Draft: Please do not quote without permission of one of the authors; comments welcome at ewestra@purdue.edu, drkelly@purdue.edu

To appear in *The Routledge Handbook of Mindshaping,* ed. T. Zawidzki

Last Changes: 9/23/24 EW
Word Count: body ~7200 words; full ~8700 words

Natural Born Jerks? Virtue Signaling and the Social Scaffolding of Human Agency

By Evan Westra and Daniel Kelly

Abstract: In this chapter, we explore a tension between the mindshaping hypothesis and commonsense Western ideas about moral agency and its relation to the social world. To illustrate this tension, we focus on the phenomenon of virtue signaling. We argue that moral intuitions about the perniciousness of virtue signaling reflect an individualistic conception of agency that we call the inside-out ideal. We argue that this ideal fits poorly with the deeply social, interactive, and regulative portrait of human nature revealed by the mindshaping hypothesis. As an alternative to the inside-out ideal, we advocate for a more outside-in conception of agency. We argue that recent work on self control fits with such a conception, and show how it provides the virtue signaler with a path towards genuine moral growth that is scaffolded by mindshaping dynamics connected to reputational incentives and communal support, and should be accepted as perfectly morally legitimate.

Keywords: Mindshaping, virtue signaling, reputation, self-control, ecological control

I. Introduction

According to the mindshaping hypothesis (Mameli, 2001; McGeer, 2007; Zawidzki, 2013), we humans owe many of our advanced capacities for cooperation and culture to our disposition to regulate the attitudes and behaviors of others, and to our susceptibility to being regulated by them in return. The framework famously challenges central orthodoxies of research on social cognition. For example, it holds that in giving voice to their own attitudes, a person is not simply reporting on the invisible, internal workings of their own mind. Rather, the primary orientation of such verbalizations is more outward facing. To say "I believe that P" is to implicitly make a public pledge, a promise to act in ways that someone who believes that P *should* act (Zawidzki, 2013, 2016) It is to commit oneself to behaving in ways that conform to the norms that govern beliefs, and also, in cases where you violate one of those norms, to explaining yourself by offering reasons that would justify your violation. Similarly, when we make folk psychological judgments about others – evaluating their characters and making inferences about their motivations – we don't primarily do so to predict and explain their actions. Rather we are keeping track of their normative obligations, ever vigilant for transgressions. In general, the mindshaping perspective construes actions like expressing our own attitudes and tracking the mental states of others as, first and foremost, acts of *interpersonal regulation* (McGeer, 2007). Engaging in such actions is a way of participating in a complex normative social system aimed at stabilizing coordination and supporting mutual intelligibility (McGeer, 2021).

As a theory of human social cognition, the mindshaping perspective offers a cohesive framework for understanding a number of different psychological capacities, from natural pedagogy (Csibra & Gergely, 2009), to meta-cognition (Zawidzki, 2021), to character judgment (Westra, 2021a), to self-interpretation (Zawidzki 2018), to communication and comparative psychology (Ross 2023). But it also paints a portrait of humanity that can be a bit … uncomfortable. In particular, it draws attention to the myriad ways in which we are all, deep down, incurably vulnerable to public opinion, to peer pressure, and to social influence, constantly striving to manage the impression we make on others. At the same time, we are also all, deep down, nosy busybodies, endlessly fussing over other people's business.

This reading does not cast us in the best light. It may also be a bit uncharitable. More charitably, the mindshaping hypothesis might be read as highlighting the deeply social character of human psychology. By focusing on the mechanisms and dynamics that allow – indeed require – us to be so interdependent and intertwined, it provides a welcome counterweight to overly individualistic styles of theorizing about the human mind. Still, it is difficult to shake the feeling that the mindshaping hypothesis also, thereby, unavoidably, highlights that most human beings are, quite naturally, status-obsessed jerks.

In this chapter, we interrogate some of the intuitions that make mindshaping theorists' portrait of humanity seem so morally unflattering. In section II, we'll argue that these intuitions can be traced to a normative conception of how moral agency should work that we call the *inside-out ideal*. We'll trace out some ways that this ideal animates folk morality, pointing out where it is at odds with the mindshaping framework's depiction of our deeply social nature. We probe this mismatch, arguing that it props up a distorted picture of human sociality. In section III, we'll begin to explore what an alternative picture might look like, one built around an ideal vision of moral agency firmly rooted in the mindshaping perspective. We draw inspiration from recent philosophical thinking about self-control that emphasizes social, situational, and ecological factors rather just internal individualistic ones. We conclude in section IV by reconsidering the inside-out ideal from this alternative perspective.

II. Inside Out Moralizing

Consider a fictional (though hopefully relatable) case:

> **Alice the Aspiring Vegan**: For a long time, Alice has thought about going vegan. It'd be a big change, but she has long found the actual arguments for veganism to be quite persuasive. Lately she's noticed that more and more of her friends are giving up animal products. She's also getting the feeling they're beginning to look down on her for not doing the same. Finally, on the first day of a new year, Alice takes the plunge. On her social media accounts, she earnestly proclaims: "I'm going vegan! After thinking long and hard about it, I've decided I will no longer live my life in a way that perpetuates such cruelty." As her posts pile up hearts and encouraging comments, Alice basks in the warm glow.
>
> But as the glow fades, it is quickly replaced by temptation. While Alice continues to get mileage out of affirming her vegan conversion in public, in private her new diet is not going so well. She slips up repeatedly, giving in guiltily to old carnivorous cravings. It continues until one day a friend catches her halfway through a meat-lovers' pizza. She feels a rush of shame as her friend gives her a pointed, mildly disgusted look.

> The next time Alice is on the verge of succumbing to the siren call of pepperoni, sausage, and mozzarella, she cringes inwardly, vividly recalling her friend's disdain. With mixed emotions – a bit of glum defeat, a bit of resolute self-exhortation – she orders a vegan pizza instead (it's …okay).
>
> Eventually, as she continues to publicly embrace her social identity as a vegan, and after a few more wobbly episodes ordering at restaurants, Alice settles into her new diet. The cravings become less frequent and intense; the temptations recede. She is surprised one day when she realizes that now even the thought of eating meat seems kind of gross. She never slips up again.

