reprisals.[32] Such costs could reasonably deter anyone from taking a public moral stand or changing their behavior. Potential reputational gains within one's reference network can offset these costs when raw courage does not suffice. Reputational benefits thus provide a motivational boost that can enable an individual to break free from the gravitational pull of the status quo.

6. Virtue signaling and hypocrisy

At this point, a critic might object that status-seeking only motivates the appearance of virtue, and not virtue itself. Because virtue signalers only care about creating the impression that they are morally good, they will be motivated to take actions designed to maintain that impression, rather than the actions that would actually warrant their neighbors' high moral esteem. In short, the motivations underlying virtue signaling seem to incentivize moral hypocrisy.

This objection misconstrues the relationship between virtue signaling and changing moral behavior. The principal utility of virtue signaling, on my account, is not that it motivates *the*

[31] Dan Sperber and Nicolas Baumard, "Moral Reputation: An Evolutionary and Cognitive Perspective," *Mind and Language* 27, no. 5 (2012): 495–518.
[32] Benoît Monin, Pamela J. Sawyer, and Matthew J. Marquez, "The Rejection of Moral Rebels: Resenting Those Who Do the Right Thing," *Journal of Personality and Social Psychology* 95, no. 1 (2008): 76–93.

virtue signalers themselves to adopt new norms (though, as we shall see shortly, it does that as well): it is that widespread virtue signaling creates social expectations among members of the virtue signaler's *audience*, which in turn motivates the adoption of and conformity to social norms; when conformity with those norms aligns with one's moral goals (as in our flight shame example), this creates an incentive for moral behavior.

'Wait,' our critic might respond, 'you said yourself that conformity to social norms depends on both normative and empirical expectations. This means that virtue signalers' behavior really does matter. If virtue signalers do not actually conform to the norms they publicly commit to, their signaling alone will not be enough to sustain people's empirical expectations. So as long as virtue signaling incentivizes hypocrisy rather than genuine moral commitment, it can't help much with establishing new norms.'

It is true that if people do not believe that enough members of their reference network are conforming to a norm, they will not conform to it either. But there are a number of forces operating within the realm of public moral discourse that push virtue signalers towards behaving in line with their claims. Central among these is the charge of hypocrisy itself: if a person very obviously fails to practice what they preach, their moral reputation will inevitably suffer, as people generally do not think highly of hypocrites.[33] We see an elegant example of this phenomenon in the 'This you?' meme that began to appear on social media feeds during the spring of 2020, when the police killing of George Floyd sparked Black Lives Matter protests across the United States and around the world. [34] As expressions of support for the Black Lives Matter movement spread across the internet, some individuals and corporate entities that tried to

[33] Jillian J. Jordan et al., "Why Do We Hate Hypocrites? Evidence for a Theory of False Signaling," *Psychological Science* 28, no. 3 (2017): 356–68.
[34] Aisha Harris, "'This You?' (It Definitely Is)," *The New York Times*, June 9, 2020.

opportunistically virtue signal about their newfound commitment to racial justice were instead met with reminders of their own racist behaviors and policies, along with a simple question: 'This you?' Thus, these hypocritical virtue signalers ultimately achieved the opposite of their intended goal: instead of enhancing their moral reputations, they were publicly called out and held to account. And far from being fooled by their empty gestures, the virtue signalers' intended audience took delight in exposing them as hypocrites.

This is not to say that hypocrisy is always called out in such a public manner, or that individuals do not sometimes get away with empty virtue-signaling. But it does show that hypocritical virtue signaling is a risky social strategy, since it is liable to a form of social sanction. Indeed, this particular form of social sanction – public shaming – can itself be a form of virtue signaling, a way of demonstrating one's own keen moral instincts. Thus, virtue signalers are not just kept in check by their audience, but by each other as well.[35]

It is thus implausible to claim that virtue signalers are unlikely to practice what they preach, since this kind of behavior would negate whatever reputational benefits the virtue signaler had hoped to accrue in the first place. Instead, reputation-seeking motives incentivize morally consistent behavior via an indirect route: a person might initially signal commitment to a norm out of a desire for enhanced moral reputation, but then feel compelled to live up to this norm because their public moral expression has made them publicly accountable.[36]

Notably, while public accountability mechanisms incentivize virtue signalers to practice what they preach, there is no similar incentive for them to *care* about what they preach. Even when

[35] For a similar point, see Levy, "Virtue Signalling Is Virtuous," 9.
[36] Brendan Dill and Stephen Darwall, "Moral Psychology as Accountability," in *Moral Psychology and Human Agency: Philosophical Essays on the Science of Ethics*, ed. Justin D'Arms and Daniel Jacobson (Oxford: Oxford University Press, 2014), 40–83.

