JOURNAL *of* ETHICS & SOCIAL PHILOSOPHY

VOLUME XXVII · NUMBER 2
April 2024

ARTICLES

OPEN ACCESS FREE FOR AUTHORS AND READERS ISSN 1559-3061

JOURNAL *of* ETHICS & SOCIAL PHILOSOPHY

VOLUME XXVII · NUMBER 2
April 2024

ARTICLES

JOURNAL *of* ETHICS & SOCIAL PHILOSOPHY
http://www.jesp.org

The *Journal of Ethics and Social Philosophy* (ISSN 1559-3061) is a peer-reviewed online journal in moral, social, political, and legal philosophy. The journal is founded on the principle of publisher-funded open access. There are no publication fees for authors, and public access to articles is free of charge and is available to all readers under the CREATIVE COMMONS ATTRIBUTION-NONCOMMERCIAL-NODERIVATIVES 4.0 license. Funding for the journal has been made possible through the generous commitment of the Division of Arts and Humanities at New York University Abu Dhabi.

The *Journal of Ethics and Social Philosophy* aspires to be the leading venue for the best new work in the fields that it covers, and it is governed by a correspondingly high editorial standard. The journal welcomes submissions of articles in any of these and related fields of research. The journal is interested in work in the history of ethics that bears directly on topics of contemporary interest, but does not consider articles of purely historical interest. It is the view of the editors that the journal's high standard does not preclude publishing work that is critical in nature, provided that it is constructive, well argued, current, and of sufficiently general interest.

Journal of Ethics and Social Philosophy
VOL. 27, NO. 2 · APRIL 2024
https://doi.org/10.26556/jesp.v27i2.2856

IS INTELLECTUAL HUMILITY COMPATIBLE WITH POLITICAL CONVICTION?

Michael Hannon and Ian James Kidd

THERE is a profound lack of respect, tolerance, and empathy in contemporary politics. Within the past few decades, political opponents have steadily grown to dislike, distrust, fear, and loathe each other; moreover, members of polarized groups perceive one another as closed minded, arrogant, and immoral.[1] However, new empirical research suggests that intellectual humility may be useful in bridging political divisions.[2] For this reason, a growing number of psychologists and philosophers maintain that intellectual humility is an antidote to some of democracy's ills.

We are enthusiastic about the potential value of intellectual humility in politics, but we also want to sound a note of caution. In a review of recent work on intellectual humility, Nathan Ballantyne reminds us that intellectual humility may have "dark sides."[3] In this paper, we develop this cautionary point by exploring three ways in which intellectual humility may threaten political conviction. In section 2, we examine how intellectually humble citizens are more likely to encounter diverse political perspectives, which may, in turn, lead to a lack of political engagement. In section 3, we argue that intellectual humility tends to facilitate empathy for political opponents, which may lead to a loss of conviction in one's own views. In section 4, we argue that intellectually humble citizens are better "epistemically calibrated" than other people but also that good epistemic calibration often demands a lack of confidence about political issues. Our argument does not spoil claims about the potential

1 See Tanesini and Lynch, *Polarisation, Arrogance, and Dogmatism*; and Iyengar "The Polarization of American Politics."

2 See Hodge et al., "Political Humility"; Hodge et al., "Political Humility and Forgiveness of a Political Hurt or Offense"; Krumrei-Mancuso and Newman, "Intellectual Humility in the Sociopolitical Domain" and "Sociopolitical Intellectual Humility as a Predictor of Political Attitudes and Behavioral Intentions"; Porter and Schumann, "Intellectual Humility and Openness to the Opposing View"; and Stanley, Sinclair, and Seli, "Intellectual Humility and Perceptions of Political Opponents."

3 Ballantyne, "Recent Work on Intellectual Humility."

benefits of intellectual humility, but it should encourage a more cautious sense of how intellectual humility might function in political life.

Toward the end of the paper, we identify some alternative ways of relating intellectual humility to political conviction. In section 5, we argue that intellectual humility could develop into a form of *political quietism* that is modeled by philosophical conservatives such as Edmund Burke and Michael Oakeshott. We describe three general features of such a quietist stance, each interpretable as a form of intellectual humility: diffidence, reticence, and modesty. The availability of these forms of quietism should complicate our thinking about the relationships between intellectual humility, political conviction, and various political stances.

1. HUMILITY AND POLITICAL CONVICTION

The systematic study of intellectual humility and its roles in politics is relatively young, doubtless buoyed by the emergence of virtue epistemology and increasing concerns about the lack of humility in recent political cultures, at least in the United States and the United Kingdom. The early empirical results, however, are encouraging. There is evidence that intellectual humility is associated with reduced affective polarization; those with higher levels of intellectual humility are less likely to derogate the character, competence, and capabilities of political outgroup members; intellectually humble people are more respectful and tolerant of others; they display greater openness to learning about rival positions; they are more empathetic toward those with whom they disagree; they report more positive experiences when discussing politics and are more likely to engage in political discussions.[4] All this indicates that intellectual humility could play a vital role in politics and public discourse. We can therefore regard intellectual humility as a democratic ideal for citizens. As Michael P. Lynch writes in *Know-It-All Society*, "intellectual humility … is a crucial attitude for inquiry and, I believe, for democracy itself."

At the same time, intellectual humility seems to be in tension with another democratic ideal—namely, *political conviction*. It is widely believed that a

4 Bowes et al., "Intellectual Humility and Between-Party Animus"; Krumrei-Mancuso and Newman, "Intellectual Humility in the Sociopolitical Domain" and "Sociopolitical Intellectual Humility as a Predictor of Political Attitudes and Behavioral Intentions"; Stanley, Sinclair, and Seli, "Intellectual Humility and Perceptions of Political Opponents"; Krumrei-Mancuso and Rouse, "The Development and Validation of the Comprehensive Intellectual Humility Scale"; Leary et al., "Cognitive and Interpersonal Features of Intellectual Humility"; Krumrei-Mancuso, "Intellectual Humility and Prosocial Values"; Porter and Schumann, "Intellectual Humility and Openness to the Opposing View"; and Johnson "Humility and the Toleration of Diverse Ideas."

flourishing democracy requires people with convictions. As Lynch puts it, "an apathetic electorate is an obviously ineffective electorate."[5] Thus, a tension emerges between the need for people to have and to act on the courage of their convictions, on the one hand, while also maintaining appropriate forms and degrees of humility about those convictions, on the other hand. A key worry is that if intellectual humility does lead to a lack of conviction, it could have sinister implications for politics. For example, a citizen who abandons their convictions may become susceptible to bad arguments, misinformation, invidious conspiracy theories, political manipulation, ideological apathy, misological incapacitation, and other politico-epistemic hazards. At the same time, too much conviction can lead to arrogance, dogmatism, dialectical incapacitation, interpersonal frustrations and tensions, and other serious problems.[6]

In the literature on intellectual humility, the standard view is the optimistic one that intellectual humility and political conviction do not conflict. (We will call this "the standard view" in what follows.) Lynch assures us that "intellectual humility is not an opponent of conviction" and "not antithetical to critical political engagement."[7] Similarly, Duncan Pritchard says it is wrong to equate humility with a lack of conviction, since individuals can be both high in intellectual humility and ideological commitment.[8] In psychology, Tenelle Porter and Karina Schumann found that "those higher in intellectual humility did not differ from others in the strength of their political views."[9] Likewise, Elizabeth J. Krumrei-Mancuso and Brian Newman found that "sociopolitical intellectual humility was distinct from political apathy and indifference."[10] In short, intellectual humility does not *require* any unwelcome changes in our political conviction.[11]

By and large, we agree with the standard view. Intellectual humility should not be *equated* with, or defined in terms of, apathy, lack of ideological commitment, or a loss or lack of conviction. An individual can simultaneously

5 Lynch, "Conviction and Humility," 139.

6 See Nadelhoffer et al., "Partisanship, Humility, and Epistemic Polarisation"; McIntyre, "Science Denial, Polarisation, and Arrogance."

7 Lynch, *Know-It-All Society*, 150–51.

8 Pritchard, "Intellectual Humility and the Epistemology of Disagreement."

9 Porter and Schumann, "Intellectual Humility and Openness to the Opposing View." For a similar view, see Hodge et al., "Political Humility and Forgiveness of a Political Hurt or Offense."

10 Krumrei-Mancuso and Newman, "Intellectual Humility in the Sociopolitical Domain."

11 According to Lynch, "intellectual humility … is an attitude that requires *confidence*" (*Know-It-All Society*, 150). For similar claims, see Kidd, "Intellectual Humility, Confidence, and Argumentation"; and Tanesini, *The Mismeasure of the Self*, pt. 2.

hold strong political beliefs and be humble, so there is no constitutive conflict between intellectual humility and political conviction. However, this sanguine claim does not, by itself, deliver a wholly optimistic conclusion. First, previous research on the relationship between intellectual humility and political conviction has ignored empirical and theoretical work indicating that intellectual humility *does* often result in apathy or lack of political conviction. Second, there are different forms or kinds of intellectual humility, which can relate to political conviction in many ways. In what follows, we highlight three ways intellectual humility could threaten conviction in political contexts.

2. ISSUE ONE: EXPOSURE TO DIVERSITY

According to new research in psychology, intellectually humble agents display greater openness to learning about rival positions and more willingness to seek out such information. For example, Porter and Schumann found that people with high intellectual humility tend to expose themselves to a greater proportion of opposing political perspectives and are more open to learning about the opposition's views during imagined disagreements.[12] Intellectual humility is also associated with a willingness to befriend political opponents, including "friending" and "following" on social media. In contrast, those low in intellectual humility are less willing to seek out and seriously consider opposing perspectives, and they are less willing to "friend" and "follow" their political opponents.[13]

Such findings support the claim that intellectual humility is good for democratic politics. Democracy is defective when citizens are insulated in echo chambers that reinforce and amplify their own perspectives, cutting them off from contrary views. Echo chambers lead to dogmatism, segregation, and polarization, which can reflect and reinforce social divisions and antagonisms. Maintaining the health of democracy requires that we foster discussions across lines of political difference.

What happens, though, when people actually interact with those who have different political views? A depressing finding in political science is that citizens who interact in these ways tend to become *less politically engaged*. Exposure to diverse perspectives and the experience of deliberating with people who hold contrary views tends to make citizens *ambivalent* and *apathetic* about politics. In *Hearing the Other Side*, Diana Mutz explores the inherent tension between

12 Porter and Schumann, "Intellectual Humility and Openness to the Opposing View."

13 Stanley, Sinclair, and Seli, "Intellectual Humility and Perceptions of Political Opponents." Of course, one could also follow one's political opponents on social media for other reasons, such as keeping an eye on what they are doing.

promoting a society with enthusiastically participative citizens and promoting one imbued with tolerance and respect for differences of opinion.[14] Drawing on abundant empirical research, she concludes that *participatory democracy* is at odds with *deliberative democracy*. Mutz writes, "although diverse political networks foster a better understanding of multiple perspectives on issues and encourage political tolerance, they *discourage* political participation."[15] In other words, the civic virtue of humility seems to pull against the democratic duty to be politically engaged. While intellectual humility encourages people to seek out diverse perspectives and form crosscutting social networks, it may also foster political ambivalence and apathy.

These results cast doubt on two common assumptions: first, that exposure to differing political views is unquestionably a good thing for democracy, and second, that there is no conflict between intellectual humility and the democratic ideal of an engaged, deliberative public. While the intellectually humble may better live up to the *deliberative* ideals of tolerance, mutual respect, and open-mindedness, this may come at the expense of the *participatory* ideal of voting, lobbying, and other ways of realizing one's political convictions. This casts some doubt on the standard view, defended by psychologists and philosophers, that intellectual humility is neither an "opponent of conviction" nor "antithetical to critical political engagement."[16] As the research above demonstrates, people who seek out diverse perspectives and deliberate with people who hold contrary views—two common symptoms of intellectual humility—tend to become ambivalent and apathetic about politics.

There may be ways to reconcile this tension, albeit with some conceptual work. We suspect the tension may partly be the result of researchers in different fields using similar notions (e.g., "engagement" and "apathy") to pick out different phenomena. For example, it may be that psychologists are characterizing "political engagement" as an openness to learning about the opposition's view, whereas political scientists measure engagement by a willingness to vote, lobby, contribute to a political campaign, and other forms of political activity.[17] Similarly, it may be that psychologists and philosophers take "apathy" to be a

14 Mutz, *Hearing the Other Side*.

15 Mutz, *Hearing the Other Side*, 3.

16 Lynch, *Know-It-All Society*, 150–51.

17 For a case from psychology, see Porter and Schumann, "Intellectual Humility and Openness to the Opposing View." For a case from political science, see Mutz, *Hearing the Other Side*. That said, Krumrei-Mancuso and Newman (in "Intellectual Humility in the Sociopolitical Domain") characterize political engagement in terms of (a) interest in politics, (b) attitudes toward participation in political discussions, and (c) likelihood to report voting in a recent election or otherwise participate in political life.

lack of confidence in belief or not caring about politics, whereas political scientists measure apathy by one's willingness to publicly express and defend one's political views.[18] Such conceptual and terminological differences may cause confusions and mistaken generalizations about what the data show. A lack of shared concepts tends to obscure important differences between the various proposals, making it difficult to know what we have learned so far.

Once we clarify the relevant notions, we may find that intellectual humility and political conviction conflict along some dimensions but not others. For example, Mutz investigates why exposure to diverse perspectives might discourage political participation.[19] She considers the idea that encountering political information that challenges one's views may lead people to be *uncertain of their own positions* and therefore less likely to take political action. If this is correct, then being exposed to diverse perspectives will threaten one's political convictions. But there is an alternative explanation: those embedded in crosscutting social networks may *feel uncomfortable taking sides* in the face of multiple constituencies. Many people dislike heated arguments, intractable debates, and all the consequent emotional strife. These people may avoid politics to maintain interpersonal social harmony. In other words, the first explanation posits an *intrapersonal conflict* about what to believe or support, while the second explanation posits *interpersonal conflict* that threatens social relationships. According to Mutz, political ambivalence is primarily due to interpersonal social concerns. If this is correct, then intellectual humility may not diminish strength in one's political views. One feels just as strongly about some political issue as one did before, but one keeps quiet about it to avoid strife. More research is needed to disentangle these hypotheses and to reconcile the alleged conflict between intellectual humility and participatory democracy.