In many ways, this case reads as a success story. Alice made a decision to change her lifestyle so that it better aligned with her considered moral principles, and, after a period of struggle, that's what she did. From this perspective, her progression is one of moral agency realized, autonomy expressed, self-governance achieved.

However, many readers will probably be dissatisfied of this reading. For them, Alice's behavior will come off as more than a little distasteful, because from their perspective Alice is *doing it wrong*. Whatever she might say about her moral motivations, it may be hard to shake the suspicion that all those arguments against eating animal products and expressions of compassion for animal suffering weren't really in the motivational driver's seat when she chose to become a vegan. Rather, her behavior seems like it is primarily driven by what Tosi and Warmke have called a *recognition desire* – a desire for social approval and status (Tosi & Warmke, 2016, 2020). If so, then when Alice goes on about her new lifestyle choice on social media, she is guilty of *virtue signaling* (Westra, 2021b). Even when she finally manages to quash her temptations and consistently live up to the ideals she is publicly broadcasting, her new, hard-won dietary habits remain tainted by their origins – by her desire to fit in, her fear of getting caught, and the guilt and shame that help her fully internalize those new ideals. In short, while Alice may have succeeded in getting to a point where her actions express her considered values, and even thereby achieved a modicum of personal moral improvement, according to this second perspective she seems to have done so in a *morally objectionable way*.

Reputations and their Discontents

We think this latter kind of reaction to Alice's story is the tip of an interesting and important iceberg. On our diagnosis, it is symptomatic of a deep ambivalence about human sociality, and especially about the sort of *social influence* that is the lifeblood of mindshaping. On the one hand, it is easy to appreciate the value of our complex social dispositions when we see them as the heart of our unmatched abilities to cooperate, and source of some of our best impulses, towards empathy, sympathy, compassion, and altruism. On the other hand, we often regard our fellow humans' facility with and susceptibility to social influence with suspicion, if not as an outright weakness, an entry point of corruption. This wariness is also present in many Western social norms, which lionize authenticity (Kelly & Morar, 2020; Leuenberger, 2021) and value individuality over conformity (Brownstein et al., 2025, Chapter 6). It is reflected in the vernacular of our common morality, which is full of connotation heavy terms like "peer pressure" and "bandwagon fan" used to denigrate social influence and those who yield to it.

Alice's case is useful exactly because it pulls us in both directions. It is a success story, but it also involves aspects of our social nature that lead us to look askance at the kind of sway we can exert over one another. Our hesitancy about Alice is an instance of our culturally calibrated tendencies to be suspicious of behaviors that might be done for the sake of belonging and social status, or that appear to be motivated by a desire for the esteem of our peers.

This ambivalence is also crystalized in many common Western attitudes towards hypocrisy. Hypocrites are, by their nature, morally ambiguous: when they advocate for one action but do another, they are often at least *half-right* (Dover, 2019). A doctor who publicly discourages smoking but is a closet smoker herself is, arguably, morally better than a doctor who simply smokes and keeps her mouth shut, not saying anything publicly that might help raise awareness of its extreme risks. But generally, Westerners do not tend to have a commensurately nuanced response to the moral ambiguity of hypocrites. Instead, we mostly just see a poser who fails to practice what they preach, and then promptly start scrutinizing the motives behind the preaching. Indeed, whatever the positive effects of a hypocrite's public pronouncements, we are dubious that they come from a place of other-directed benevolence or a sincere altruistic desire to contribute to the public good. Rather, we tend to assume they are driven by self-interest, usually a desire for some kind of reputational gain (Jordan et al., 2017). Moreover, when someone publicly endorses a norm but then violates that same norm in private, we don't *just* see them as being in the wrong for breaking the norm. We see them as morally worse and more deserving of punishment than someone who just violates the norm without saying anything publicly about it at all (Effron et al., 2018).[1]

It gets curiouser. Even when someone acts in a straightforwardly prosocial manner without *any* accompanying pronouncement, others will view that person with distrust, or even outright hostility, if they suspect the person's action was influenced by social incentives like a concern for their own reputation (Raihani & Power, 2021). As a result, such actions are often evaluated less positively when they are made in public view than when they are made anonymously (De Freitas et al., 2019). Sometimes, this leads to rather perverse judgments about prosocial actions. In one study, participants evaluated examples of charitable acts that differed in how effective they were, and in whether or not the actor seemed to be reputationally motivated. Participants favored those cases where they thought the actor was not motivated by reputational concerns, *even when* the act was less effective in bringing about charitable effects (De Freitas et al., 2019). Ironically, the social cost of being seen as overly concerned with "what other people think" ends up creating an alternative set of reputational incentives that reward agents who try to (or appear to try to) conceal their prosocial actions. Deflecting attention from one's altruism becomes an effective way to cultivate a favorable reputation for modesty (Hoffman et al., 2018; Quillien, 2019).

We see a variation on our theme here. Taken together, this pattern of judgments shows another side of our ambivalence about social influence: a default mistrust of the way reputational incentives interact with putatively moral behavior. On a basic level, we value prosocial actions. We also tend to think that those who do good *should* have a good name, that they deserve the reputational rewards they get. But mixed in with this is a stubborn worry that even do-gooders might be doing the right thing for the wrong reasons. We are so paranoid about the possibility that some agents might act prosocially for the sake of reputational rewards that our suspicion bleeds into other areas. It colors our evaluation of people's moral speech, of their public endorsements of altruistic acts and verbal expressions of

[1] Interestingly, this pattern of judgments does not arise in more interdependent, non-Western cultures (Dong et al., 2022; Effron et al., 2018).

prosocial attitudes. It can lead us to approve more highly of forms of prosociality that are "properly motivated" even if they are less effective, and it can even push us to *punish* prosocial actions if they don't appear to have the "right kind" of motive. This, in turn, can have the strange effect of incentivizing prosocial actors to cloak themselves in a self-effacing veil of humility, even if they are anything but humble. All this amounts to another facet of our overall uneasiness with or outright disapproval of moral agency that, whatever its aims or outcomes, isn't "done the right way".