virtue signalers' actions are aligned with their moral claims, their underlying attitudes are not, which is its own kind of hypocrisy. Virtue signalers would have others believe that they are motivated by moral beliefs when they are really acting out of reputational concerns. Since this kind of motivation is constitutive of virtue signaling, it would seem to follow that all virtue signalers are hypocrites in this sense.[37]

However, this kind of hypocrisy is in fact consistent with the kind of moral progress that virtue signaling can help us achieve. Because public accountability mechanisms motivate virtue signalers to conform to a new norm – however hypocritically – their behavior will provide evidence for others' empirical expectations, which in turn support the diffusion of new positive social norms. As long as virtue signalers are held accountable, leading them to practice what they preach, they can still contribute to moral progress, even though they are not virtuous themselves.

7. Virtue signaling and epistemic vigilance

These worries about hypocrisy also recall one of Tosi and Warmke's major concerns about the negative effects of widespread virtue signaling: its potential to cause widespread cynicism. On their account, as people come to suspect that many expressions of moral attitudes are in fact disingenuous or self-serving, they will become increasingly skeptical of the moral claims that people raise. And when virtue signalers display excessive amounts of moral outrage, this ends up lowering the evidential value that such displays normally carry; thus, onlookers can no longer use the fact that people are morally outraged about something as a sign that it is a matter of significant moral concern. In short, as virtue signaling spreads, so will broader forms of

[37] Thanks to one of the editors of this journal for raising this point.

skepticism about public moral discourse, which will make it harder for people to learn from one another.

As I have already noted, this argument relies upon an epistemic understanding of how public moral discourse produces moral progress: people make arguments about moral issues, which cause other people to update their moral beliefs and change their behavior. Until now, my argument has focused on highlighting a non-epistemic way that public moral discourse can lead to moral progress, and the role that virtue signaling has to play in this process. In doing so, I have sidestepped concerns about the epistemic downsides of virtue signaling. But now we are in a position to address them.

As in the case of hypocritical virtue signaling, I suggest that false, inaccurate, or otherwise epistemically ill-founded acts of virtue signaling also risk reputational harms. Tosi and Warmke argue that reputation-seeking motives lead virtue signalers to try to outdo one another by 'ramping up' the severity of one's moral claims, each trying to appear more righteous and more outraged than the last. A purported consequence of this dynamic is that contributions to moral discourse become increasingly disconnected from the truth. But this picture ignores the fact that being perceived as reasonable and as a reliable source of information is quite important for one's reputation.[38] Being exposed as epistemically unreliable or incoherent would be just as damaging for one's reputation as being exposed as a moral hypocrite, albeit for different reasons: while the hypocritical virtue signaler undermines their moral credibility, the unreasonable virtue signaler undermines their epistemic credibility. This means that the need to be perceived as reasonable will also place plausibility constraints on the content of a person's virtue signals. As these

[38] Sacha Altay, Anne Sophie Hacquin, and Hugo Mercier, "Why Do so Few People Share Fake News? It Hurts Their Reputation," *New Media and Society*, 2020.

expressions become increasingly extreme or untethered from the communal common ground, they will be less effective in achieving their status-seeking ends.

This idea is reflected in Hugo Mercier and Dan Sperber's argumentative theory of reasoning.[39] According to this theory, one of the primary functions of discursive reasoning is to shore up the reliability of communication and testimony. From an evolutionary perspective, communication offers individuals the opportunity for fitness-enhancing forms of cooperation, but also exposes them to the risk of being exploited by free-riders. To protect ourselves from deception and misinformation, human beings have developed psychological mechanisms for *epistemic vigilance*, which enable us to monitor the coherence of a person's testimony and reasoning, and to keep track of their overall epistemic reliability. These mechanisms kick into action not when we reason in a solitary fashion – in that regard, people are relatively lazy and tend to conserve cognitive resources – but when critically evaluating the arguments of others.[40] This asymmetry in the reasoning process amounts to a form of interactive quality control that enables groups of people to reject bad arguments and converge upon sound ones.[41] In the process, it forces individuals to generate better arguments for their conclusions than they would have on their own.