3. ISSUE TWO: EMPATHY

A second optimistic claim is that intellectual humility also facilitates empathy.[20] Of course, empathy is a complicated concept, and there are many different accounts of what it is. On one popular account, empathy is the set of capacities that enables one person to take on, or share, the perspective of another. This connects rather naturally to common conceptions of intellectual humility.

18 Contrast Krumrei-Mancuso and Newman, "Intellectual Humility in the Sociopolitical Domain"; Porter and Schumann, "Intellectual Humility and Openness to the Opposing View"; and Mutz, *Hearing the Other Side*.

19 Mutz, *Hearing the Other Side*, 102–22.

20 Krumrei-Mancuso, "Intellectual Humility and Prosocial Values"; and Johnson, "Humility and the Toleration of Diverse Ideas."

The humble person is better able to imagine the world from different points of view, distinct from their own, which is an ability abjectly lacking in their arrogant and dogmatic counterparts. Insofar as one can "think" or "feel" one's way into a distinct perspective, one might be less likely to regard those who disagree with them as immoral, stupid, lazy, or dishonest. Instead of characterizing others simplistically or in caricature, intellectually humble agents tend to see the humanity in people on the other side of the political spectrum.[21] If so, intellectual humility fosters more constructive deliberation, cooperation, and the ability to work toward common goals. For these reasons, empathic understanding is deeply important to political life.[22]

Determining whether empathy is compatible with political conviction will ultimately depend on how one conceives of empathy. On some recent accounts, empathy is not perfectly compatible with political conviction. For example, Olivia Bailey has argued that it is difficult for us to sustain empathic representation without regarding that perspective as to some degree appropriate.[23] My realization that I empathize with my colleague's frustration about some issue, for instance, will incline me to think she is right to feel frustrated. According to Bailey, it is not only possible but also *likely* that an empathic attitude toward another's perspective will incline us toward that perspective. If empathy inclines us to see the validity of the other's perspective, then empathy seems in tension with retaining one's conviction. We cannot retain our original convictions if empathy leads us to adjust them in the direction of the perspective and the person with whom we empathize.

We can respond to this worry in different ways. An obvious option is to reject the claim that empathizing with a perspective *P* typically or usually leads one to regard *P* as appropriate. Person *A* could emphasize with person *B* and come to find *B*'s perspective *intelligible* but not appropriate. I could see *why* you are so angry with a colleague—given a history of tension, failing to "gel," and professional conflicts—without also coming to consider your responses *appropriate*. In other words, I may *get* why you acted as you did but not think you *should* have acted as you did. Intelligibility and appropriateness are quite different things. Another option is to reject the conception of empathy as "perspective taking." Drawing on the phenomenological tradition, Matthew Ratcliffe argues that empathy is more akin to a perception-like exploration of a person's perspective

21 This is not unquestionably a good thing. It may be a moral failing to successfully empathize with truly horrific outlooks, even if it is an epistemic achievement of sorts. However, we set such cases aside.

22 Morrell, *Empathy and Democracy*; and Hannon, "Empathetic Understanding and Deliberative Democracy."

23 Bailey, "Empathy and Testimonial Trust."

or world, one that presupposes differences as well as similarities between the experiential worlds of the empathizer and the empathizee.[24] There can be many people with whom we empathize whose experiences we cannot "simulate" in any real sense—those with chronic psychiatric illnesses, intense suffering, or life experiences too radically different from anything we have experienced ourselves.[25] We cannot "take on" or simulate those experiences, but we can *explore* them as one would an unfamiliar place, through sustained, tentative processes sustained by interactions, imagination, and trust. Indeed, this conception of empathy as mutual exploration of the experiential world of another arguably embeds a distinctive kind of intellectual humility that is rooted in recognition of radical differences in the structures and contents of different people's experiences.

Suppose we stick with the account of empathy as perspective taking. We then run into another *prima facie* tension. On the one hand, intellectual humility is vital to political life insofar as it helps us deliberate with those who have different convictions and outlooks. On the other hand, intellectual humility facilitates the development of empathy, which may lead one to agree, to some extent, with the perspective of others, including those with very different political views. Here, we should distinguish between two claims. First, one might argue it is a *constitutive* feature of empathy that we regard the target perspective as to some degree *appropriate*. Adam Smith bound empathy to an appreciation of the "propriety" of others' emotions.[26] Alternatively, one might argue that empathy often *inclines us without necessitating us* toward the perspectives of others. When we empathize with someone, we *sometimes* adjust our beliefs in their direction, but there is no necessity to do so. This is a general psychological claim about human behavior, not a constitutive claim about the nature of empathy, and the extent to which it actually obtains may be shaped by wider contextual factors. Of course, the psychological claim is much weaker. It is possible for empathizers to stay confident in their political convictions, at least in some cases. In contrast, the constitutive claim would challenge the compatibility between empathy and conviction, as it would be impossible to empathize with a person without seeing the validity of their perspective.

To dispel confusion about this issue, one must articulate whether the compatibility of intellectual humility and conviction is to be interpreted as a claim about what is *possible* versus what is *likely*. As a matter of psychological fact, empathy may often lead individuals to lose some degree of confidence in their initial view; but this does not necessarily imply that empathy is incompatible

24 Ratcliffe, "Empathy without Simulation."

25 Carel and Kidd, "Suffering as Transformative Experience."

26 Smith, *The Theory of Moral Sentiments*.

with conviction. Much turns on precisely what one means by "incompatible" and "conviction."[27] Even if empathizers *are* often inclined toward the perspectives of others, this may not be a *necessary* feature of empathy. Even if empathizers always decrease their confidence in belief to some degree, they may still hold their beliefs with a high level of conviction. Moreover, we could consider different conceptions of empathy, such as the explorationist account offered by Ratcliffe. Given these options, we should be more cautious before endorsing the hopeful thought that intellectual humility fosters empathy.

4. Issue Three: Epistemic Calibration

According to many theorists, intellectually humble people are *better epistemically calibrated*, meaning they more accurately assess the plausibility of evidence and arguments, are better at forming beliefs on the basis of the evidence, and have a more accurate sense of their cognitive limits and fallibility.[28] Nancy Snow maintains that humility is a form of self-knowledge of one's limitations.[29] Allan Hazlett says it requires a proper assessment of the epistemic statuses of one's first-order doxastic attitudes.[30] In other words, an intellectually humble agent is disposed to believe responsibly—on the basis of available evidence—and disposed to form largely accurate evaluations of their own epistemic standing.[31] In contrast, those who lack intellectual humility are disposed to bad epistemic conduct: they believe irresponsibly and form inaccurate evaluations of their own epistemic strengths and weaknesses.

What is the relationship between epistemic calibration and confidence in belief? It is often said that awareness of one's epistemic limits is not associated

27 We also suspect that scholars mean different things by "empathy," but we set this point aside.

28 The idea that intellectual humility fundamentally involves what I call "epistemic calibration" is widely defended. See Snow, "Humility"; Church and Barrett, "Intellectual Humility"; Hoyle et al., "Holding Specific Views with Humility"; Leary et al., "Cognitive and Interpersonal Features of Intellectual Humility"; Porter and Schumann, "Intellectual Humility and Openness to the Opposing View"; Bowes et al., "Intellectual Humility and Between-Party Animus"; Krumrei-Mancuso et al., "Links between Intellectual Humility and Acquiring Knowledge." Whitcomb et al. (in "Intellectual Humility") dismiss the idea that intellectual humility consists in a disposition to form proper beliefs about the epistemic statuses of one's beliefs; but for a reply, see Snow, "Intellectual Humility."

29 Snow, "Humility."

30 Hazlett, "Higher-Order Epistemic Attitudes and Intellectual Humility."

31 Thus, intellectual humility does not require us to undervalue our capabilities and ourselves. A person might recognize their accomplishments, skills, talents, etc., and yet still be humble about them.

with less confidence in belief but rather with how one *interacts* with one's beliefs.[32] To be intellectually humble means we need to be *thoughtful* in choosing our convictions—to not be more confident than the evidence supports and to form our beliefs and decisions on the basis of the evidence. It does not require us to give up on the ideas we love or believe in. It simply requires us to reconsider our viewpoint *when warranted*. Duncan Pritchard puts the point this way:

> Intellectual humility . . . is entirely compatible with sticking to one's guns, even in the face of disagreement from those around you. Of course, it is not compatible with *always* sticking to one's guns in light of disagreement, as that would indeed be dogmatism. But in cases where one is *legitimately* confident of one's judgments—where one knows that one has special expertise or knowledge that those around one lacks, say, or where this is simply a topic that one knows one has put a due level of thought into—then having the conviction of one's opinions is entirely compatible with one not being dogmatically or intellectually arrogant.[33]

According to this sort of view, there is no essential tension between confidence in belief and intellectual humility. If we are attentive to the quality of the evidence on which our beliefs are based, and we are properly cognizant of our own limitations in obtaining and evaluating relevant information, then we may be both intellectually humble and have justified confidence in our views.[34]

Our question is whether people typically *are* justified in having much confidence in their political beliefs. We suggest that such confidence is often illegitimate. In particular, we will highlight two epistemic *defeaters* of political belief. These provide reasons to think one's political beliefs are unlikely to be true—or at least to be suspicious of one's ground for them. Assuming that intellectual humility increases a person's willingness and ability to revise a belief or reduce confidence in it when one learns of defeaters, it follows that intellectually humble people are more likely to have relatively low confidence in their political beliefs.

Consider two epistemic defeaters: *complexity* and *partisanship*. Starting with complexity, it is a truism that many, if not all, political issues are vastly epistemically complex.[35] Think of health care, nuclear disarmament, the economy, trade and tariffs, educational policy, international relations, social justice concerns,

32 Deffler Leary, and Hoyle, "Knowing What You Know"; and Krumrei-Mancuso and Rouse, "The Development and Validation of the Comprehensive Intellectual Humility Scale."

33 Pritchard, "Educating for Intellectual Humility and Conviction," 405 (emphasis added).

34 Indeed, our confidence is better justified than those who lack intellectual humility.

35 Lippmann, *Public Opinion*; Rawls, *Political Liberalism*; Somin, *Democracy and Political Ignorance*; and Friedman, *Power without Knowledge*.

taxation, and whatever other issues concern you. Obviously, the citizens of democratic societies tend to disagree very strongly about these issues. Consider disputes about the nature, causes, significance, and appropriate responses to global warming, drugs, poverty and inequality, terrorism, racism, the gender wage gap, criminal behavior, illicit immigration, and so on. All sorts of complex social and political factors influence each of these issues, making it reasonable for any ordinary citizen to have, at best, very little confidence in any belief about the best way to ameliorate them. These issues involve so many stakeholders and affect so many lives that any solution proposed to alleviate them ought to be met with doubt and extreme caution. In the face of such widespread complexity, what ought the intellectually humble person believe?

A truism in epistemology is that the strength of one's belief should derive in large part from the strength of one's epistemic position. If you lack good evidence for your belief, you should not be very confident in it. However, a key element of intellectual humility is an accurate assessment of one's epistemic standing and an ability to acknowledge gaps in one's knowledge.[36] Thus, the intellectually virtuous agent should have little conviction about these complex political topics, instead adopting a low credence or even suspending judgment about the best political decision.[37] While intellectual humility is perfectly compatible with having the courage of one's convictions where that is epistemically appropriate, it may rarely be epistemically appropriate to hold one's beliefs with confidence in the political domain. The complexity of the social world may frequently undermine justified confidence in pursuing one end or policy over another.

Epistemic complexity is not the only defeater of our political views. Another is what we might call *partisanship*. In a recent article, Hrishikesh Joshi points out that many people's political beliefs cluster around two main camps, despite the fact that these issues are rationally orthogonal.[38] In the United States, for example, the ordinary voter's views about abortion, climate change, immigration, gay marriage, minimum wage, gun control, affirmative action, and business regulation are strongly correlated. This raises an epistemic challenge for the politically

36 Tangney, "Humility."

37 Three caveats are needed. First, low confidence would not be required if the humble agent were also a genuine expert. The existence of expert peer disagreement, however, may still warrant a reduction in confidence. Second, a justified lack of confidence need not push one all the way to uncertainty or suspension of belief. But for highly complex issues, it will likely result in a significant reduction in confidence. Third, the intellectually humble are not always or necessarily less confident than those lacking humility. For instance, the intellectually timid are not humble and yet lack confidence in belief.

38 Joshi, "What Are the Chances You're Right about Everything?"

partisan. There is no compelling explanation for why one political side would get things reliably wrong with respect to a wide range of orthogonal issues.[39] Anyone who finds themselves having the beliefs that are typical of one of the clusters of political opinion therefore ought to reduce their confidence in these beliefs, since the fact of clustering provides an epistemic defeater of these beliefs. The orthogonality of these issues makes it likely that one's beliefs are the result of problematic irrelevant influences or a biased subset of evidence. Think of the ways that certain political identities, such as *being a Democrat* or *being a Republican*, tend to impose normative expectations about the positions one holds, independently of one's actual and perhaps highly particular convictions. This puts rational pressure on people to reduce their confidence in political propositions.[40]

Crucially, this epistemic challenge applies to *anyone* whose opinions tend to be clustered in this way. It is not just a problem for the intellectually humble. However, the intellectually humble are disposed to believe responsibly and to form largely accurate evaluations of their own epistemic standing. Indeed, if their intellectual humility is a self-conscious feature of their political and epistemic identity, then those dispositions will be especially important to them. As a result, they will tend to be less confident than unhumble individuals about these political issues.