The Inside-Out Ideal

This puzzling pattern of evaluative judgments also reveals something important about what Western folk morality sees as the correct, properly *authentic* way to behave. Call this the *inside-out ideal*. By its lights, exercising one's moral agency the right way requires protecting one's autonomy from the interference of outside social influence. Expressions of genuine moral agency should originate within an agent, from stable psychological properties like their character traits and deeply held moral beliefs. In the simplest cases, these psychological properties issue in a variety of prosocial actions. Those actions are then socially evaluated by others; they become the subject matter of third-party moral judgments made by members of the agent's community. The ideal designates right and wrong ways to talk the talk while walking the walk as well. *Public expressions* of moral conviction are optional. But as we've noted, when they do happen, they aren't only seen as of secondary importance, but are default suspect and worthy of scrutiny. To be acceptable, public avowals of moral principles must also emanate from the *same* internal, stable, psychological properties as properly motivated actions. Third parties evaluate the content of such avowals, but they also pay especially close attention to what seems to motivate them, and whether or not they coincide with the agent's non-verbal actions.

The prominence of the inside-out ideal clarifies why Alice's road to veganism might strike some as objectionable. In its starkest and most caricatured form, the ideal sees social influence as inherently corrupting, and any behaviors motivated by social incentives like reputational concerns as morally tainted.[2] Third-party moral judgments—like those that push Alice to take the plunge of embracing veganism or the one expressed by the friend who catches her scarfing down a meat-lover's pizza—are not supposed to drive her behavior if she is to be a properly autonomous, self-governing moral agent. To be fully authentic, such an agent's core moral attributes should be insulated from the costs and rewards associated with what other people think. When choosing her path, Alice should not have been swayed by public opinion but guided by her inner moral compass alone, and she should have reshaped her dietary habits through her own fortitude and internal willpower. Moral agency ideally flows from the inside-out, never from the outside-in.

We suspect that many who find Alice's moral transformation dissatisfying are at least somewhat sympathetic to the inside-out ideal. We also think the ideal sheds light on why it seems natural to not even give a hypocrite partial credit for talking the talk, and why we scrutinize public moral speech and conspicuously performed prosocial actions so closely for ulterior, reputation-based motives. We are immersed in a culture and commonsense morality deeply informed by the inside-out ideal, so we are primed to be vigilant for violations of it, and of what it portrays as ersatz instances of moral agency.

Against the Inside-Out Ideal

[2] For an expression of the inside-out ideal that comes close to this, see Emerson's enormously influential "Self-Reliance" (1841).

Even from a commonsense point of view, there's something awfully unforgiving about this ideal. The vast majority of us *do* care about how we are perceived by our peers. We *do* pay attention to how the things we do and say will affect the way others perceive us. And we *are* reliably motivated by what others do and think. Moreover, all of this isn't just normal but, according to the mindshaping account of the actual character of human psychology, completely *natural* as well. Our social sensitivities are deeply ingrained and not easily suppressed. The putatively virtuous indifference to status and reputation prescribed by the inside-out ideal is difficult and only rarely achieved (Miller, 2013). The inside-out ideal pits us against our own nature.

Moreover, it's not clear that we should *want* ourselves or others to fully achieve this ideal. Going in one direction, someone who was *completely* oblivious, indifferent, or impervious to the opinion of others would be viscerally alarming. Their behavior, which is likely to come off as irritatingly inconsiderate in the best of times, could easily tip into genuine unkindness and more extreme forms of immorality. Going in the other direction, intentionally foregoing *all* concern for status or what others in your community think of you seems like it would render your life – like the lives of Wolf's moral saints – "strangely barren" (Wolf, 1982). Surely, it is both normal and good for us to value the esteem of our family, friends, and at least some of our peers. Indifference to what the people in our lives think of us seems almost pathological.

The mindshaping perspective also makes clear that fully embracing the inside-out ideal can lead us to throw out some of the baby with the bathwater. Its insistence on internally rooted moral agency obscures the benefits furnished by our sensitivity to opinion and social influence. For example, many researchers think that our propensity to gossip about others, and our directly related concern with what others are saying about us – our concern for our own reputation – has played an important role in the evolution of fairness and cooperation (Baumard et al., 2013; Dunbar, 2004). Gossip serves as an informal channel for the circulation of information crucial for choosing between potential collaborative partners (Feinberg et al., 2012; Sperber & Baumard, 2012; Wu et al., 2016b). In other words, it's no accident that we mindshapers are so obsessed with monitoring and regulating the minds, words, and actions of those around us, or that we are alert and intricately responsive to having ours monitored and regulated in turn. We would be a much less prosocial species, and much worse at cooperation and culture, if we weren't.

In short, we believe that the inside-out ideal is something of a trap. This prominent strand in the fabric of Western social norms and moral sensibilities puts us at war with ourselves. It can blind us to the upside of social influence, and casts normal, natural, and often straightforwardly positive human social motivations as morally worrisome weaknesses that we should strive to overcome. To escape this trap, we will try to distance ourselves from the inside-out ideal. We turn next to developing a few building blocks for a more expansive conception of agency. Such an alternative model will be less internally oriented and individualistic, and so more consistent with our nature as a mindshaping species, giving it the advantage of fitting better with a conception of human flourishing firmly rooted in our extreme interdependence.