Thus, the claim that virtue signaling is unconstrained from the truth ignores the fact that it occurs within a discursive ecosystem filled with epistemically vigilant agents, where poor arguments and unfounded claims are subject scrutiny. If social comparison motives push virtue signalers to 'ramp up' and make increasingly unwarranted claims, then the presence of vigilant audiences

[39] Hugo Mercier and Dan Sperber, *The Enigma of Reason* (Cambridge, MA: Harvard University Press, 2017); Hugo Mercier and Dan Sperber, "Why Do Humans Reason? Arguments for an Argumentative Theory," *Behavioral and Brain Sciences*, 2011.
[40] Emmanuel Trouche et al., "The Selective Laziness of Reasoning," *Cognitive Science* 40, no. 8 (2016): 2122–36.
[41] Deanna Kuhn, Victoria Shaw, and Mark Felton, "Effects of Dyadic Interaction on Argumentive Reasoning," *Cognition and Instruction* 15, no. 3 (1997): 287–315; Emmanuel Trouche, Emmanuel Sander, and Hugo Mercier, "Arguments, More than Confidence, Explain the Good Performance of Reasoning Groups," *Journal of Experimental Psychology: General* 143, no. 5 (2014): 1958–71.

serves as a counterweight that forces them to adequately justify those claims. This is not to say that such constraints are airtight: a virtue signaler skilled in rhetoric and sophistry could succeed in spreading unwarranted moral claims in order to enhance their moral reputation. But like moral hypocrisy, this is a risky social strategy. More prudent virtue signalers should be prepared to back up their moral claims with plausible arguments, or else they might be dismissed as fools.

This picture helps to contextualize the cynicism that Tosi and Warmke are worried about: far from undermining the integrity of public discourse, the fact that we are disposed to be cynical about the claims of others actually ensures that the discursive process as a whole remains reliable. A little bit of cynicism keeps virtue signalers on their toes.

8. Virtue signaling and polarization

Another worry about the escalatory dynamics of virtue signaling is that it contributes to polarization. Out of a desire to outdo one's peers in the competition to appear morally pure, this argument goes, virtue signalers on different sides of a debate are liable to adopt increasingly extreme ideological positions, and to engage in ever more severe condemnations of the outgroup, both of which undermine the possibility for political compromise.[42] One way that others have responded to this concern is to point out that there is nothing inherently harmful about this kind of moral polarization, and that it is sometimes entirely appropriate. When there are clear and compelling moral reasons for adopting a position that is 'extreme' relative to the group norm, one should adopt it regardless of its extremity. In these cases, it is not polarization that we should

[42] Tosi and Warmke, *Grandstanding: The Use and Abuse of Moral Talk*, 143–53.

view with suspicion, but rather the assumption that moral truth lies somewhere in the middle ground.[43]

However, the biggest concern about polarization does not hinge on the claim that it is necessarily bad to be polarized, or that compromise is always the path of virtue. It is rather that intergroup polarization exacerbates forms of social bias and motivated reasoning that make us irrational.[44] This argument is consistent with the observation that in certain circumstances, one end of a polarized debate might be morally right. The problem is instead that polarized agents *invariably* believe that their side is correct *regardless* of whether this is really true, and become increasingly entrenched in these beliefs despite all counterevidence. To the extent that a particular group arrives at the truth via this process, it will be due to mere happenstance, and not because they were reasoning in an epistemically sound manner.

This argument against polarization echoes Tosi and Warmke's more general concern about the unreliability of virtue signaling. In the previous section, I argued that virtue signalers are epistemically constrained by the vigilance of their interlocutors and concern for their epistemic reputations. However, these sorts of constraints might not be as effective in contexts of intergroup polarization. After all, virtue signalers are not necessarily worried about their reputations in the population writ large, but rather their reputations within their reference networks. Making dubious claims about members of the outgroup might diminish one's standing in their eyes, but this will not matter as long as it improves one's reputation within the ingroup. And when members of their audience engage in politically motivated reasoning, virtue signalers are unlikely to face much critical resistance in this regard. If polarization makes people less

[43] Levy, "Virtue Signalling Is Virtuous"; Coady, "Philosophy & Public Affairs Discussion at PEA Soup: Justin Tosi and Brandon Warmke's 'Moral Grandstanding.'"
[44] Tosi and Warmke, *Grandstanding: The Use and Abuse of Moral Talk*, 74–76.

critical in their assimilation of information, this will create conditions where the most epistemically harmful forms of virtue signaling might run rampant.

However, there are at least two ways of cashing out this epistemic worry about polarization, one very troubling and one much less so. According to the more troubling version of this picture, intergroup contexts like polarization short-circuit our normal rational safeguards, such as deliberation, seeking out more information, and habits of critical thinking, instead redeploying these processes in the service of rationalizing politically convenient beliefs. If intergroup biases corrupt our best defenses against bad reasoning in this way, then polarization really does put us in a very bad epistemic position, and should make us skeptical about even our most well-thought out moral and political convictions. The less worrisome possibility is that politically motivated reasoning is just an ordinary, run-of-the-mill form of sloppy reasoning that afflicts some people more than others, and is mitigated by deliberation and critical thinking. In this case, polarization might make some people prone to bias, but these biases can be overcome through rational scrutiny. This is far less troubling, because it suggests that the epistemic problems associated with polarization – while real – are entirely manageable, and not a cause for skepticism.