A justified lack of confidence is not necessarily a bad thing. It is often useful to have insight into one's areas of ignorance, distinguishing what one knows from what one does not know. An intellectually humble individual is deliberative, careful to weigh evidence, and disposed to monitor whether they are jumping to conclusions that exceed the available evidence.[41] But these benefits of intellectual humility are perfectly compatible with the claim that intellectual humility fosters a (warranted) lack of confidence in one's political beliefs. In this regard, we may view it as a threat to political conviction. Moreover, intellectual humility may lead to other problematic consequences. As Joshi points out, strong partisanship may have practical benefits, including "promoting a sense of solidarity and community, facilitating engagement in long-term political projects and commitments, and helping to sustain motivation."[42] Indeed, there can be interpersonal costs to one's attempts to exercise intellectual humility about political issues; for example, an intellectually humble person might

39 Joshi, "What Are the Chances You're Right about Everything?" 43–48.

40 As Joshi writes, "the partisan has higher-order evidence that some of her first-order political beliefs are mistaken (or, alternatively, some of her credences are inaccurate)" ("What Are the Chances You're Right about Everything?" 50).

41 Deffler, Leary, and Hoyle, "Knowing What You Know"; and Leary et al., "Cognitive and Interpersonal Features of Intellectual Humility."

42 Joshi, "What Are the Chances You're Right about Everything?" 54.

note that the arguments of "their" side are weaker than they are being presented, which might provoke ire, or they may note that those on their side are self-servingly ignoring important counterevidence to some favored policy.

Beyond these philosophical worries, there is empirical evidence that intellectual humility moderates belief strength. In two studies, Adam Hodge and colleagues found that political humility was positively related to openness but negatively associated with political commitment.[43] Likewise, Shauna Bowes and her team found that politics-specific intellectual humility is negatively associated with political belief strength and certainty.[44] This makes sense, given that intellectual humility moderates affective polarization and affective polarization is most pronounced in those who hold the strongest political beliefs.[45] Relatedly, previous research has shown that a strong theistic or nontheistic commitment is related to lower levels of intellectual humility.[46] In general, those who score low on intellectual humility tend to express greater certainty in their views than those who score higher.[47] However, the evidence in this area is mixed, and more research is needed.

5. Humility and Quietism

So far, we have discussed three ways in which intellectual humility could threaten one's political convictions. In section 2, we examined the relationship between intellectual humility, exposure to diversity, and political apathy. In section 3, we looked at connections between intellectual humility, empathy, and loss of conviction. In section 4, we argued that intellectually humble citizens are better "epistemically calibrated," but this may result in a justified lack of confidence. Importantly, none of these arguments presume that intellectual humility and political conviction are *necessarily* incompatible. We agree with defenders of the standard view that intellectual humility should not be equated with loss of conviction, apathy, or lack of ideological commitment. Nevertheless, there are reasonable grounds to doubt the optimistic view that intellectual

43 Hodge et al., "Political Humility"; and Hodge et al., "Political Humility and Forgiveness of a Political Hurt or Offense."

44 Bowes et al., "Intellectual Humility and Between-Party Animus." To assess political conviction, they asked participants to indicate "the strength of your political beliefs" on a sliding scale from 0 (not at all strong) to 100 (extremely strong). Hodge et al. ("Political Humility") measured political commitment by the Ideological Obligation Questionnaire.

45 Bougher, "The Correlates of Discord."

46 Hopkin, Hoyle, and Toner, "Intellectual Humility and Reactions to Opinions about Religious Beliefs."

47 Leary et al., "Cognitive and Interpersonal Features of Intellectual Humility."

humility is no threat to political conviction. We close by considering the possibility of forms of political engagement that express kinds of intellectual humility in ways that are *quietist*. One aspect of those forms of quietism is diminished willingness to participate in the kinds of energetic debate that are integral to modern democratic political *ethoi*.

Our main claim is that there are forms of intellectual humility that encourage attitudes and actions integral to forms of *political quietism*, which we understand as a certain *stance* on the political world. We take the idea of a "stance" from Bas van Fraassen, who characterizes it as a set of attitudes, commitments, approaches, and propositional attitudes, such as beliefs, wishes, and hopes.[48] We think there are also political stances. Think of the person with cooperative attitudes toward rivals, who is strongly committed to democratic government, values being epistemically well calibrated, displays empathetic understanding of others, and has a lucid sense of the epistemic complexity of modern political life.

Although this is only a sketch of this political stance, consider two features. First, there are clearly other stances that a person could adopt. Some are less committed to democratic governance, or they are committed for prudential or epistemic rather than principled reasons. Some people do not approach complexity with epistemically arduous exercises of circumspection and diligence. Some people do not place epistemic priority on being well calibrated; others might regard that aim as being in tension with other values, such as trust in inherited tradition or respect for religious authority. Some do not put value on empathy and might even see it as morally dangerous.[49] Some people are extremists or fanatics who abhor moderation, balance, and compromise.[50] All this has implications for how we understand and value the varieties of intellectual humility. If humility requires attitudes such as openness or fallibility, and if debating and empathizing with rivals is a source of humility, then we can appraise stances in terms of their conduciveness to forms of intellectual humility. So, although there is value in studying intellectual humility in isolation, we also need to understand different conceptions of humility within the wider stances a person takes on the political world.[51]

To see this, consider a stance of *political quietism*. It differs from the more active stance common to most contemporary scholars who write about

48 Van Fraassen, *The Empirical Stance*, 47–48. For a discussion of stances as "epistemic policies," see Teller, "What Is a Stance?"

49 Cassam, "The Epistemology of Terrorism and Radicalisation."

50 On extremism, see Cassam, *Extremism*. On fanaticism, see Townsend et al., *The Philosophy of Fanaticism*.

51 Conceptions of intellectual humility can also be rooted in worldviews or metaphysical visions. See Cooper, *The Measure of Things*; and Kidd, "Deep Epistemic Vices."

humility and political life, who typically value, *inter alia*, "engagement" and "participation," debate, interaction with rivals, the expression and discussion of convictions, and ambitious styles of political activity directed toward substantive goals. Indeed, the value placed on such activist stances is a major reason for wanting to reconcile humility and conviction.[52] Moreover, being active in this sense—authentic, ambitious, engaged, committed, passionate—resonates with widespread tendencies within much of modern moral and social culture. However, there are alternative quietist stances with different conceptions of intellectual humility. We want to sketch out three general features of such a quietist stance, each interpretable as a form of intellectual humility. These features are *diffidence*, *reticence*, and *modesty*. Diffidence regulates the potential tensions between our political goals and our commitment to epistemic standards. Reticence concerns our interpersonal politico-epistemic behavior. Modesty is an active sensitivity to the complexity and changeability of the political world and the consequent difficulties of becoming and remaining properly informed and cognizant. Collectively, these features converge in kinds of *political quietist stance*, and to see why, it is worth sketching them more fully.

First, *diffidence*, in the sense of a principled commitment to reserve or cautiousness when it comes to taking on epistemically complex goals or commitments. For a diffident quietist, the epistemic costs of participation are highly salient, as are the high epistemic standards. Confronted with political events or decisions, a diffident quietist wants to do due epistemic diligence and so highlight the dull-sounding procedural epistemic virtues, such as assiduousness, carefulness, thoroughness, and other dispositions that align personal epistemic conduct with the ideal of *epistemic conscientiousness*.[53] If diffidence urges us to go slowly and work diligently, it is set against many of those tendencies that corrupt modern political culture, such as polarization and demonization of our rivals. Michael Oakeshott, for one, recommended diffidence as a means of resisting our tendencies to "attribute to our enemies a homogeneity which in fact they do not possess."[54] In practice, then, a diffident quietist declines many opportunities to engage politically out of a keen recognition that they cannot properly perform due diligence, rather than out of apathy.

52 This is not the only reason, though. Ever since David Hume's castigation of humility as a "monkish virtue," three main criticisms of humility are that it requires ignorance, entails self-abnegation, and leads to paradoxical self-attributions (Hume, *An Enquiry concerning the Principles of Morals*, 258). A classic statement is Julia Driver's "The Virtues of Ignorance."

53 On epistemic conscientiousness, see Montmarquet, *Epistemic Virtue and Doxastic Responsibility*, 23.

54 Oakeshott, *What Is History?* 162. For similar claims about Oakeshott, see Craiutu, *Faces of Moderation*, ch. 5.

Second, *reticence*, a principled reluctance to debate complex issues due to the reticent quietist's appreciation of the enormous epistemic demands of preparing for and performing such debates.[55] A reticent quietist desires broad, deep understanding and is therefore highly resistant to underprepared participation in debates about complex, contentious, and important political topics. Practically, they may confine their discourse to some well-defined set of issues or will demand sufficient time to prepare for debate as a condition of participation. On other matters outside a well-defined area of confidence, they maintain principled silence—an attitude markedly different from those keen to chip in on whatever topics are "hot" or trending at that moment. Moreover, reticence is a guard against what David Hume called "enthusiasm," the overactive energy that shows itself in those "excited by novelty" and "animated by opposition."[56] Such reticence is consistent with voicing and defending positions but in a way set against the temptations to engage in lightning commentaries, "hot takes," rapid judgments, "universal punditry," and other failures of reticence.[57]

A third epistemic feature of political quietism is *modesty* about one's epistemic capacities to attain and maintain a sufficiently detailed, up-to-date, and critically tested knowledge and understanding of political issues. Oakeshott emphasizes the roles of slow, careful "initiation" into traditions of thought, reflection, and sensibility that affords us the capacities for "judgment."[58] Modesty functions to remind us that it is difficult to remain sufficiently informed about a complicated changing world. Understanding is fragile and transient, liable to become outdated, constantly at risk of being undermined by new empirical or conceptual developments, and so on.[59] A modest quietist is alert to these possibilities and so is averse to epistemic overconfidence. Edmund Burke, for one, discerned overconfidence in the taste of many people for radical and rapid reforms of complex institutions and social "arrangements." It is, he argued, very difficult to achieve a perspicacious understanding of the things one wants to reform: the effects of current arrangements are not always

55 See Smith, "On Diffidence" and "The Virtues of Unknowing."

56 Hume, "Of Superstition and Enthusiasm," 48.

57 On universal punditry, see Kitcher, *Science in a Democratic Society*, secs. 34 and 36. A reticent quietist will also honor Michel de Montaigne's advice to "soften and moderate" the typical "rashness" of our speech using qualifiers—such as "perhaps," "I think," "as far as I know" (Montaigne, *Essays*, 1165).

58 Oakeshott, "The Voice of Liberal Learning," 59, 66.

59 This kind of epistemic modesty can be intensified by other dispositions, such as the cynical anticipation that social institutions often operate according to concealed aims and mechanisms, the identification of which requires new and demanding kinds of epistemic work. See Kidd, "Institutional Cynicism and Civic Virtue."

obvious, and they may provide "remoter benefits" invisible to us, such that it is only with "infinite caution" that we should "venture upon pulling down" any complicated edifice."[60]

A quietist stance, then, is characterized by a set of attitudes, commitments, and beliefs that include specific forms of diffidence, reticence, and modesty. Together, these inflect a different conception or style of intellectual humility that we see modeled by philosophical conservatives from Burke to Oakeshott. If this is right, there are other options for those who want to explore the relations of humility and politics. There are forms of political quietism that recommend more diffident and reticent styles of political life and engagement and that emphasize a more modest conception of the breadth and depth of understanding to which individual political agents could seriously aspire. In these politically quietist stances, there are different ways of operationalizing intellectual humility. Here, humility gets hooked into a set of attitudes, commitments, and beliefs that include diffidence and reticence, cautiousness and conscientiousness, acute suspicion of the temptations of subtle forms of overconfidence, and ardent resistance to what Oakeshott called "dauntlessness," the enthusiasm for "plans that involve the transformation of the world" that are epistemically suspect because they are rooted in a "preoccupation with what is large and distant," a tendency he regarded as "the intellectual vice against which we have to guard at the present time."[61]

Whatever one thinks of these stances of political quietism, they offer alternative ways of thinking about how intellectual humility can relate to political life. For many people, a reticent and diffident stance on the world seems a compelling way of coping with the deeply complex and contested character of the social world and the polarized, pugnacious mood of political discourse. For others, such quietism may be pragmatically inadvisable and even deleterious. Some contemporary character epistemologists have argued that the normative status of character traits *as* virtuous or vicious may sometimes be dependent on the social location of the epistemic agent. José Medina uses the concept of a *predicament*, which is the whole dynamic structure of concerns, dangers, obstacles, and resources that structures a person's experience of the social world.[62] One's predicament can determine what sorts of character traits and stances are salient for a person in relation to the project of understanding and coping with the world. If so, the status of diffidence and reticence as virtues—or as

60 Burke, *Reflections on the Revolution in France*, 152.

61 Oakeshott, *What Is History?* 161. For a similar sketch of a politically quietist figure, see McPherson, *The Virtues of Limits*, 120–21.

62 Medina, *The Epistemology of Resistance*.

subvirtues in an intellectual humility cluster—might be contingent.[63] Indeed, that claim can even be extended to traits usually classified as vices; for example, it has been argued that the trait of closedmindedness can function as a virtue for members of marginalized groups living within epistemically hostile environments.[64] Our aim is not to adjudicate these different possibilities but rather to emphasize their existence and urge further study of them. There are other ways to think about how intellectual humility may relate to political conviction, for instance, some of which can inform political stances that have a more quietist character. Exploring such alternative possibilities gives us a richer overview of the connections between the varieties of intellectual humility and the many forms of political conviction.[65]

University of Nottingham
ian.kidd@nottingham.ac.uk
michael.hannon@nottingham.ac.uk

REFERENCES

Bailey, Olivia. "Empathy and Testimonial Trust." *Royal Institute of Philosophy Supplements* 84 (November 2018): 139–60.

Ballantyne, Nathan. "Recent Work on Intellectual Humility: A Philosopher's Perspective." *Journal of Positive Psychology* 18, no. 2 (2023): 200–220.

Battaly, Heather. "Can Closed-Mindedness Be an Intellectual Virtue?" *Royal Institute of Philosophy Supplements* 84 (November 2018): 23–45.

Bougher, Lori D. "The Correlates of Discord: Identity, Issue Alignment, and Political Hostility in Polarized America." *Political Behavior* 39, no. 3 (September 2017): 731–62.