III. Beyond the Inside-Out Ideal: Ecological Control and Socially Scaffolded Moral Agency

What would a more expansive, mindshaping-friendly conception of moral agency look like? We find inspiration in recent advances in research on self-control. Traditionally, psychologists working on self-control have understood their subject matter in a way that recalls the inside-out ideal, through an

individualistic lens tightly focused on internal factors. According to one familiar analogy, self-control is like a *mental muscle*: a capacity that can be drawn on to effortfully suppress prepotent desires and impulses, that can become tired after bursts of usage but also strengthened with consistent training (Baumeister et al 1998; Bratslavsky et al., 1998; Muraven et al., 2006). It is no accident that failures of self-control are often described as *weakness* of will. This muscular model also fits well with the phenomenology of impulse control, which is typically experienced as effortful, aversive, and exhausting (Inzlicht et al., 2014).

Self-control from the Outside-In: Socially-Mediated Ecological Control

However, many researchers in this literature have begun to move beyond these narrow notions that see self-control as merely an in-the-moment exertion of mental brawn, and have been exploring a range of other facets of our capacity for self-regulation. Most do not deny we have this kind of internal impulse-suppressing power (Ainslie, 2021; Sripada, 2021), but also countenance a number of other strategies that individuals use to shape their own behavior without having to flex any inhibitory "muscles" (Duckworth et al., 2016). Some of these strategies can still be rather internally oriented. For example, people can try to change the way they deal with objects of temptation. Spotting a delicious looking charcuterie plate at a work event, aspiring vegans like Alice might endeavor to keep their attention trained fully on their conversation partner to avoid staring at all that tasty meat and cheese. Or rather than simply trying to resist the urge to graze, they might aim to undermine it by reappraising the objects of temptation, actively reminding themselves of the disgusting conditions the meat and cheese came from.

Other strategies of self-control seem to work from the *outside-in* rather than the inside-out. Instead of taking situations as given and trying to draw on a reservoir of internally sourced influence over how we think about or respond to them, these strategies use our ability to prospectively select and shape situations themselves. For example, Alice might be proactive and ask her employer beforehand to make sure that there will be vegan options available at the shindig. If that's not happening, she could avoid the temptation of the charcuterie board by avoiding the situation entirely and just skipping the event. Strategies like these can be understood as a form of *ecological control* (Clark, 2007; Setman & Kelly, 2021).

The exercise of this form of self-regulation tends to be more diffuse. Ecological control is not the flexing of an inhibitory muscle at the moment an urge arises, but is rather the result of distributed, decentralized processes, and its supporting elements are typically assembled and arranged over time. A person seeking to shape their own thought and behavior using ecological control can change their physical circumstances. For example, someone trying to get over a tough heartbreak might remove all the reminders of their ex from their apartment and phone. They can opt into and out of situations with an eye towards the kinds of emotional effects those are likely to induce, avoiding places haunted by memories of lost love. Similarly, someone trying to quit drinking can stop frequenting bars; someone trying to quit smoking can cut back on the amount of time they spend around other smokers. In doing so, they aren't using or building up inhibitory muscles, they're removing themselves from situations where they need to use those muscles at all. Indeed, as Pascal pointed out centuries ago, someone who wants to believe in God but harbors recalcitrant doubts would do well to take this kind of tack in reverse, *joining* a religious community, spending time socializing with its members and talking to them about their religious beliefs. In all these cases, people regulate their own behavior and mental lives by exerting control over their own ecology. In many cases, by actively choosing and molding the external circumstances they put themselves in, they are also thereby choosing and molding the kinds

of social influences they are exposing themselves to. They are using ecological control, and enlisting other people to help shape their own minds (Brownstein et al., 2025, chapter 7; Holroyd & Kelly, 2016).

Externally focused strategies like these do not seem to fit well with how self-control is commonly conceptualized. Empirical evidence supports this impression. When asked to generate examples of strategies for self-control, people are more likely to come up with internal, intrapsychic approaches. They also tend to see these more familiar forms of self-control as more effective than external strategies, and also evaluate them as more moral (Bermúdez et al., 2023; Murray et al., 2023). This seems to suggest that people not only see internal forms of self-control as more paradigmatic, but that they are also inclined to see use of external strategies as indicative of some kind of moral weakness or character flaw. This stands in stark contrast with the evidence about the *effectiveness* of ecological control across multiple domains, including those in which people typically struggle to successfully regulate their own behavior, from substance abuse, to studying and academic work, to saving for retirement (Duckworth et al., 2016). It also stands in contrast to findings that suggest individuals who rate themselves as higher in self-control actually tend to employ fewer internal strategies and more external strategies in their daily lives (Ent et al., 2015; Imhoff et al., 2014).

Here we see another variation on our theme. This research about self-control, together with the research on folk psychological beliefs about self-control, solidifies our impression that this is yet another domain where common forms of thought are guided by a heavily internal and individualistic ideal of moral agency. Here again the inside-out ideal appears to be an obstacle to our flourishing, standing in the way of effective self-regulation and successful pursuit of meaningful moral projects. As such, expanding our notion of self-control to include expressions of agency that are not grounded in stable, inner attributes will be both a theoretical and practical step forward. We think it also points the way to an analogous shift we should make, one that can reduce the grip the inside-out ideal holds on how we think about reputationally motivated behaviors like Alice's. Specifically, we propose that Alice's "virtue signaling" is better thought of as an exercise of *socially mediated ecological control.*

Self-Regulation and Virtue Signaling

For those still under the sway of the inside-out ideal, it's easy to interpret Alice's public embrace of veganism solely in terms of the social benefits she hopes to gain, what might seem to be her less-than-noble motives. But this interpretation leaves out something important: by making her moral decision public, she has also makes herself *accountable* to the people around her in new ways. A private commitment to veganism would be quite easy to violate with no one the wiser. Lapses might result in a little self-castigation, but they would not incur any extra social costs. Her public *commitment*, however, opens Alice up to public *opinion*. In talking the talk, Alice proactively summons a specific kind of social influence onto herself, which changes the reputational stakes of different actions she might take going forward. Others will think less of her if they catch her eating meat; they might gossip about, ostracize, or otherwise punish her if they catch her violating her new principles.