Which of these two pictures is correct? While there is some evidence supporting the more pessimistic view of polarized agents,[45] recent research suggests that on the whole, the more mundane picture is probably closer to the truth.[46] For example, Bago and colleagues recently found that opportunities for deliberation make people better at judging whether or not a headline

[45] Dan M. Kahan, "Ideology, Motivated Reasoning, and Cognitive Reflection," *Judgment and Decision Making* 8, no. 4 (2013): 407–24; Steven A. Sloman and Nathaniel Rabb, "Thought as a Determinant of Political Opinion," *Cognition* 188, no. March (2019): 1–7.

[46] Ben M. Tappin, Gordon Pennycook, and David G. Rand, "Rethinking the Link between Cognitive Sophistication and Politically Motivated Reasoning.," *Journal of Experimental Psychology: General*, 2020; Ben M. Tappin, Gordon Pennycook, and David G. Rand, "Thinking Clearly about Causal Inferences of Politically Motivated Reasoning: Why Paradigmatic Study Designs Often Undermine Causal Inference," *Current Opinion in Behavioral Sciences* 34 (2020): 81–87.

is fake news, regardless of whether or not it is concordant with their political views.[47] Indeed, there is evidence that only a tiny fraction of people actually fall for fake news, partisan biases notwithstanding.[48] Meanwhile, Mosleh and colleagues found that Twitter users who score higher on the Cognitive Reflection Test – a widely used measure of analytic thinking that predicts performance on heuristics and biases tasks[49] – tend to follow and share posts from more reliable sources.[50] The same research group has also found that putative effects of cognitive sophistication on motivated reasoning are better explained as the effects of priors beliefs about the question at hand, a confound ignored by previous studies.[51] All this suggests that deliberation and critical thinking are not, in fact, corrupted by polarization. This is not to say that politically motivated reasoning is not real – just that it does not hobble our best epistemic defenses against false information.

As far as virtue signaling is concerned, the threat of polarization turns out to be a little less disturbing than it first appeared. Polarization does not seem to seriously undermine epistemic vigilance, at least for more reflective individuals. This means that when a virtue signaler makes an implausible but politically convenient claim about the outgroup, whether or not they are believed will depend upon whether their audience is disposed towards critical thinking, not just their partisan allegiance. Virtue signalers spreading such falsehoods cannot count on their most cognitively sophisticated audience members to rationalize their dubious claims. Polarization does

[47] Bence Bago, David G. Rand, and Gordon Pennycook, "Fake News, Fast and Slow: Deliberation Reduces Belief in False (but Not True) News Headlines," *Journal of Experimental Psychology: General* 149, no. 8 (2020): 1608–13.
[48] Altay, Hacquin, and Mercier, "Why Do so Few People Share Fake News? It Hurts Their Reputation"; Nir Grinberg et al., "Fake News on Twitter during the 2016 U.S. Presidential Election," *Science* 363, no. 6425 (2019): 374–78.
[49] Maggie E. Toplak, Richard F. West, and Keith E. Stanovich, "The Cognitive Reflection Test as a Predictor of Performance on Heuristics-and-Biases Tasks," *Memory and Cognition* 39, no. 7 (2011): 1275–89.
[50] Mohen Mosleh et al., "Cognitive Reflection Correlates with Behavior on Twitter," *Nature Communications*, 2021.
[51] Tappin, Pennycook, and Rand, "Rethinking the Link between Cognitive Sophistication and Politically Motivated Reasoning."

not make bad reasoners of us all – it is just one more context where bad reasoners might reveal themselves.

9. From social norms to moral beliefs

One of the premises of my argument has been that changes in moral belief are often not enough to promote moral progress. This claim has rested in part on cases where people persisted in immoral practices despite knowing better, because the practice in question was a social norm (as in the case of dueling). These are cases where moral beliefs might *precede* but do not precipitate changes in moral behavior. Virtue signaling and norm change offer an alternate, social path to achieving these changes. Yet one might worry that the kind of progress achieved by norm change is too shallow and precarious to count as genuine moral progress: while conformity to these norms might bring people's behavior in line with principles of right action, these behaviors are not themselves motivated by moral reasons. In this section, I address this worry by sketching out how prior changes in social norms can precipitate thicker, more robust forms of moral progress.