Bowes, Shauna M., Madeline C. Blanchard, Thomas H. Costello, Alan I. Abramowitz, and Scott O. Lilienfeld. "Intellectual Humility and Between-Party Animus: Implications for Affective Polarization in Two Community Samples." *Journal of Research in Personality* 88 (October 2020): 103992.

Bowes, Shauna M., Thomas H. Costello, Caroline Lee, Stacey McElroy-Heltzel, Don E. Davis, and Scott O. Lilienfeld. "Stepping Outside the Echo Chamber: Is Intellectual Humility Associated with Less Political Myside Bias?"

63 See Dillon, "Critical Character Theory"; and Kidd, "Epistemic Corruption and Social Oppression" and "From Vice Epistemology to Critical Character Epistemology."

64 Battaly, "Can Closed-Mindedness Be an Intellectual Virtue?"

65 We are grateful for the constructive and probing comments of two anonymous referees.

Personality and Social Psychology Bulletin 48, no. 1 (January 2022): 150–64.

Burke, Edmund. *Reflections on the Revolution in France, and on the Proceedings in Certain Societies in London Relative to that Event. In a Letter Intended to Have Been Sent to a Gentleman in Paris*. London: J. Dodsley, 1790.

Carel, Havi, and Ian James Kidd. "Suffering as Transformative Experience." In *Philosophy of Suffering: Metaphysics, Value, and Normativity*, edited by David Bains, Michael Brady, and Jennifer Corns, 165–79. Abingdon: Routledge, 2019.

Cassam, Quassim. "The Epistemology of Terrorism and Radicalisation." *Royal Institute of Philosophy Supplements* 84 (November 2018): 187–209.

———. *Extremism: A Philosophical Analysis*. Abingdon, UK: Routledge, 2021.

Church, Ian M., and Justin L. Barrett. "Intellectual Humility." In *Handbook of Humility: Theory, Research, and Applications*, edited by Everett L. Worthington Jr., Don E. Davis, and Joshua N. Hook, 78–91. Abingdon, UK: Routledge, 2016.

Cooper, David E. *The Measure of Things: Humanism, Humility, and Mystery*. Oxford: Oxford University Press, 2007.

Craiutu, Aurelian. *Faces of Moderation: The Art of Balance in an Age of Extremes*. Philadelphia: University of Pennsylvania Press, 2018.

Deffler, Samantha A., Mark R. Leary, and Rick H. Hoyle. "Knowing What You Know: Intellectual Humility and Judgments of Recognition Memory." *Personality and Individual Differences* 96 (July 2016): 255–59.

Dillon, Robin. "Critical Character Theory: Towards a Feminist Theory of 'Virtue' (and 'Vice')." In *Out from the Shadows: Analytical Feminist Contributions to Traditional Philosophy*, edited by and Sharon L. Crasnow and Anita M. Superson, 83–114. Oxford: Oxford University Press, 2012.

Driver, Julia. "The Virtues of Ignorance." *Journal of Philosophy* 86, no. 7 (July 1989): 373–84.

Friedman, Jeffrey. *Power without Knowledge*. Oxford: Oxford University Press, 2019.

Hannon, Michael. "Empathetic Understanding and Deliberative Democracy." *Philosophy and Phenomenological Research* 101, no. 3 (November 2020): 591–611.

Hazlett, Allan. "Higher-Order Epistemic Attitudes and Intellectual Humility." *Episteme* 9, no. 3 (September 2012): 205–23.

Hodge, Adam S., Joshua N. Hook, Daryl R. Van Tongeren, Don E. Davis, and Stacey E. McElroy-Heltzel. "Political Humility: Engaging Others with Different Political Perspectives." *Journal of Positive Psychology* 16, no. 4 (2021): 526–35.

Hodge, Adam S., David K. Mosher, Cameron W. Davis, Laura E. Captari, Joshua

N. Hook, Don E. Davis, and Daryl R. Van Tongeren. "Political Humility and Forgiveness of a Political Hurt or Offense." *Journal of Psychology and Theology* 48, no. 2 (June 2020): 142–53.

Hopkin, Cameron R., Rick H. Hoyle, and Kaitlin Toner. "Intellectual Humility and Reactions to Opinions about Religious Beliefs." *Journal of Psychology and Theology* 42, no. 1 (March 2014): 50–61.

Hoyle, Rick H., Erin K. Davisson, Kate J. Diebels, and Mark R. Leary. "Holding Specific Views with Humility: Conceptualization and Measurement of Specific Intellectual Humility." *Personality and Individual Differences* 97 (July 2016): 165–72.

Hume, David. *An Enquiry concerning the Principles of Morals*. Edited by Geoffrey Sayre-McCord. Indianapolis: Hackett, 2006.

———. "Of Superstition and Enthusiasm." In *Political Essays*, 46–50. Cambridge: Cambridge University Press, 1994.

Iyengar, Shanto. "The Polarization of American Politics." In *The Routledge Handbook of Political Epistemology*, edited by Michael Hannon and Jeroen de Ridder, 90–100. London: Routledge, 2021.

Johnson, C. R. "Humility and the Toleration of Diverse Ideas." In *The Routledge Handbook of Philosophy of Humility*, edited by Mark Alfano, Michael P. Lynch, and Alessandra Tanesini, 148–56. Abingdon, UK: Routledge, 2020.

Joshi, Hrishikesh. "What Are the Chances You're Right about Everything? An Epistemic Challenge for Modern Partisanship." *Politics, Philosophy and Economics* 19, no. 1 (February 2020): 36–61.

Kidd, Ian James. "Deep Epistemic Vices." *Journal of Philosophical Research* 43 (2018): 43–67.

———. "Epistemic Corruption and Social Oppression." In *Vice Epistemology*, edited by Ian James Kidd, Heather Battaly, and Quassim Cassam, 69–86. Abingdon: Routledge, 2021.

———. "From Vice Epistemology to Critical Character Epistemology." In *Social Virtue Epistemology*, edited by Mark Alfano, Colin Klein, and Jeroen de Ridder, 84–102. Abingdon: Routledge, 2022.

———. "Institutional Cynicism and Civic Virtue." In *The Epistemology of Democracy*, edited by Hana Samaržija and Quassim Cassam, 152–69. New York: Routledge, 2023.

———. "Intellectual Humility, Confidence, and Argumentation." *Topoi* 35, no. 2 (October 2016): 395–402.

Kitcher, Philip. *Science in a Democratic Society*. New York: Prometheus, 2011.

Krumrei-Mancuso, Elizabeth J. "Intellectual Humility and Prosocial Values: Direct and Mediated Effects." *Journal of Positive Psychology* 12, no. 1 (2017): 13–28.

Krumrei-Mancuso, Elizabeth J., and Brian Newman. "Intellectual Humility in the Sociopolitical Domain." *Self and Identity* 19, no. 8 (2020): 989–1016.

———. "Sociopolitical Intellectual Humility as a Predictor of Political Attitudes and Behavioral Intentions." *Journal of Social and Political Psychology* 9, no. 1 (February 2021): 52–68.

Krumrei-Mancuso, Elizabeth J., Megan C. Haggard, Jordan P. LaBouff, and Wade C. Rowatt. "Links between Intellectual Humility and Acquiring Knowledge." *Journal of Positive Psychology* 15, no. 2 (2020): 155–70.

Krumrei-Mancuso, Elizabeth J., and Steven V. Rouse. "The Development and Validation of the Comprehensive Intellectual Humility Scale." *Journal of Personality Assessment* 98, no. 2 (2016): 209–21.

Leary, Mark R., Kate J. Diebels, Erin K. Davisson, Katrina P. Jongman-Sereno, Jennifer C. Isherwood, Kaitlin T. Raimi, Samantha A. Deffler, and Rick H. Hoyle. "Cognitive and Interpersonal Features of Intellectual Humility." *Personality and Social Psychology Bulletin* 43, no. 6 (June 2017): 793–813.

Lippmann, Walter. *Public Opinion*. New York: The Free Press, 1922.

Lynch, Michael P. "Conviction and Humility." In *The Routledge Handbook of Philosophy of Humility*, edited by Mark Alfano, Michael P. Lynch, and Alessandra Tanesini, 139–47. Abingdon, UK: Routledge, 2020.

———. *Know-It-All Society: Truth and Arrogance in Political Culture*. New York: Liveright Publishing, 2019.

McIntyre, Lee. "Science Denial, Polarisation, and Arrogance." In Tanesini and Lynch, *Polarisation, Arrogance, and Dogmatism*, 193–211.

McPherson, David. *The Virtues of Limits*. Oxford: Oxford University Press, 2021.

Medina, José. *The Epistemology of Resistance: Gender and Racial Oppression, Epistemic Injustice, and Resistant Imaginations*. Oxford: Oxford University Press, 2012.

Montaigne, Michel de. *Essays: A Selection*. Edited and translated by M. A. Screech. London: Penguin, 1991.

Montmarquet, James. "Epistemic Virtue and Doxastic Responsibility." *American Philosophical Quarterly* 29, no. 4 (October 1992): 331–41.

Morrell, Michael E. *Empathy and Democracy: Feeling, Thinking, and Deliberation*. University Park: The Pennsylvania State University Press, 2010.

Mutz, Diana C. *Hearing the Other Side: Deliberative versus Participatory Democracy*. Cambridge: Cambridge University Press, 2006.

Nadelhoffer, Thomas, Rose Graves, Gus Skorburg, Mark Leary, and Walter Sinnott-Armstrong. "Partisanship, Humility, and Epistemic Polarisation." In Tanesini and Lynch, *Polarisation, Arrogance, and Dogmatism*, 175–92.

Oakeshott, Michael. "The Voice of Liberal Learning." In *Michael Oakeshott on Education*, edited by Timothy Fuller, 62–104. New Haven: Yale University

Press, 1989.
———. *What Is History? and Other Essays*, edited by Luke O'Sullivan. Exeter: Imprint Academic, 2004.
Porter, Tenelle, and Karina Schumann. "Intellectual Humility and Openness to the Opposing View." *Self and Identity* 17, no. 2 (2018): 139–62.
Pritchard, Duncan. "Educating for Intellectual Humility and Conviction." *Journal of Philosophy of Education* 54, no. 2 (April 2020): 398–409.
———. "Intellectual Humility and the Epistemology of Disagreement." *Synthese* 198, no. 7 (April 2021): 1711–23.
Ratcliffe, Matthew. "Empathy without Simulation." In *Imagination and Social Perspectives: Approaches from Phenomenology and Psychopathology*, edited by Michela Summa, Thomas Fuchs, and Luca Vanzago, 199–220. London: Routledge, 2017.
Rawls, John. *Political Liberalism*. New York: Columbia University Press, 1993.
Smith, Adam. *The Theory of Moral Sentiments*, edited by D. D. Raphael and A. L. Macfie. Indianapolis: Liberty Classics, 1982.
Smith, Richard. "On Diffidence: The Moral Psychology of Self-Belief." *Journal of Philosophy of Education* 40, no. 1 (February 2006): 51–62.
———. "The Virtues of Unknowing." *Journal of Philosophy of Education* 50, no. 2 (May 2016): 272–84.
Snow, Nancy E. "Humility." *Journal of Value Inquiry* 29 (1995): 203.
———. "Intellectual Humility." In *The Routledge Handbook of Virtue Epistemology*, edited by Heather Battaly. London: Routledge, 2019.
Somin, Ilya. *Democracy and Political Ignorance: Why Smaller Government Is Smarter*. 2nd ed. Stanford: Stanford University Press, 2016.
Stanley, Matthew L., Alyssa H. Sinclair, and Paul Seli. "Intellectual Humility and Perceptions of Political Opponents." *Journal of Personality* 88, no. 6 (December 2020): 1196–216.
Tanesini, Alessandra. *The Mismeasure of the Self: A Study in Vice Epistemology*. Oxford: Oxford University Press, 2021.
Tanesini, Alessandra, and Michael Lynch. *Polarisation, Arrogance, and Dogmatism: Philosophical Perspectives*. New York: Routledge, 2021.
Tangney, Julie. "Humility." In *The Oxford Handbook of Positive Psychology*, edited by Shane J. Lopez and C. R. Snyder, 483–90. Oxford: Oxford University Press, 2009.
Teller, Paul. "What Is a Stance?" *Philosophical Studies* 121, no. 2 (November 2004): 159–70.
Townsend, Leo, Ruth Rebecca Tietjen, Hans Bernhard Schmid, and Michael Staudigl, eds. *The Philosophy of Fanaticism: Epistemic, Affective, and Political Dimensions*. New York: Routledge, 2022.

Van Fraassen, Bas C. *The Empirical Stance*. New Haven: Yale University Press, 2002.

Whitcomb, Dennis, Heather Battaly, Jason Baehr, and Daniel Howard-Snyder. "Intellectual Humility: Owning Our Limitations." *Philosophy and Phenomenological Research* 94, no. 3 (May 2017): 509–39.

Journal of Ethics and Social Philosophy
VOL. 27, NO. 2 · APRIL 2024
https://doi.org/10.26556/jesp.v27i2.2788

ADDING INSULT TO INJURY

IS CENSORSHIP INSULTING?

Sebastien Bishop

THE GOVERNMENT bans the praise of terrorist attacks because it is worried that such praise will inspire further terrorist acts. The government censors arguments in favor of holding racist, sexist, and otherwise discriminatory views for fear that the dissemination of such views will promote discrimination and fortify preexisting prejudices. In a bid to get more parents vaccinating their children the government introduces restrictions on the publication of misleading anti-vaccination propaganda.

Certain restrictions on speech are accepted without much controversy. Virtually all agree, for instance, that the government may restrict speech acts that threaten to cause imminent and clear harm, e.g., true threats, blackmail, speech that violates a nondisclosure agreement, speech uttered at such a high volume that it will burst the eardrums of those who walk by. But cases of what we might call harmful advocacy—examples of which are listed in the opening paragraph—are more complicated. These cases involve the government restricting the expression of "corrupting" arguments that it reasonably fears might persuade citizens to think and act in harmful ways. Strong free speech supporters insist that citizens should be free to engage in and hear harmful advocacy, arguing that restrictions are deeply objectionable at best, and at worst wholly impermissible.