With her initial announcement Alice sets into motion a sequence of changes to her environment that will eventually result in changes to her own attitudes and actions. She exerts ecological control over her own behavior, and that control is overtly socially mediated. She is employing a strategy that proactively engineers her social situation. Her public commitment changes how the members of her community think about her, and so will change how they will interact with her, how they will evaluate and regulate her behavior. In doing this, she has reshaped the social incentives she herself faces. She

has drawn elements external to herself, including other people and their reactive attitudes, into the distributed loop she is arranging and using to regulate herself. This is self-regulation, even if the pathway that regulation takes is indirect and extends out beyond her skin and skull before returning, even if the machinery supporting it is external, decentralized, distributed, and social. It is self-control from the outside-in.

Notably, Alice is ultimately regulating her own behavior by first mindshaping other people. From the mindshaping perspective, Alice's initial avowals of her newfound veganism are immediately identifiable as publicly expressed normative commitments. Instead of keeping her intentions to herself, she broadcasts them to members of her social network. Those members change their minds about Alice in response, recalibrating their expectations of her. If Alice were to loudly avow that she is a vegan, and then go on to gorge herself at an all-you-can-eat seafood buffet, her community would correctly criticize her and downgrade their assessment of her reputation. "Virtue signaling" can yield social benefits, of course; but there is a tradeoff. It also opens the signaler up to considerable risk and social costs.

Once Alice makes her commitment public, the only way for her to retain the reputational benefits she has gained by signing on to veganism will be to do what vegans are expected to do: comply with vegan principles (in public, at least – more on this in a moment). But that isn't all. Alice's public commitment will also expose her to a surprisingly wide range of new normative demands, all of which will require Alice to engage in extensive reputation management. She may find that reaping the social rewards of her virtue signaling requires making more changes than she initially realized, as the implications of those principles send ripple effects into domains of her life beyond her dietary choices. Her fashion choices will be differently evaluated. Going out in her beloved leather jacket will no longer be a good look, regardless of how bomb it is. It may raise eyebrows if she refrains from taking a position against medical research on animals. If she finds herself arguing with others about veganism, she'll be expected to have arguments on hand to defend the position, ideally ones that are persuasive to the people around her, and will risk a loss of social standing is she comes up short (Mercier & Sperber, 2011, 2017). To do this well, she may need to change her reading habits to stay up to speed. She may become inclined to change the social circles she runs in, as she seeks out other vegans who hold similar views and whose reasoning she finds convincing, and who are receptive to her own arguments in turn. Perhaps she'll even start expending mental energy rehearsing those arguments to herself, in anticipation of the next time she's called on to defend the cause.

Of course, if Alice were to take a public stand on a less controversial moral issue – say, believing in women's suffrage – she would probably not experience the same kind of reputational management challenges that she confronts going vegan. Because the rightness of women's suffrage is a commonplace moral belief, it's unlikely to require as much change or attract as much public scrutiny. It is also much harder to violate, since women's suffrage is already a widely accepted moral norm and is formally enshrined in federal law. But for these same reasons, if Alice had loudly avowed women's suffrage, she wouldn't have stood to gain anywhere near the sort of reputational benefits she can gain from her embrace of a more *avant-garde* moral position like veganism. When a moral belief is taken for granted, avowing it publicly is unlikely to yield any kind of social reward. (It can even be discomfiting, a way of protesting too much.) In contrast, by signaling her intention to hold herself to a higher and less widespread moral standard, Alice places herself in a position to reap some real social rewards, and also in a position of great normative pressure, scrutiny, and risk.

Here our discussion of self-control and virtue signaling dovetails with our earlier discussion of hypocrisy. When Alice avows her moral beliefs under the watchful gaze of her community, she calls forth a host of skeptics eager to catch her in a misstep. The threat of being exposed as a hypocrite and fraud works in tandem with other motivations Alice has to live up to her public moral commitments (Good & Shaw, 2021; Wu et al., 2016b, 2016a). This in turn casts a new light on aspects of our ordinary moral thinking that seemed somewhat puzzling before. While it may seem strange that people are so hostile towards hypocrites, "do-gooders," and "moral rebels," these attitudes motivate a level of social vigilance that ensures public commitments like Alice's are not simply cheap talk. If everyone around Alice were indifferent to her moral proclamations, nobody would notice if Alice deviated from her moral commitments. But because Alice lives in a community of skeptical busybodies likely to closely scrutinize anyone who would have others view them as morally exceptional, Alice can reliably expect deviations from her public commitments to be both directly and indirectly penalized. The public nature of Alice's action thus rearranges her social ecology, actively creating a social situation that will further scaffold her agency and advance her pursuit of moral self-improvement.

There is a positive side to this story as well. Adopting a socially distinctive set of public moral commitments opens Alice up to criticism if she fails to live up to them, but it also allows her to plug into new sources of social support. By publicly laying claim to a new social identity, she takes a step towards becoming part of a new community. Whereas being a solitary vegan might have been a lonely and difficult path, announcing her veganism enables Alice to become a member of a group structured around vegan norms (Bar-On 2024). In a social environment where veganism is the rule rather than the exception, she is likely to find staying on the straight and narrow path much easier. When she goes out for dinner with her new social group, it will be taken for granted that they'll go to a vegan-friendly restaurant. When one of them throws a dinner party, Alice can rest assured it won't involve the temptation of another charcuterie board. When she overhears a comment mocking her veganism, she knows her new friends will be a sympathetic audience for her gossipy complaints about it later. In short, when Alice selects into a new moral community of mutual mindshapers and like-minded social regulators, she doesn't just reap the rewards of an enhanced reputation. She has chosen to expose herself to *supportive* social influence, and to place herself in social situations in which she will be better able to live up to her moral commitments.