Consider again the flight-shaming movement. Suppose that this campaign has been successful, and that a flight-shame norm is now in place. People who conform to this norm do not do so because they believe that air travel is wrong. They conform to it because they have the empirical expectation that others in their reference networks do so as well, and because they believe that others in their reference networks think that people ought to conform to it. These are social reasons, not moral ones. Here, virtue signaling has made the world a better place, but it has not led to any improvements in people's moral beliefs.

Once the new norm is in place, however, it can create a set of social conditions where it becomes easier for people to bring their moral beliefs in line with their actions, and their actions in line

with their moral beliefs. To see how this is possible, consider a world (not unlike the real one) where there is no flight-shame norm, but where many people privately believe that air travel has bad moral consequences. As we have seen, it might be very difficult for these people to act on those moral beliefs. They might feel social pressure to go on that vacation to Rome, or to attend a prestigious conference at Oxford. They might want to refuse to go on these trips, but they know that doing so would make them seem excessively moralistic and prudish. They might also miss out on important social and professional opportunities or damage their personal relationships. In this kind of normative environment, acting on their moral beliefs would be quite costly. To follow through on one's moral convictions, one would have to be particularly insensitive to these social costs.[52]

Now let us return to that other imaginary world where the flight-shaming movement has been successful. Here, the costs of acting on one's environmentalist values are greatly reduced. By changing the social norm, we have enabled people to bring their actions in line with their moral beliefs without risking ostracism. In this new normative environment, our reluctant flyers now have the social freedom to act on their moral convictions. Where previously social pressures might have posed a barrier to moral action, now they facilitate it.

Another effect of the new social norm might be that people who were previously resistant to moral arguments against air travel can now become open to them. Before the implementation of the new norm, some people exposed to environmentalist arguments against air travel might have been strongly motivated to dismiss them, either as a way of maintaining their self-image, or out of resistance to the sacrifices that those arguments prescribed. Accordingly, they might have

[52] In *Norms in the Wild*, Bicchieri argues that such individuals, whom she calls 'trendsetters,' can play an important role in the broader diffusion of social norms, since they are able to weather the costs of violating existing normative standards.

found ways to rationalize away any moral qualms about their European conference travel: 'I am just one person,' they might think. 'My actions cannot possibly make a difference.' But in the new normative environment, people who would have previously resisted moral arguments against trans-Atlantic air travel might find themselves more open-minded. The new social norm creates a psychological permission structure for these people to change their moral beliefs.

Finally, consider a child who develops in this new normative environment. This child has never had to change their behavior or give anything up to conform to our new norm, since it is all they have known. Instead, they have grown up in a community where people avoid excessive air travel, where those who do are criticized in moral terms, and where the moral problems with this form of transportation can be openly acknowledged. In short, all the social signals this child receives will tell them that air travel is morally problematic (even if, deep down, the adults are really just conforming to a social norm). This is the kind of developmental context where a child might come to internalize flight shame as a genuine moral value, rather than conform to it out of mere social pressure.[53] Spread out across an entire generation, this developmental process might represent a much deeper and wider shift in moral beliefs – albeit one that has lagged a little behind an earlier, morally superficial shift in social norms. This scenario represents the aspirational goal of any campaign for social change: just as we now intuitively view dueling or foot-binding as morally repugnant, those seeking to massively reduce harmful practices like air travel or meat eating ultimately hope to make those practices seem unthinkably wrong to future generations. But first, they must change the social norms that govern the current generation.

10. Conclusion

[53] For a developmental account of how children come to internalize the norms of their community as objective moral obligations, see Tomasello, "The Moral Psychology of Obligation."

Nobody likes to be accused of virtue signaling, and it is normal to find virtue signaling annoying. But virtue signaling is not evil, and it is not a great threat to the integrity of public moral discourse. It is simply a vehicle for the diffusion of social norms, positive or negative. The worst excesses of virtue signaling imagined by its critics are held in check by the vigilance of the virtue signaler's audience, to whom they are socially accountable. Virtue signaling can also be an invaluable instrument for social change, especially when moral arguments alone have fallen short. Worries about the dangers of virtue signaling are misplaced: we should not see it as a dangerous impediment to moral progress, but as a potential asset.

Acknowledgments: Thanks to Dong An, Kristin Andrews, Jennifer Nagel, Adam Westra, and the editors of *Philosophy and Public Affairs* for their comments on earlier drafts of this paper, and to audiences at the University of Toronto and the Southern Society for Philosophy and Psychology for discussion; special thanks to Brendan de Kennessey for ongoing discussions and feedback throughout this project. This research was supported by the Social Sciences and Humanities Research Council of Canada.