To support their position, strong free-speech supporters have offered a wide range of arguments and ideas. One of the most interesting arguments revolves around the idea that restrictions on harmful advocacy (henceforth simply "censorship") are deeply insulting to citizens. The worry, broadly understood, is that censorship fails to properly respect or recognize the intellectual capacities of citizens. As such, even when censorship is effective in preventing harms to citizens, it nonetheless comes at the significant cost of failing to properly respect the citizenry at large. By contrast, so the thought goes, an alternative political scheme that allows for no censorship or permits censorship only in exceptional cases does a better job of respecting citizens as independent, rational, morally responsible agents. This alternative political system may be less effective at

preventing speech harms, but it is at least one where citizens can hold their heads high.

One response to the above is to concede that while censorship is insulting, in at least some cases it is nonetheless all-things-considered justified. This is a promising response. Those pursuing this response will argue that the censorship of dangerous speech helps prevent significant harms. Not only may these harms be life changing, but they tend to be distributed unevenly across society, usually falling on the heads of already marginalized groups.[1] Preventing these harms and the unfairness associated with them should be weighed against (and often in fact outweighs) whatever moral bad is involved in disrespecting citizens. One might also point out that while censorship may be insulting, often this censorship curbs speech that is itself insulting. Thus, one might challenge the view that censorship's insulting nature renders it impermissible by pointing out that even insulting censorship may be the lesser of two evils.

This paper offers a different but complementary line of response—that censorship is not in fact insulting in the ways that have been suggested, or that at any rate the insult involved in censorship has been exaggerated. As this paper argues, critics of censorship have been too quick in assuming that censorship is an afront to the intellectual capacities of citizens. Instead, we would do well to reflect on the various ways in which censorship may be framed as a way to take those intellectual capacities seriously.

To this point, the paper considers and rejects three versions of the worry that censorship is insulting. Section 1 explores the idea that censorship is insulting *qua* involving a negative appraisal of the citizens being interfered with. The key idea here is that censorship involves a lack of what Stephen Darwall terms "appraisal respect," insofar as the government is suggesting that citizens cannot be trusted to manage their own beliefs and intentions. Drawing on the work of Thomas Nagel, section 2 explores the idea that censorship diminishes the political status of citizens. Finally, section 3 explores the suggestion that censorship is incompatible with a full appreciation of the thinking nature of citizens, and thus involves a lack of what Darwall would term "recognition respect."

Ultimately the paper argues that the worry that censorship is insulting has been overstated. The best kind of censorship stems from an appreciation of the diverse needs of citizens, as well as the need for cooperation if societal flourishing is going to be achieved on a large scale. Granted, such a vision involves acknowledging the imperfections and liabilities of citizens—at least when compared to the rather solitary, highly intellectual creature one sometimes

1 Matsuda, *Words That Wound*; Brison "The Autonomy Defense of Free Speech"; Brown, *Hate Speech Law*.

finds in the philosophical literature. Still, such a vision of citizens as imperfect falls well short of being genuinely insulting. To err is human. And there is nothing insulting about being told that you are human.

1. APPRAISAL RESPECT

Perhaps the most straightforward way in which some form of conduct or expression can be insulting is in involving a negative appraisal of another agent. For instance, I might insult my neighbor by suggesting that his moral character is lacking in some way, or that his athletic abilities are subpar. One way of framing such insults is to say that they involve a lack of what Darwall calls "appraisal respect."[2] On this view, I insult and show a lack of respect for another person when I appraise their actions, character, capabilities, etc., and find them wanting. Can this help shed light on the suggestion that censorship is insulting?

It is probably fair, albeit a little blunt, to say that the government's decision to engage in censorship implies that it believes its citizens to be wanting in some respect. After all, the government that engages in censorship does so because it fears that at least some citizens, left to their own devices, and left free to hear all available arguments, will end up developing harmful beliefs and intentions that place others in danger. In this way, the government judges that citizens cannot be altogether trusted to manage their own beliefs and intentions by themselves and would benefit from state interference. What is more, it is easy to see why some might find this insinuation to be insulting. Certainly, many take pride in being able to decide for themselves what is right and wrong, as well as more generally what kind of beliefs and intentions are worth having.

At this point it will be useful to quickly canvass and reject a potential objection a reader might have to this suggestion. The government, so the objection goes, frequently and seemingly without much controversy suggests that its citizens lack competency. After all, traffic safety laws, contract laws, food-packaging laws, and almost any other law one cares to mention all imply that citizens sometimes need help from the government and would likely come up short if left to their own devices. Along these lines, it is tempting to conclude that censorship is no more insulting than the most mundane and uncontroversial forms of government regulation.

But we should be wary of trying to draw an analogy between censorship and more run-of-the-mill regulation cases. To see why, we need only remind ourselves that agents take special pride in managing certain areas of their lives compared to others. For instance, most agents are more likely to be insulted

2 Darwall, "Two Kinds of Respect."

by the suggestion that they cannot adequately take care of their own children than they are by the suggestion that they cannot drive safely without the government's help. Similarly, it is generally more insulting for someone to suggest that you cannot look after your own health than for them to suggest that you sometimes come up short reading the fine print on a contract. Along these lines, those who take censorship to betray a special lack of appraisal respect for citizens will likely claim that judging for one's self (in light of all available arguments) what is morally right and wrong, or which ideas are true and false, is an especially important part of what it means to operate as a mature and responsible autonomous agent. Insinuations of incompetence and attempts to take over the management of this area are thus revealed to be especially insulting in a way that distinguishes censorship from more mundane government interferences with our lives.

This point will serve to preserve the suggestion that censorship is at least *prima facie* distinctively insulting to autonomous agents. What distinguishes censorship as a special case of insulting government interference, so the argument goes, is that it involves the government suggesting that citizens cannot be trusted to manage an area of their lives so fundamental to what it means to function as a responsible and autonomous agent, as deciding for one's self what to think. Have we, then, arrived at a persuasive account of why censorship is so objectionable? I think the answer is no. Here is why.

Whether the suggestion that John is falling short in some endeavor counts as insulting depends on how we understand the challenges involved with that endeavor, and whether we can reasonably expect errors to be made.[3] Take the case where John's PhD supervisor reflects upon the work he has produced and points to places where that work could be improved. We can imagine that the supervisor has marked in red pen dozens of places where John's work could be better. But even granting that the supervisor is making a negative appraisal of John and his work, this negative appraisal need not strike us as insulting so long as it takes place in the context of a recognition of the challenges and expected difficulties involved in completing a PhD. If John's supervisor knows that producing a PhD thesis is difficult, and that making mistakes is just part and parcel

3 Or at least, it partly depends on how we understand the challenges involved. All the same, I think we should agree that maliciously intentioned or callous suggestions that John is falling short may count as insulting. If a stranger repeatedly tells John that he is out of shape and unhealthy, just because the stranger wishes to cause John distress, then this plausibly will count as insulting. So malicious suggestions about how someone is falling short in some respect may well be insulting. But the kind of censorship we are discussing does not involve this kind of malice. At most, it involves the government thinking that its citizens are failing in some respect and that government interference is required.

of completing a PhD, then there need not be anything wounding or insulting when he points to places where John's work might be improved.

Now let us apply this idea to the case of censorship. The government might view the development of misguided (and even harmful) beliefs and intentions as just part and parcel of human reasoning and to be expected when we have a free flow of different arguments and ideas. Censorship then, rather than stemming from a supercilious or insulting vision of citizens as incompetent, might instead be rooted in a view that acknowledges the difficulties involved in coming to develop the correct (or at least harmless) kind of beliefs and intentions. In other words, the censorious government may merely be acknowledging the imperfect nature of human reasoning, and the likelihood that without censorship citizens will sometimes come to develop harmful beliefs and intentions.[4]

Such a view would be well-justified. For instance, no citizen can claim to be a perfect reasoner or a flawlessly rational being. Indeed, a good deal of research has gone into showing how our decisions about what to believe and how to act are often, without our realizing it, influenced by a number of biases and prejudices outside of our control—confirmation bias, framing effects, groupthink, projection bias, self-serving bias, and anchoring bias, for instance, are all well-known biases that impact upon our decision-making in a way that would seem to undermine the rationality and control we have in these areas.[5] Moreover, sometimes the nuances and technical difficulties associated with a subject can lead us to make errors about what to believe and how to act. The point is made clear when we reflect on disputes concerning immigration, the connections between religious groups and terrorism, climate change, vaccinations, and the threat posed by our political adversaries. All of these topics involve various complexities, and disputes often turn on the more technical aspects of these topics—disagreements about what the empirical data tells us, what kind of data should be used, and the extent to which we can rely on "experts" to help fill in gaps in our own understanding. What is more, charismatic speakers are often adept at exploiting these complexities and our own lack of expertise. In short then, even ostensibly reasonable citizens, reflecting upon the various arguments and ideas they are exposed to, can come to mistaken and harmful conclusions.

Even leaving aside the issues surrounding our reasoning biases and the technical difficulties associated with certain subjects, arriving at the right beliefs and intentions can be made considerably more difficult simply by the emotional

4 De Marneffe, "Avoiding Paternalism."

5 Caputo, "A Literature Review of Cognitive Biases in Negotiation Processes"; Murata, Nakamura, and Karwowski, "Influence of Cognitive Biases in Distorting Decision Making and Leading to Critical Unfavorable Incidents"; Loewenstein, O'Donoghue, and Rabin, "Projection Bias in Predicting Utility."

context in which we find ourselves. Consider recent calls to censor particularly exploitative kinds of "seed faith" appeals. Seed faith appeals involve religious leaders encouraging their followers to donate (sometimes very large) sums of money to the church, while assuring their followers that these donations will be rewarded by God in this lifetime. Sometimes these appeals prey on highly vulnerable people. For instance, religious leaders have been known to convince seriously ill believers to donate thousands of dollars to the church, rather than spending that money on medical treatment, convincing them that they stand a better chance of being cured by God than by a doctor. Calls to ban these kinds of seed money appeals are sometimes framed as insulting. But there is nothing insulting about recognizing that those suffering from potentially deadly medical conditions might be vulnerable to exploitation, or that those in desperate situations are prone to making desperate decisions.

These points all tell in favor of a relatively simple thesis: there need not be anything insulting in the government's suggestion that its citizens occasionally need help developing the right kind of beliefs and intentions. Granted, the government that engages in censorship implies that some of its citizens are in danger of erring in some way. But, as indicated above, to err is human, and there is nothing insulting about being told you are human.[6]

2. NAGEL AND STATUS

Nagel considers the potential insultingness of censorship from a rather different perspective. In his 1995 essay "Personal Rights and Public Space," Nagel considers whether this kind of censorship threatens the political status of citizens.[7] In particular Nagel is concerned with attempts by the government to curb speech that it judges likely to reinforce sexist, racist, homophobic, and other kinds of identity-based prejudices.[8] Drawing on the work of Frances

6 Granted, there are other kinds of worries one might have with government censorship. For instance, one might worry that, even if censorship is not necessarily insulting, nonetheless it is likely to be ineffective in practice. One might likewise worry that governments cannot be trusted to regulate speech, or that acceptable censorship today will lead via a slippery slope to unacceptable censorship tomorrow. In response to these worries, I note two points. First, these worries are separate to a worry about whether censorship is insulting. Second, these are worries about certain cases of censorship—censorship that is ineffective, censorship engaged in by untrustworthy governments, censorship that will lead to worse censorship in the future—and as such do not tell decisively against censorship *per se*, or make sense of the principled worry many register toward censorship. What these concerns leave unexplained, in other words, is why we might object even to effective censorship.

7 Nagel, "Personal Rights and Public Space."

8 Nagel, "Personal Rights and Public Space," 96.

Kamm and Warren Quinn, Nagel suggests that such censorious interferences render citizens intellectually violable in a way that we have compelling reason to reject.[9] How does his argument work?

To understand Nagel's thoughts on freedom of speech and censorship, one must reflect upon his larger approach to normativity and human status. Nagel claims that the inherent moral status of persons determines what kinds of freedoms those persons ought to enjoy, as well as placing limits on what interferences with those persons are permissible. When it comes to assessing the government's interferences with our lives, Nagel is interested in how the moral status of citizens determines what kind of burdens the government may impose upon them and what kind of justificatory reasons they may appeal to when imposing these burdens. Nagel's core idea here is that beings that enjoy a higher moral status have powerful claims against certain kinds of interferences, even though these interferences might be acceptable when dealing with beings of a lower moral status. For instance, perhaps certain kinds of animals can be permissibly killed for food or hunted for sport, but human beings—possessing an elevated moral status—cannot be used and abused in this way. Or consider the case where we must decide whether to kill one person in order to save five (different) people from being killed. On a straightforward consequentialist analysis of this case, we ought to kill the one in order to save the many. After all, why protect just one person from being killed when you can protect five? But Nagel demurs, viewing persons as enjoying the kind of elevated moral status that means they may not be compelled to sacrifice their life in the name of the greater good. In other words, Nagel tells us that persons enjoy an inherent "inviolability" possessed by higher moral beings.

At the foundation of Nagel's work lies a vision of persons as belonging to a special category of creature. On this view, unlike say insects or livestock, persons may not be compelled to sacrifice life and limb in the name of promoting the greater good. This is because persons enjoy an altogether higher moral status that grants them normative immunity from making these sacrifices. Nagel's suggestions here speak to an almost sublime, quasi-religious understanding of persons as belonging to an order of moral significance that reaches beyond the material confines of this world, and that grounds a seemingly undefeatable moral claim against being sacrificed in this manner. Nagel's suggestions here also enjoy a certain plausibility. As a matter of intuition, for instance, I imagine that virtually all of us would agree that animals can sometimes be sacrificed in ways that persons cannot; diseased livestock, for instance, may be regrettably slaughtered to protect the rest of the herd from becoming infected, in a way that would be utterly

9 Kamm, *Rights, Duties, and Status*; Quinn, "Actions, Intentions, and Consequences."

unacceptable were we dealing with people. Nagel's message is that our intuitions here reflect a deeper insight about what it means to be a person.