Objections

All this talk of socially mediated ecological control might strike some readers as thin gruel, morally speaking.[3] While reputational considerations may indeed play a role in motivating Alice to adhere to her newfound veganism, this may not strike everyone as an especially praiseworthy moral achievement, let alone one that is likely to be sustainable. Surely, a cynic might complain, a system of incentives based on purely external sources of motivation could not lead to a stable moral change to Alice's behavior. Instead, we should expect Alice to modulate her behavior in a way that would maximize reputational benefits while minimizing personal sacrifices. Maybe she would continue to avow her veganism and abstain from eating meat in public, while continuing to indulge her meat-eating desires in private. When the incentives are absent, the inside-out inspired thought goes, her behaviors will revert to those that express her true, inner, probably carnivorous self. In short, one might worry that moral achievements achieved from the outside in are inherently untrustworthy.

[3] It's still *vegan* gruel, though.

Though there are undoubtedly individuals who fit this cynical description, it is implausible as a broad generalization. For most people, maintaining such a charade would simply be too effortful, the risks of exposure too high. We also mustn't assume that Alice is some kind of sociopath: if the fictional Alice is like most of us, she is motivated by a mix of both egoistic and altruistic concerns (Miller, 2013), and the balance between those will change over time. Surely it is a good idea to be realistic and clear-eyed about her reputational motivations, it's also perfectly plausible that her veganism is also driven by genuine moral reasons. Thus, while she may be more likely to succumb to temptation in private, it is unfair and unfounded to simply assume her veganism is nothing more than a public performance.

Put another way, skeptics might worry that a transformation like Alice's can only be superficial. While she may have reformed her behavior, it is implausible that she will ever be able to acquire the kind of deep, stable moral dispositions required by the inside-out ideal. Because reputational, social motivations played a substantial part in Alice's change, one might harbor the suspicion that these changes don't reflect her "true self" (Strohminger et al., 2017).

We agree it is important to understand the long-term effects that can be brought about by uses of socially mediated ecological control. We are bullish on the prospects of social influence to bring about sustainable changes, though. If Alice keeps deciding to put herself in situations in which she is consistently incentivized to stick to her vegan guns, it's plausible that the initially uncomfortable decisions become increasingly less so over time, until eventually they become habitual, made without effort or notice. This familiar idea traces back to Aristotle: "Men become builders by building houses, harpers by playing on the harp. Similarly, we become just by doing just acts, temperate by doing temperate acts, brave by doing brave acts' (*NE*.1103a32-b2). According to this widely accepted perspective on moral learning, while we cannot simply *will* ourselves to become virtuous, we can acquire virtue by repeatedly engaging in virtuous activities. We do this as children, but we may also continue to do it as adults as well (Sanderse, 2020). Whatever her initial motivations might have been, consistently *doing* vegan things will ultimately give Alice the experience she needs to *internalize* her avowed vegan norms, and adopt veganism as a set of stable, well-entrenched habits.

Conclusion: The Inside-Out-Ideal, Revisited

Let's recap. Alice begins her journey with a mix of moral reasons and reputational motivations to become a vegan. She first acts on both by "virtue signaling" about her new lifestyle choice. In doing so she begins creating a new regime of self-regulation via the exercise of ecological control. Her announcement exposes her to a new kind of social influence, rearranging her social circumstances, reshaping the minds of others in her community so that they react to her in new but predictable ways. These reactions present Alice with new social incentives and reputational stakes, thus engaging her concern for social status, recognition, identity, and belonging (Walton & Brady, 2017). These function to scaffold her motivation to live up to her moral commitment. Her initial avowal sets into motion the creation of a learning environment in which Alice is likely to develop something that was not there beforehand (Kelly, 2020, 2022): habitual, intrinsic motivation to act in accordance with her veganism. Overall, she achieves an increased capacity for effective agency, from the outside-in.

Ironically, although the route there was much different, the final result of this narrative resembles the starting point prescribed by the inside-out ideal: Alice comes to actually possess a stable disposition to act according to her moral principles. The difference, of course, is that she did not arrive there through a brute exercise of will while remaining totally indifferent to the judgments of those around

her. Instead, she harnessed social influence and put it to her own ends, and acted within the parameters of a community governed by mindshaping. She achieved her goal by actively recalibrating an extant social system of normative expectations, punishments, and rewards. This involved the exercise of self-control on her part, but the way she regulated her own behavior occurred against a background – indeed recruited and put to use – her own natural, normal, powerful aversion to criticism and social sanction and craving for social approval and belonging. But far from corrupting or undermining her moral agency, these social motivations and the social influence she brings to bear on them simply reflect the fact that Alice belongs to a species of mindshapers.

Notice, however, that on this retelling of Alice's story, the inside-out ideal has not been fully expunged. On the contrary, it has remained one of the main characters. We think a new interpretation is required to correctly appreciate its significance, however. With Alice's public avowal of her new moral principles, she started out acting *as if* the inside-out ideal was accurate. So too did all the cynical members of her audience who immediately began monitoring her for signs of hypocrisy, whose attitudes Alice carefully monitored in return as she kept tabs on her own reputation. So too did Alice's supporters and allies, who took her nominal moral conversion seriously even as they rewarded her with encouragement, approval, and a sense of belonging. When Alice eventually internalizes her veganism and becomes intrinsically motivated by its associated norms, she is also likely, if she has also internalized the folk psychology of her WEIRD culture, to reify the inside-out ideal herself, updating her narrative identity (Harrelson, 2016; McAdams, 2019; McLean et al., 2020) to depict her conversion to veganism as an act of will. Throughout this process, we can see the inside-out ideal exerting its influence through a bundle of powerful social norms and overly individualistic folk psychological concepts. These play a genuine role in Alice's story. Together they did, in fact, help guide many of the actions and reactive attitudes that scaffold her transformation. As such, we can identify the inside-out ideal as a piece of WEIRD social technology, one that broadly functions as a regulative ideal. However, humans are also, in fact, deeply social, unavoidably interdependent mindshapers. This juxtaposition reveals that the inside-out ideal is a *moral fiction* (Alfano, 2013; Joyce, 2024).