The next key move in Nagel's argument is to argue that these reflections upon the moral status of persons ought to inform how the government treats its citizens—and, crucially, what kind of interferences with those citizens it is willing to engage in. In particular the government ought to grant citizens the kind of political status (that is, the kinds of political rights and liberties) that they are owed as beings possessing an elevated moral status. When it comes to the government that allows its citizens to be sacrificed, the main problem is not so much that this deals a material blow to the freedoms of its citizens (although this might well be problematic to some degree). Rather, what has centrally gone wrong is that the government has failed to fully recognize the elevated moral status of the persons they rule. As Nagel puts it:

> What is good about the public recognition of such a status is that it gives people the sense that their inviolability is appropriately recognized. Naturally they're gratified by this, but the gratification is due to recognition of the value of the status, rather than the opposite—i.e., the status does not get its value from the gratification it produces. . . . It may be that we get the full value of inviolability only if we are aware of it and it is recognized by others, but the awareness and the recognition must be of something real.[10]

So that is a rough sketch of the foundation of Nagel's account. But what has this got to do with censorship? The answer is that, having reflected upon the elevated moral status of persons and the kinds of claims that this establishes, Nagel now shifts his focus toward our status as independent, thinking beings and what kinds of claims this establishes. While our general moral status grants us inviolability with regard to the sacrifice of life, Nagel argues that our status as independent thinking beings grants us a kind of intellectual inviolability that rules out censorship.

> That the expression of what one thinks and feels should be overwhelmingly one's own business, subject to restriction only when clearly necessary to prevent serious harms distinct from the expression itself, is a condition of being an independent thinking being. It is a form of moral recognition that you have a mind of your own. . . . The sovereignty of each person's reason over his own beliefs and values requires that he be permitted to express them, expose them to the reactions of others, and defend them against objections. It also requires that he not be protected

10 Nagel, "Personal Rights and Public Space," 93.

> against exposure to views or arguments that might influence him in ways others deem pernicious, but that he have the responsibility to make up his own mind about whether to accept or reject them.[11]

For Nagel, the problem with censorship is a problem of recognition and status. The government that establishes a stringent legal right against censorship treats its citizens in line with their (supposed) metaphysical status as independent thinking beings. By steadfastly ruling out censorship even in those cases where it might prevent harms or benefit the populous, the government recognizes that its citizens possess, by nature, a powerful claim against others telling them what to think. Again, there is the hint of the sublime present in this conception of persons as having a deep-seated claim to sovereignty over their own minds, even when material considerations (e.g., harm prevention, social welfare) tell against this. By contrast, the political system that permits the government to engage in censorship bestows on citizens a lower political status.[12]

Some readers may balk at this suggestion. In particular they might argue that, even if we agree with Nagel that our status as independent thinking beings means that what we think and feel should be overwhelmingly our own business, this does not mean that the public expression of our thoughts and feelings is simply our own business. But note that this objection slightly misreads Nagel—or at least the version of Nagel I am discussing. When Nagel talks about our status as thinking beings, he is not primarily referring to our ability to express

11 Nagel, "Personal Rights and Public Space," 96.

12 One preliminary worry with Nagel's account is that there is something overblown about all this talk of the government undermining the status of its citizens as higher moral beings. After all, the government routinely controls and restricts the choices of citizens. Consider the uncontroversial restrictions on murder, speeding, stealing, etc., that the government imposes. If these restrictions do not degrade our status, then why worry that restrictions on speech (assuming those restrictions are similarly effective in preventing harm) pose a threat to our status?

As I read him, Nagel's response is that our laws against murder, speeding, stealing, etc., only interfere with our physical autonomy (our ability to act). These laws do not interfere with our intellectual autonomy (our ability to think for ourselves). It is the way that censorious laws seek to subvert our intellectual autonomy, and make certain ideas unthinkable, that marks these laws out as especially troubling. I suspect that many readers will intuitively agree with Nagel here. It is one thing, so the thought goes, to restrict John's freedom to steal Joan's apple—perhaps through placing Joan's apple behind a locked door or punishing John for stealing the apple. But it is another thing entirely to manipulate what ideas and arguments John has access to such that he never even gets the chance to consider stealing Joan's apple. At the very least, I suspect that many will agree with the basic intuition that the latter case involves an interference with a more fundamental and private part of John's person. However, as I argue in the main text, one can concede all of this to Nagel and still dispute that censorious interferences are significantly degrading.

ourselves to others, but rather our ability to listen to what others have to say and then judge their arguments and ideas for ourselves. The problem with a censorious political system on this listener-based account is that citizens are stripped of the ability to judge arguments for themselves in order to prevent the spread of dangerous ideas. As such, citizens within this political system—all of them—belong to a lower echelon of person that only sometimes gets to judge arguments for themselves, since they may apparently be stripped of this power whenever the material considerations call for it.[13]

The implications of all of this are subtle: even if we agree, *arguendo*, that censorship is sometimes justified in order to prevent serious harms from befalling innocent persons (and to be clear, Nagel is reluctant to even admit this much), the Nagelian picture still holds that permitting censorship nonetheless comes at the significant cost of demeaning our political status within that society. A democratic system where the government routinely engages in censorship may well be a safer place to reside. But we who reside in this society will no longer be quite the same sublime thinking creatures we sometimes imagine ourselves to be.

Before canvassing some of my worries with all of this, let me say that there is much to like about the Nagelian picture. Nagel begins by suggesting that many, including himself, register a deep intuitive unease with censorious interferences.

13 Note that, according to Nagel, censorship undermines the equal status of all citizens. With this in mind, Nagel frames free speech as a matter of protecting the equal status of all. Some readers will be understandably skeptical of Nagel here. In particular they will point out that embracing strong free speech rights and refusing to engage in censorship will affect different groups in rather different ways. Granted, perhaps in some sense all groups will benefit from having their higher moral status affirmed. But this higher moral status may involve an increase in speech-related harms throughout society. And crucially, these speech harms will not be distributed evenly throughout society. On the contrary, the harms that flow from (e.g.) hate speech tend to rather predictably fall on the heads of certain already marginalized groups, while other groups are left untouched. The worry, then, is that for all his talk of taking the equal status of persons seriously, Nagel is rather overlooking how certain groups will have to bear the brunt of his free speech policies. Worse still, by ignoring this, Nagel may even be guilty of unfairly prioritizing those groups who are least likely to be negatively affected by free speech policies, thus creating a new problem to do with the status of citizens in society.

While I think this is a reasonable worry, I will not discuss it much further. Nagel is likely to respond to this worry by insisting that what really matters when we are designing our speech policies is how these policies affect our higher moral status as thinking beings. That is to say, that our status as higher moral beings trumps considerations to do with harm prevention. Some will understandably balk at Nagel's prioritization of our alleged higher moral status. However, I want to undermine Nagel's argument via a slightly different route—that even if we grant Nagel's assumptions about the importance of our higher moral status, this does not ground an argument against censorship.

His account is in an attempt to unpack and situate these intuitions. Notice that Nagel deploys a striking argumentative strategy here, reflecting not so much on the interference itself (e.g., its harmful or disruptive qualities), but rather on the nature of the persons being interfered with. Of course, Nagel's analysis of persons has implications for how we understand things like the forcible sacrifice of life in the name of saving others, and censorship. But his starting point is definitively the moral qualities of the persons being interfered with in these cases.

Still, I want to discuss a fundamental problem with Nagel's account. Once we take a closer look at the mechanics of Nagel's objection to censorship, we see that his key argumentative move—that censorship is bad because it bestows on citizens a violable political status—can be interpreted in two subtly different ways. Unfortunately for Nagel, neither interpretation is promising.

The key move in Nagel's argument against permitting censorship is his suggestion that censorship objectionably bestows on citizens the political status of intellectually violable beings—i.e., beings who may permissibly have their intellectual freedom violated. But how precisely should we understand Nagel's complaint here? The 1995 paper touches upon at least two potential readings. The first reading construes Nagel as objecting to the way censorship bestows on citizens a lower political status than they might otherwise have enjoyed. That is to say, given that citizens might have otherwise enjoyed the political status (and concomitant liberties and legal rights) of intellectually inviolable beings, we have reason to lament the government's decision to permit censorship and bestow on citizens the lower political status of beings that may have their intellectual autonomy undermined. On the first reading then, it is the comparative loss of status involved with permitting censorship that is objectionable.

The main problem with this first reading is that, even granting that permitting censorship changes our political status, it is doubtful that this change involves a significant loss of status. Two points in particular are worth emphasizing here. First, if there is a loss of status involved in the move from intellectual inviolability to intellectual violability, it is likely a subtle one. The government that engages in precisely worded, narrowly framed speech regulation need not suppose that their citizens may have their intellectual autonomy undermined willy-nilly. On the contrary, they may hold that citizens generally have a strong claim to exercise their intellectual autonomy. It is just that this government also supposes that there are certain select cases where the harms involved are such that interferences with the intellectual autonomy of citizens, while regrettable, are nonetheless all-things-considered justified. Whatever else we might say about this change then, it is a subtle one, and thus may not support the powerful objection to censorship that the likes of Nagel wish to establish.

Nagel might reply that when it comes to losing inviolability, there are no subtle changes. Here the thought is that losing inviolability always involves a substantial loss. But this kind of argument works best when made in the context of our having an inviolable claim against torture, or an inviolable claim against being made to sacrifice our life for others, or other such uncontroversially dehumanizing interferences. This argument functions less well when made in the context of whether we can sometimes have our access to dangerous persuasive arguments partly blocked. Construing agents as violable in this way does not seem to strike the same demonstrable blow to the status of citizens as, say, permitting torture might. Indeed, we can strengthen this point by reflecting on how even the likes of Nagel and other strong free-speech supporters will admit of some acceptable cases of interference with our intellectual autonomy, e.g., in the form of suppressing incitement to imminent violence.

Second, while in one respect censorship may lower the political status of citizens, in another respect it raises their political status. In his response to Kamm's suggestion that the status of all persons would be degraded were the government to permit persons to be killed in order to save the lives of others, Shelly Kagan argued that a decreased level of inviolability simultaneously secures for each agent an increased level of "saveability."[14] In this sense, so the response goes, permitting persons to be sacrificed does not necessarily diminish or degrade the status of citizens. For whatever loss of status is experienced as a result of being treated as a being that can sometimes be sacrificed may be made up for by being treated as a citizen with an increased claim to being saved.

An analogous argument can be made concerning censorship. Permitting censorship may well treat citizens as intellectually violable. But this decreased inviolability secures an increase in one's state protection and the claims one has against others engaging with arguments that one might be endangered by. The point here is not a consequentialist one to do with balancing the benefits and burdens of censorship. It is that bestowing on citizens powerful legal claims against others endangering them speaks to a certain way of valuing those citizens that enhances their status within the political community. Those citizens subject to censorship may possess fewer political liberties in one sense, but in exchange they enjoy greater legal protection. Take the example of the regulation of (e.g., racist and sexist) hate speech. This regulation, while diminishing the status of citizens in one respect, simultaneously bolsters the protection citizens have against the harms that flow from hate speech, and in this way

14 Kagan, "Replies to My Critics."

bolsters the status of citizens as beings worth protecting. So understood, censorship both diminishes and enhances our political status in various respects.[15]

The implications of all of this for our discussion of whether censorship is insulting are subtle and worth drawing out carefully. One thought is that the above implies that censorship is in no way whatsoever harmful to the status of citizens. After all, even if in some respect the status of citizens is lowered due to censorship, that same status is simultaneously raised by the protection afforded by the said censorship. The net result of this, one might think, is that the status of citizens is unchanged. Some readers may find this a little hard to swallow. One worry is that I am mistakenly assuming that the losses and gains in status brought about by censorship cancel each other out. Some readers will reject this canceling-out model and insist that, even if censorship raises our status in some respects, it nonetheless lowers our status in other respects. With this in mind perhaps the better lesson to draw from our discussion is that, even if censorship does lower our status in some respects, this lowering is at least somewhat compensated by our status being raised in other respects. We might concede then that censorship is in some respects insulting, but nonetheless point out that this insult has been exaggerated by the likes of Nagel, who has overlooked the significant and compensating benefits to our status that censorship secures. Censorship may well lower the status of citizens in some respect, but it also offers them compensation in kind.

Let us move on to the second way of understanding the key move in Nagel's argument. On this second reading the problem with censorship is not that it bestows on citizens a lower political status than they might have otherwise enjoyed, but that this lower status is unfitting. Persons, so the thought goes, naturally possess a special kind of moral status that makes them intellectually inviolable. As such, the government fails to properly recognize its citizens and their true moral status when it leaves citizens intellectually violable. It fails, in other words, to grant citizens the kind of elevated political status that is appropriate and right for beings like us.

This second reading echoes the argumentative strategies we find in Nagel's chief inspirations (Kamm and Quinn). Quinn, for instance, writes that "it is not that we think it fitting to ascribe rights because we think it a good thing that rights be respected. Rather we think respect for rights a good thing precisely because we think people actually have them—and, if my account is correct, that

15 Connectedly, Anne-Sofie Greisen Hojlund suggests that the government's decision not to engage in welfare-promoting, lifesaving regulation may convey a variety of objectionable attitudes, including neglect, indifference, and unwillingness to give appropriate weight to the strong interests of others. See Greisen Hojlund, "What Should Egalitarian Policies Express?"

they have them because it is fitting that they should."[16] The main problem with this second reading is that it boils down to a brute claim about the underlying inviolable status of persons. Nagel's account began as an attempt to contextualize and unpack a certain kind of intuitive worry that many (including Nagel) register with censorship. Rather than leave the argument at the level of intuition, however, Nagel suggested that we can productively unpack and even help justify this intuitive response by interpreting it as a worry about status and recognition. The problem is that at this point in Nagel's argument we now find ourselves with the brute claim that persons just are the kinds of beings that are intellectually inviolable. Without further independent argument in favor of this brute claim, those of us who do not already find ourselves drawn to this striking vision of citizens as naturally intellectually inviolable will find Nagel's account unpersuasive.

Moreover, one might even worry that this brute approach to status risks lapsing into an overly selfish, individualistic view of persons. Nagel may view any interference with our intellectual autonomy as degrading. But we should be wary of accepting this assumption too quickly. Granted, censorship involves bestowing burdens on citizens—citizens may now have restricted access to certain kinds of arguments. Some of these burdens may be simply for the benefit of other citizens. And some of these burdens may be unpleasant. But there is nothing necessarily degrading about taking on burdens for others. Recall that we are here reflecting on the status and nature of people. Even if it is the case that persons sometimes have to make unpleasant sacrifices for one another, this hardly implies that the persons themselves are unpleasant or thereby belong to a lower echelon of creature. Perhaps even the most wonderful creatures may sometimes have to help each other out.

Nagel may have room to respond here. One thought is that, while citizens should sometimes take on burdens for one another, the government nonetheless degrades citizens when it enforces these burdens. This thought is strengthened if one views censorship as compelling citizens to shoulder burdens for others. This is an interesting line of thought, but I offer two responses. First, some will contest the suggestion that censorship compels citizens or in some sense "makes the choice for them." An alternative and milder way of characterizing censorious laws is that they give citizens additional reasons to act in a particular way. On this milder way of characterizing censorious regulation, such regulation falls short of wholly determining what citizens do. Second, if censorship is degrading in this way, then note that a whole host of other relatively uncontroversial government regulations are also degrading. Taxation, restrictions on playing music loud late at night, and anti-monopoly laws can

16 Quinn, *Morality and Action*, 173.

all be framed as the government forcing citizens to shoulder burdens for one another. Are we to conclude that these laws are also degrading? Even if one is tempted to answer this question with a yes, the objection to censorship we are considering seems less like an objection to censorship *per se* and more like a general anarchist worry with government regulation.

3. RECOGNITION RESPECT

To end this discussion, I consider one final way in which censorship might be thought to deliver a special kind of insult. Perhaps censorship is insulting insofar as it involves a failure to properly recognize that citizens are thinking beings. Or, as Darwall might put, censorship involves a lack of "recognition respect" on the government's part for its citizens and their fundamental thinking capacities. Of course, this is close to the Nagelian worry canvassed about how censorship might bestow on citizens an unfittingly low political status. But the worry here is not so much about political status as it is about the extent to which a government can simultaneously interfere with the intellectual capacities of its citizens in as direct a way as is involved in censorship, while still having a proper appreciation for those citizens and their intellectual capacities. Perhaps censorious governments are so concerned with pointing to the shortcomings and dangers associated with the intellectual capacities of citizens that recognition respect for citizens falls out of the picture.

Unlike appraisal respect, recognition respect does not involve appreciating some achievement or excellence of character on a person's part.[17] Instead, it primarily involves giving a person the due consideration and respect that is owed them simply in light of their being a person.[18] It is common to hear governments who engage in radical rights-violating behaviors being accused of lacking "recognition respect" for their citizens. Such a government, so the thought goes, fails to recognize the basic human capacities of its citizens and how these capacities ought to inform how the government treats these citizens. But we need not reserve this objection simply for such extreme cases. Jonathan Quong, for instance, has argued that paternalism involves a failure to recognize the nature and capacities of those being paternalized.[19] Can a similar argument be constructed in order to problematize censorship?

17 Darwall, "Two Kinds of Respect."

18 I say "primarily" because Darwall also thinks that recognition respect can be granted by responding appropriately to someone's "presented self." However, I will say no more about this aspect of recognition respect.

19 Quong, *Liberalism without Perfectionism*. Echoing Rawls, Quong argues that agents have two crucial moral powers, the second of which is the "capacity to form, revise, and

A relatively straightforward version of such an argument goes as follows: censorship involves a failure on the government's part to recognize the fundamental capacity for moral assessment that its citizens possess, and the importance of their exercising this capacity free from outside interference. The government that engages in censorship may have the best of intentions and aim only to prevent innocent persons from being harmed. But its pursuit of these aims through censorious regulation reflects a failure to recognize that its citizens are thinking beings, capable of arriving at their own conclusions. This species of argument underpins Ronald Dworkin's widely cited suggestion that the government "insults its citizens . . . when it decrees that they cannot be trusted to hear opinions that might persuade them to dangerous or offensive convictions."[20]

This straightforward version of the recognition respect worry is unlikely to win many admirers. Its key claim is that censorship involves a failure on the government's part to recognize that its citizens are thinking beings that possess the capacity to assess for themselves what kind of beliefs and intentions they ought to develop. However, not only is censorship compatible with the government recognizing that its citizens are thinking beings, such recognition is in fact necessary for engaging in censorship in the first place. After all, the reason the government engages in censorship is that it is worried that citizens, left to their own devices, will be exposed to arguments and ideas that persuade them to develop harmful thoughts. This government, then, is fully aware that its citizens are capable of coming to their own conclusions about what kind of thoughts are worth having. Indeed, that is the whole problem! It is the fact that citizens have this kind of intellectual power, and may use it unwisely, that explains why intervention is necessary.

In reply, one might argue that all the above really shows is that censorship involves a formal recognition of the fact that citizens have certain intellectual capacities. As such, a more sophisticated version of the recognition respect worry pushes the thought that true recognition involves more than this. The sadistic murderer who takes special delight in slowly extinguishing the sentience of his victims may well formally acknowledge the humanity of those he kills. Indeed, this kind of formal recognition is part of his sadistic motivation for killing (he enjoys seeing his victims' humanity extinguished). Yet at the same time he fails to fully recognize the moral significance of their humanity and how it is supposed to modify his behavior. In a similar vein, while censorship may well be compatible with a formal recognition of the thinking nature of citizens, its critics might argue

rationally pursue [one's] conception of the good" (2). The problem with paternalism, so Quong suggests, is that it treats agents as though they lack this second Rawlsian power; it treats them as though they lack this capacity.

20 Dworkin, *Freedom's Law*, 200.

that it is incompatible with a richer appreciation of their thinking nature. This richer appreciation, so the argument goes, involves at the very least attaching some significant normative weight to the intellectual autonomy of citizens.

This is an interesting and challenging objection to governmental censorship. Here I offer two initial responses. First, depending on how one understands what it means to recognize and appreciate our intellectual capacities, promoting our intellectual capacities in the long run may sometimes involve interfering with those same capacities in the short run. In some ways this is a straightforward idea—we all know, for instance, that promoting a patient's long-term health may involve giving him medicines that make him unwell in the short term. A similar point arguably applies when it comes to our intellectual capacities. For instance, certain persuasive appeals may help reinforce an environment that is hostile to certain marginalized groups. As a result, members of these marginalized groups may be deterred from both expressing themselves in public and from engaging with popular arguments and ideas.[21] Moreover, hostile environments may present obstacles to agents developing their intellectual skills and pursuing their intellectual interests. Drawing these thoughts together, we see that an appreciation for the value of agents utilizing their intellectual capacities may in fact establish a case in favor of governmental censorship. Governments that engage in censorship may be taking the intellectual autonomy of their citizens very seriously—it is just that they think, with some justification, that the value of intellectual autonomy tells both for and against censorship, and sometimes more tellingly for censorship.

Second, plausibly the government can recognize the importance of its citizens' intellectual capacities and how citizens generally have a powerful claim against censorship, while nonetheless holding that certain cases of censorship are all-things-considered justified. Such a government might recognize the moral importance of its citizens' intellectual capacities and the *pro tanto* interest they therefore have in being free from censorship, while also judging that sometimes other moral factors (such as harm prevention and the promotion of well-being) have even greater moral weight. Such a government may step back from granting the intellectual autonomy of its citizens infinite (or trumping) moral weight. Nonetheless the government recognizes that our intellectual autonomy has significant moral value.

What options are left for the critic of censorship who wishes to insist that censorious governments fail to properly recognize the intellectual capacities of their

21 Williams, "Stress and the Mental Health of Populations of Color"; Kwate and Meyer, "On Sticks and Stones and Broken Bones"; Priest et al., "A Systematic Review of Studies Examining the Relationship between Reported Racism and Health and Wellbeing for Children and Young People."

citizens? One option would be to insist that the only way for the government to properly recognize the moral importance of our intellectual capacities is to grant citizens a claim against censorship that cannot be outweighed by other moral considerations—i.e., to grant the intellectual autonomy of its citizens infinite, trumping moral weight. But this suggestion is vulnerable to the kind of bruteness worry we canvassed earlier when discussing Nagel. After all, this suggestion simply assumes that the only proper way to appreciate an agent's intellectual capacities is to steadfastly refuse ever to interfere with her intellectual autonomy.

A more promising strategy would be for the critic of censorship to suggest that there is a gap in the argument of those of us who think censorship is compatible with recognition, and that this gap needs to be filled. In particular they might ask, with some justification, just how appreciative a vision the government can have of its citizens and their intellectual capacities when it openly admits that these capacities are limited, sometimes harmful, and sometimes worth limiting in the name of other values. The worry here is subtle. Think of a child who grows up in awe of the beauty of music and who would not give up their dream of becoming a musician for the world. Then the child grows up and learns that, not only is being a musician more frustrating, mundane, and technical than they had imagined it to be, but that sometimes other things in life are more important. We might well think that, from this person's perspective, being a musician and music more generally have lost some of their luster. Similarly, those of us who think that the government can simultaneously appreciate the intellectual capacities of its citizens and engage in censorship should reflect carefully on just what kind of appreciation we are really left with. Are we left with a vision of citizens and their intellectual autonomy that, while appreciative to some degree, have also lost much of its luster? At the very least, it seems that we should try to provide some description of how such a government views the citizenry that it censors.

At the close of our discussion, then, the key question is what kind of vision of its citizens and their intellectual capacities a censorious government really possesses. I end this essay with a four-point sketch of the conception of citizens and their place in a political community that might underpin a government's decision to engage in censorship. The sketch draws together several insights touched upon already in this essay, and has one main aim: to demonstrate that a government that engages in censorship may nonetheless be committed to a genuinely appreciative and attractive vision of its citizens.

First, the government recognizes that the capacity citizens possess for intellectual autonomy (i.e., the ability to assess arguments and ideas and form beliefs and intentions in light of this assessment) is to some extent flawed. That is to say, the government reasonably views its citizens as liable to discharge their intellectual capacities in ways that may be unwise, affected by bias, self-defeating,

liable to be a cause for regret in the future, etc., and thus conclude that citizens are liable to at least sometimes arrive at imperfect beliefs and intentions.[22]

Second, the government recognizes that its citizens have a deep interest in exercising their intellectual autonomy. But it also rejects the simplified vision of citizens as merely intellectual beings whose only or predominant interest is in enjoying intellectual inviolability. Instead, the government embraces a more holistic vision of citizens as having a range of interests and capacities—some of which are intellectual, but others of which may be more accurately characterized as emotional, social, relational, physical, etc.[23]

Third, the government holds that whether citizen interests are met depends on their environment. For instance, as thinkers, we benefit greatly from being able to share our ideas with sympathetic audiences who are happy to respond with their own critical reflections on our ideas. We likewise benefit from being able to engage with the ideas and arguments of others.[24] However, the speech of others can also be both indirectly and directly threatening to our interests. The proliferation of hate speech, for instance, may inspire listeners to harass, discriminate, and assault certain people. Hate speech may also inspire listeners to stop engaging with the ideas and speech of those groups that the hate speech vilifies.[25] As if this was not bad enough, being the target of hate speech is correlated with displaced aggression, avoidance, social withdrawal, decreased political participation, alcoholism, suicide, and increased levels of stress and anxiety.[26] In addition, being targeted by hate speech may make one less likely

22 Caputo, "A Literature Review of Cognitive Biases in Negotiation Processes"; Murata, Nakamura, and Karwowski, "Influence of Cognitive Biases in Distorting Decision Making and Leading to Critical Unfavorable Incidents"; Rabin, "Projection Bias in Predicting Utility."

23 For instance, citizens have a deep interest in, e.g., being physically safe and healthy, successfully pursuing their goals, having an adequate sense of self-worth, and having a suitable range of functioning capabilities (e.g., bodily health, bodily integrity, an adequate range of emotional capabilities, a sense of self-respect, the ability to pursue play and leisure activities). See Westlund, "Rethinking Relational Autonomy"; Benson, "Free Agency and Self-Worth,"; Govier, "Self-Trust, Autonomy, and Self-Esteem"; Nussbaum, *Women and Human Development* and "Capabilities as Fundamental Entitlements"; Sen, *Commodities and Capabilities*, "Development as Capability Expansion," and *The Idea of Justice*.

24 Shiffrin, *Speech Matters*.

25 Williams, "Stress and the Mental Health of Populations of Color"; Kwate and Meyer, "On Sticks and Stones and Broken Bones"; Priest et al., "A Systematic Review of Studies Examining the Relationship between Reported Racism and Health and Wellbeing for Children and Young People."

26 Matsuda, "Public Response to Racist Speech" and *Words That Wound*; Williams, "Stress and the Mental Health of Populations of Color"; Lewis, Cogburn, and Williams, "Self-Reported Experiences of Discrimination and Health"; Brown, *Hate Speech Law*.

to engage in public discourse and the sharing of ideas, which we have already suggested is important for the development of one's intellectual capacities.[27]

Fourth, the government views citizens as sometimes liable to take on burdens for one another. Of course, there is a limit on the kinds of burdens citizens can be expected to take on in the service of their fellow citizens. All the same, there will be occasions when citizens will be expected to shoulder moderate burdens for one another. Given that, as discussed above, the proliferation of certain kinds of arguments may strike a blow against the interests of citizens, the government recognizes that imposing certain limits on expression will sometimes help protect the interests of certain citizens.

This vision of citizens and of their role in the political community is neither insulting nor unappreciative. Granted, this vision conceptualizes citizens as flawed reasoners, and as sometimes liable to harm one another as a result of exposure to certain arguments. Likewise, this vision suggests that citizens possess a certain kind of vulnerability, and that some degree of cooperation is required if we are to truly thrive. But these suggestions stem from an accurate and grounded understanding of how our intellectual capacities function in practice. Critics of censorship may insist that we embrace a more flattering vision of citizens—one that conceptualizes us as highly competent, independent, and self-sufficient thinking beings. But this more flattering vision risks being so detached from the real-world functioning of people as to lapse into a kind of vanity.