We are not sure what to make of this. It seems to leave us with a paradox about the value of the inside-out ideal. On the one hand, its picture of moral agency is descriptively false and potentially harmful. It seems to deny or denigrate our nature as mindshapers. It obscures and disparages many effective modes of agency and ways to achieve self-governance. Too often, it acts as an obstacle to moral improvement. And yet, even from that same mindshaping perspective, it seems to be serving a critical, perhaps indispensable function. Without a shared commitment to this moral fiction, it is not clear that many of the key social mechanisms that scaffolded Alice's moral journey would have been in place to usher her along.

And so, we once again find ourselves in a place of, if not full on *aporia*, certainly ambivalence. Now though, it is not the corrupting effects of social motivations and influence that give us pause. It is rather the set of individualistic intuitions that previously made us recoil from our own nature as a mindshaping species.

References

Ainslie, G. (2021). Willpower with and without effort. *Behavioral and Brain Sciences*, *44*, e30. https://doi.org/10.1017/S0140525X20000357

Alfano, M. (2013). *Character as Moral Fiction* (p. 238). Cambridge University Press.

Baumard, N., André, J. B., & Sperber, D. (2013). A mutualistic approach to morality: The evolution of fairness by partner choice. *Behavioral and Brain Sciences*, *36*(1), 59–78. https://doi.org/10.1017/S0140525X11002202

Bermúdez, J. P., Murray, S., Chartrand, L., & Barbosa, S. (2023). What's inside is all that counts? The contours of everyday thinking about self-control. *Review of Philosophy and Psychology*, *14*(1), 33–55. https://doi.org/10.1007/s13164-021-00573-2

Bratslavsky, E., Muraven, M., & Tice, D. M. (1998). Ego depletion: Is the active self a limited resource? *Journal of Personality and Social Psychology*, *74*(5), 1252–1265.

Brownstein, M., Madva, A., & Kelly, D. (2025). *Somebody Should Do Something: How Anyone Can Help Create Social Change*. MIT Press.

Clark, A. (2007). Soft Selves and Ecological Control. In D. Ross, D. Spurrett, H. Kincaid, & G. L. Stephens (Eds.), *Distributed Cognition and the Will* (pp. 101–122). The MIT Press. https://doi.org/10.7551/mitpress/7463.003.0007

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences*, *13*(4), 148–153. https://doi.org/10.1016/j.tics.2009.01.005

De Freitas, J., DeScioli, P., Thomas, K. A., & Pinker, S. (2019). Maimonides' ladder: States of mutual knowledge and the perception of charitability. *Journal of Experimental Psychology: General*, *148*(1), 158–173. https://doi.org/10.1037/xge0000507

Dong, M., van Prooijen, J.-W., Wu, S., & van Lange, P. A. M. (2022). Culture, Status, and Hypocrisy: High-Status People Who Don't Practice What They Preach Are Viewed as Worse in the United States Than China. *Social Psychological and Personality Science*, *13*(1), 60–69. https://doi.org/10.1177/1948550621990451

Dover, D. (2019). The walk and the talk. *Philosophical Review*, *128*(4), 387–422. https://doi.org/10.1215/00318108-7697850

Duckworth, A. L., Gendler, T. S., & Gross, J. J. (2016). Situational Strategies for Self-Control. *Perspectives on Psychological Science*, *11*(1), 35–55. https://doi.org/10.1177/1745691615623247

Dunbar, R. I. M. (2004). Gossip in Evolutionary Perspective. *Review of General Psychology*, *8*(2), 100–110. https://doi.org/10.1037/1089-2680.8.2.100

Effron, D. A., Markus, H. R., Jackman, L. M., Muramoto, Y., & Muluk, H. (2018). Hypocrisy and culture: Failing to practice what you preach receives harsher interpersonal reactions in independent (vs. Interdependent) cultures. *Journal of Experimental Social Psychology*, *76*(December 2017), 371–384. https://doi.org/10.1016/j.jesp.2017.12.009

Emerson, R. W. (1841). *Self-Reliance*. CreateSpace Independent Publishing Platform.

Ent, M. R., Baumeister, R. F., & Tice, D. M. (2015). Trait self-control and the avoidance of temptation. *Personality and Individual Differences*, *74*, 12–15. https://doi.org/10.1016/j.paid.2014.09.031

Feinberg, M., Willer, R., Stellar, J., & Keltner, D. (2012). The virtues of gossip: Reputational information sharing as prosocial behavior. *Journal of Personality and Social Psychology*, *102*(5), 1015–1030. https://doi.org/10.1037/a0026650

Good, K., & Shaw, A. (2021). Achieving a good impression: Reputation management and performance goals. *WIREs Cognitive Science*, *12*(4), e1552. https://doi.org/10.1002/wcs.1552

Harrelson, K. J. (2016). Narrative Identity and Diachronic Self-Knowledge. *Journal of the American Philosophical Association*, *2*(1), 164–179. https://doi.org/10.1017/apa.2015.30

Hoffman, M., Hilbe, C., & Nowak, M. A. (2018). The signal-burying game can explain why we obscure positive traits and good deeds. *Nature Human Behaviour*, *2*(6), 397–404. https://doi.org/10.1038/s41562-018-0354-z

Holroyd, J., & Kelly, D. (2016). Implicit bias, character, and control. In J. Webber & A. Masala (Eds.), *From Personality to Virtue* (pp. 106–133). Oxford University Press.

Imhoff, R., Schmidt, A. F., & Gerstenberg, F. (2014). Exploring the Interplay of Trait Self–Control and Ego Depletion: Empirical Evidence for Ironic Effects. *European Journal of Personality*, *28*(5), 413–424. https://doi.org/10.1002/per.1899

Inzlicht, M., Schmeichel, B. J., & Macrae, C. N. (2014). Why self-control seems (but may not be) limited. *Trends in Cognitive Sciences*, *18*(3), 127–133. https://doi.org/10.1016/j.tics.2013.12.009

Jordan, J. J., Sommers, R., Bloom, P., & Rand, D. G. (2017). Why Do We Hate Hypocrites? Evidence for a Theory of False Signaling. *Psychological Science*, *28*(3), 356–368. https://doi.org/10.1177/0956797616685771

Kelly, D. (2020). Internalized norms and intrinsic motivations: Are normative motivations psychologically primitive? *Emotion Researcher*, 36–45.

Kelly, D. (2022). Two Ways to Adopt a Norm: The (Moral?) Psychology of Internalization and Avowal. In M. Vargas & J. M. Doris (Eds.), *The Oxford Handbook of Moral Psychology*. Oxford University Press.

Kelly, D., & Morar, N. (2020). Bioethical ideals, actual practice, and the double life of norms. *The American Journal of Bioethics*, *20*(4), 86–88.

Leuenberger, M. (2021). What is the Point of Being Your True Self? A Genealogy of Essentialist Authenticity. *Philosophy*, *96*(3), 409–431. https://doi.org/10.1017/S0031819121000012

Mameli, M. (2001). Mindreading, Mindshaping, and Evolution. *Biology and Philosophy*, *16*(5), 597–628. https://doi.org/10.1023/A:1012203830990

McAdams, D. P. (2019). "First we invented stories, then they changed us": The evolution of narrative identity. *Evolutionary Studies in Imaginative Culture*, *3*(1), 1–18.

McGeer, V. (2007). The regulative dimension of folk psychology. In D. D. Hutto & M. Ratcliffe (Eds.), *Folk psychology re-assessed* (pp. 137–156). Springer.

McGeer, V. (2021). Enculturating folk psychologists. *Synthese*, *199*(1–2), 1039–1063.

McLean, K. C., Syed, M., & Lowe, L. (2020). Narrative identity in the social world. In *Cambridge Handbook of Personality Psychology*.

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, *34*, 57–111. https://doi.org/10.1017/S0140525X10000968

Mercier, H., & Sperber, D. (2017). *The Enigma of Reason*. Harvard University Press.

Miller, C. B. (2013). *Moral character: An empirical theory*. Oxford University Press.

Muraven, M., Shmueli, D., & Burkley, E. (2006). Conserving self-control strength. *Journal of Personality and Social Psychology*, *91*(3), 524.

Murray, S., Bermúdez, J. P., & De Brigard, F. (2023). Moralization and self-control strategy selection. *Psychonomic Bulletin & Review*, *30*(4), 1586–1595. https://doi.org/10.3758/s13423-023-02257-7

Quillien, T. (2019). Universal modesty in signal-burying games. *Proceedings of the Royal Society B: Biological Sciences*, *286*, 20190985. https://doi.org/10.1098/rspb.2019.0985

Raihani, N. J., & Power, E. A. (2021). No Good Deed Goes Unpunished: The social costs of prosocial behaviour. *Evolutionary Human Sciences*, 1–21. https://doi.org/10.1017/ehs.2021.35

Sanderse, W. (2020). Does Aristotle believe that habituation is only for children? *Journal of Moral Education*, *49*(1), 98–110. https://doi.org/10.1080/03057240.2018.1497952

Setman, S., & Kelly, D. (2021). Socializing willpower: Resolve from the outside in. *Behavioral and Brain Sciences*, *44*.

Sperber, D., & Baumard, N. (2012). Moral Reputation: An Evolutionary and Cognitive Perspective. *Mind and Language*, *27*(5), 495–518. https://doi.org/10.1111/mila.12000
Sripada, C. (2021). The atoms of self-control. *Noûs*, *55*(4), 800–824. https://doi.org/10.1111/nous.12332
Strohminger, N., Knobe, J., & Newman, G. (2017). The True Self: A Psychological Concept Distinct From the Self. *Perspectives on Psychological Science*, *12*(4), 551–560. https://doi.org/10.1177/1745691616689495
Tosi, J., & Warmke, B. (2016). Moral grandstanding. *Philosophy and Public Affairs*, *44*(3), 197–217. https://doi.org/10.1111/papa.12075
Tosi, J., & Warmke, B. (2020). *Grandstanding: The Use and Abuse of Moral Talk*. Oxford University Press.
Walton, G. M., & Brady, S. T. (2017). The Many Questions of Belonging. In A. J. Elliot, C. S. Dweck, & D. S. Yeager (Eds.), *Handbook of Competence and Motivation* (2nd ed., pp. 271–293). Guilford Press.
Westra, E. (2021a). Folk personality psychology: Mindreading and mindshaping in trait attribution. *Synthese*, *198*(9), 8213–8232. https://doi.org/10.1007/s11229-020-02566-7
Westra, E. (2021b). Virtue Signaling and Moral Progress. *Philosophy and Public Affairs*, *9999*. https://doi.org/10.1111/papa.12187
Wolf, S. (1982). Moral Saints. *The Journal of Philosophy*, *79*(8), 419–439. https://doi.org/10.2307/2026228
Wu, J., Balliet, D., & Van Lange, P. A. M. (2016a). Gossip Versus Punishment: The Efficiency of Reputation to Promote and Maintain Cooperation. *Scientific Reports*, *6*(1), 23919–23919. https://doi.org/10.1038/srep23919
Wu, J., Balliet, D., & Van Lange, P. A. M. (2016b). Reputation management: Why and how gossip enhances generosity. *Evolution and Human Behavior*, *37*(3), 193–201. https://doi.org/10.1016/j.evolhumbehav.2015.11.001
Zawidzki, T. W. (2013). *Mindshaping: A New Framework for Understanding Human Social Cognition* (p. 320). MIT Press.
Zawidzki, T. W. (2016). Mindshaping and self-interpretation. In J. Kiverstein (Ed.), *The Routledge Handbook of Philosophy of the Social Mind* (pp. 495–513). Routledge.
Zawidzki, T. W. (2021). A new perspective on the relationship between metacognition and social cognition: Metacognitive concepts as socio-cognitive tools. *Synthese*, *198*(7), 6573–6596. https://doi.org/10.1007/s11229-019-02477-2