4. Conclusion

Censorship then, despite what its critics might say, is not deeply insulting. The best kind of censorship, far from being premised on a derisive or disrespectful view of citizens, rather proceeds from a holistic appreciation of the varying interests, capacities, limitations, and vulnerabilities of citizens, as well as the need for cooperation. Those who consider this vision to be insulting because of the way it acknowledges certain imperfections and duties of people may well be guilty of a kind of vanity—a vision of themselves and people more generally that, while rather flattering, is detached from reality.

That said, it is worth reflecting on the fact that it is only the "best kind" of censorship that avoids deeply insulting its citizens. Critics of censorship may be justified in suggesting that a good deal of censorship does not in fact proceed from such a holistic view of the citizenry and is therefore insulting. For instance, it may be that, were citizens left free to engage in certain kinds of hate

27 Bishop and Simpson, "Disagreement and Free Speech."

speech, this would result in only minor harms that would not normally justify government intervention. Perhaps citizens would, by and large, simply reject hate speech as the nonsense it in fact is. Were the government to engage in censorship in this case, based on an exaggerated fear about the harmful fallout of this kind of hate speech being permitted, it would plausibly count as insulting its citizens.[28] Similarly, it may be that certain governments engage in censorship largely because they undervalue (or are simply unconcerned about) the value of citizens being free to exercise their intellectual capacities. Again, such a government would plausibly qualify as holding an unacceptably insulting view of its citizenry.

What we should take from our discussion is that there is a need to consistently scrutinize and challenge censorious governments. Two governments might both decide to engage in censorship, but if one does so on the basis of an accurate recognition of the harms involved and the imperfect nature of human reasoning, while the other does so out of disdain for its citizens and a lack of concern for their intellectual autonomy, then we are dealing with two very different cases of censorship. Censorship may be an effective tool for harm prevention (alongside other tools, such as education). But the price we pay for wielding this tool, it would seem, is eternal vigilance.[29]

Oxford University
sebastien.bishop@philosophy.ox.ac.uk

REFERENCES

Benson, Paul. "Free Agency and Self-Worth." *Journal of Philosophy* 91, no. 12 (December 1994): 650–68.

Bishop, Sebastien. "Back to School: Matthew Kramer's *Freedom of Expression as Self-Restraint*." *Modern Law Review* 86, no. 2 (March 2023): 564–87.

Bishop, Sebastien, and Robert Mark Simpson. "Disagreement and Free Speech."

28 Kramer, *Freedom of Expression as Self-Restraint*. Though, as I have argued, this is partly an empirical matter to do not only with whether designing an excellent education system that makes censorship redundant is theoretically possible, but also whether designing such an education system is sufficiently difficult as to leave governments blameless if they try and fail on this score; see Bishop, "Back to School."

29 I would like to thank the two anonymous reviewers who provided comments on an earlier draft of this article, as well as Rosa Bell, Alexander Brown, Rachel Fraser, Alice Harberd, Jeffrey Howard, Chong-Ming Lim, William Russ, and Tatiana Sitnikova. Special thanks to Jessica Fischer and Robert Mark Simpson.

In *The Routledge Handbook of Disagreement*, edited by Maria Baghramian, Adam Carter, and Richard Rowland. New York: Routledge, forthcoming.

Brison, Susan. "The Autonomy Defense of Free Speech." *Ethics* 108, no. 2 (January 1998): 12–39.

Brown, Alexander. *Hate Speech Law: A Philosophical Examination*. New York: Routledge, 2015.

Caputo, Andrea. "A Literature Review of Cognitive Biases in Negotiation Processes." *International Journal of Conflict Management* 24, no. 4 (2013): 374–98.

Darwall, Stephen. "Two Kinds of Respect." *Ethics* 88, no. 1 (October 1977): 36–49.

De Marneffe, Peter. "Avoiding Paternalism." *Philosophy and Public Affairs* 34, no. 1 (Winter 2006): 68–94.

Dworkin, Ronald. *Freedom's Law: The Moral Reading of the American Constitution*. Cambridge, MA: Harvard University Press, 1996.

Govier, Trudy. "Self-Trust, Autonomy, and Self-Esteem." *Hypatia* 8, no. 1 (Winter 1993): 99–119.

Greisen Hojlund, Anne-Sofie. "What Should Egalitarian Policies Express? The Case of Paternalism." *Journal of Political Philosophy* 29, no. 4 (December 2021): 519–38.

Kagan, Shelly. "Replies to My Critics." *Philosophy and Phenomenological Research* 51, no. 4 (December 1991): 919–28.

Kamm, Frances. *Rights, Duties, and Status*. Vol. 2 of *Morality, Mortality*. Oxford: Oxford University Press, 1996.

Kramer, Matthew. *Freedom of Expression as Self-Restraint*. Oxford: Oxford University Press, 2021.

Kwate, Naa Oyo, and Ilan Meyer. "On Sticks and Stones and Broken Bones: Stereotypes and African American Health. *Du Bois Review: Social Science Research on Race* 8, no. 1 (Spring 2011): 191–98.

Lewis, Tené, Courtney Cogburn, and David Williams. "Self-Reported Experiences of Discrimination and Health: Scientific Advances, Ongoing Controversies, and Emerging Issue." *Annual Review of Clinical Psychology* 11 (2015): 407–40.

Loewenstein, George, Ted O'Donoghue, and Matthew Rabin. "Projection Bias in Predicting Utility." *Quarterly Journal of Economics* 118, no. 4 (November 2003): 1209–48.

Matsuda, Mari. "Public Response to Racist Speech: Considering the Victim's Story." *Michigan Law Review* 87, no. 8 (1989): 2320–38.

———. *Words That Wound: Critical Race Theory, Assaultive Speech, and the First Amendment*. Boulder, CO: Westview Press, 1993

Murata, Atsuo, Tomoko Nakamura, and Waldemar Karwowski. "Influence of

Cognitive Biases in Distorting Decision Making and Leading to Critical Unfavorable Incidents." *Safety* 1, no. 1 (2015): 44–58.

Nagel, Thomas. "Personal Rights and Public Space." *Philosophy and Public Affairs* 24, no. 2 (Spring 1995): 83–107.

Nussbaum, Martha. "Capabilities as Fundamental Entitlements: Sen and Social Justice." *Feminist Economics* 9, nos. 2–3 (2003): 33–59.

———. *Women and Human Development: The Capabilities Approach*. Cambridge: Cambridge University Press, 2000.

Priest, Naomi, Yin Paradies, Brigid Trenerry, Mandy Truong, Saffron Karlsen, and Yvonne Kelly. "A Systematic Review of Studies Examining the Relationship between Reported Racism and Health and Wellbeing for Children and Young People." *Social Science and Medicine* 95 (2013): 115–27.

Quinn, Warren. "Actions, Intentions, and Consequences: The Doctrine of Doing and Allowing." *Philosophical Review* 98, no. 3 (July 1989): 287–312.

———. *Morality and Action*. Cambridge: Cambridge University Press, 1994.

Quong, Jonathan. *Liberalism without Perfectionism*. Oxford: Oxford University Press, 2010.

Sen, Amartya. *Commodities and Capabilities*. Amsterdam: North-Holland, 1985.

———. "Development as Capability Expansion." *Journal of Development Planning* 19 (1989): 41–58.

———. *The Idea of Justice*. London: Allen Lane, 2009.

Shiffrin, Seana. *Speech Matters: On Lying, Morality, and the Law*. Princeton: Princeton University Press, 2014.

Westlund, Andrea. "Rethinking Relational Autonomy." *Hypatia* 24, no. 4 (Fall 2009): 26–49

Williams, David. "Stress and the Mental Health of Populations of Color: Advancing Our Understanding of Race-Related Stressors." *Journal of Health and Social Behaviour* 59, no. 4 (December 2018): 466–85.

Journal of Ethics and Social Philosophy
VOL. 27, NO. 2 · APRIL 2024
https://doi.org/10.26556/jesp.v27i2.2907

THE RIGHT TO MENTAL AUTONOMY

ITS NATURE AND SCOPE

William Ratoff

Let us suppose that you are an anti-vaxxer who has decided against receiving any of the effective COVID-19 vaccinations. Suppose further that your flight has been delayed and that you are sleeping in the departure lounge of an airport. I sneak up on you and deploy my transcranial magnetic stimulation (TMS) technology to interfere with your mind. This technology works by emitting magnetic fields that induce electrical events in your brain. Let us suppose that I implant a desire in you to get vaccinated against COVID-19.[1] After you land at your destination airport, you immediately rush to get vaccinated.

Intuitively, I have wronged you here. But how? I have not harmed you. If anything, I have made your life better: you now have antibodies against COVID-19. One promising explanation of the wrongness of this action is that my action was wrong because it violated your right to bodily autonomy. After all, I induced electrical events in your brain without your consent. And your brain is clearly part of your body and thus falls under the protection of your right to bodily autonomy. In this way then, through appeal to your right to bodily autonomy, we can seek to explain the wrongness of my action of using a TMS device to induce in your mind a desire to get vaccinated.

But this explanation has struck many as being incomplete.[2] How so? Well, suppose I used my TMS device not to interfere with your thinking, but rather to induce a bowel movement in you—perhaps by surreptitiously waving it over your stomach while you were sleeping.[3] Intuitively, I have again wronged you in so acting. And, very plausibly, the wrongness of my action should again be explained through an appeal to your right to bodily autonomy. I have violated

1 Of course, TMS technology does not allow such precise interventions as the implantation of a desire. For better or worse, such interventions remain firmly in the realm of science fiction. (If the reader prefers, please substitute all instances of "TMS technology" or "TMS device" for instances of "sci-fi ray gun" or similar).

2 Douglas and Forsberg, "Three Rationales for a Legal Right to Mental Integrity."

3 As far as I am aware, TMS cannot be used to induce a bowel movement. But let us suppose, for the sake of inducing some relevant moral intuitions, that it can be.

your right to bodily autonomy by inducing physical events in your bowels without your consent with my TMS device—events that, in turn, triggered a bowel movement. This latter action of mine seems *less wrong* than my former action (of using my TMS device to implant a desire in you to get vaccinated), but both my actions, it seems, are equally severe violations of your right to bodily autonomy. After all, in each case, I induce events in your body without your consent by waving my TMS device over you. Consequently, it looks to follow that something else must explain the *additional* wrongness of my former act of interfering with your mind, something over and above the violation of your right to bodily autonomy.

The most natural explanation, I contend, of this extra wrongness is that my former action, but not my latter one, violated your right to mental autonomy—that is, your right against significant, nonconsensual interference with your mind.[4] Only you have the right to *directly* change your thinking about any arbitrary matter or to directly change your plans of action. I cannot permissibly attempt to change your mind without your consent by using TMS—or some other sci-fi method of mind control—to directly change your beliefs, desires, or intentions.[5] Such actions violate your right to mental autonomy—often in addition to their violating your right to bodily autonomy. This, I suggest, is why my former action of interfering with your mind is more wrong than my latter action, which interferes only with the functioning of your body.

A number of moral philosophers and legal scholars have now recognized the existence of a natural, or moral, right to mental autonomy and called for its legal recognition.[6] This right is standardly characterized as your right against significant, nonconsensual interference with your mind. It is your right to make up your own mind for yourself, so to speak. But the precise scope of this right remains thus far undertheorized: What limits does this right place on the morally permissible ways of influencing someone's thinking? What ways of seeking to change someone's mind manifest appropriate respect for their right to mental autonomy? Why would it be permissible for me to attempt to change

4 Bublitz and Merkel, "Crimes against Minds"; Douglas and Forsberg, "Three Rationales for a Legal Right to Mental Integrity."

5 Ienca and Andorno, "Towards New Human Rights in the Age of Neuroscience and Neurotechnology"; Douglas and Forsberg, "Three Rationales for a Legal Right to Mental Integrity."

6 Bublitz and Merkel, "Crimes against Minds"; Bublitz, "Means Matter"; Douglas and Forsberg, "Three Rationales for a Legal Right to Mental Integrity." The right to mental autonomy is also known as "the right to mental self-determination" (Bublitz and Merkel, "Crimes against Minds"), "the right to cognitive liberty" (Ienca and Andorno, "Towards New Human Rights in the Age of Neuroscience and Neurotechnology"), and "the right to mental integrity" (Douglas and Forsberg, "Three Rationales for a Legal Right to Mental Integrity").

your mind about policy *P* by presenting you (nonconsensually, even) with the reasons for favoring policy *P*, but impermissible for me to change your mind about *P* by zapping you with my TMS mind control device?

Here I make the case that the right to mental autonomy is to be correctly analyzed as the right to form attitudes in light of reasons. You form an attitude *autonomously* just when you form it in light of reasons.[7] Consequently, I contend, we should think that the right to mental autonomy just is the right to form attitudes in light of reasons. Once understood this way, we can see why this right protects its holder against all (nonconsensual) "nonrational" interventions on their thinking—including, but not limited to, nonconsensual neurosurgery, pharmacological manipulations, sci-fi mind control, subliminal messaging, and non-reason-giving advertising or nudging. Rather, the only fully permissible ways to seek to influence someone's thinking—those ways that do not violate the right to mental autonomy—are through methods that seek to engage their rational faculties. This result, I claim, accords with our moral intuitions—our ultimate data in this region of philosophical space.[8]

The structure of the rest of this paper goes like this: in section 1, I argue that there is good reason to believe that we (adult humans) possess a natural, or moral, right to mental autonomy. Then, in section 2, I make my case that this right can be correctly analyzed as the right to form attitudes in light of reasons and investigate the precise limits that this right places on the morally permissible ways of influencing someone's thinking. Last, in section 3, I consider various problematic cases that might be thought to pose a challenge for my analysis.

1. A RIGHT TO MENTAL AUTONOMY?

Why think that we possess a natural, or moral, right to mental autonomy—a right to make up our minds for ourselves?

First, a couple of distinctions: I am here concerned only with a natural, or moral, right to mental autonomy, not the legal recognition of such a right—that is, a *legal right* to mental autonomy. We rational agents possess natural or moral rights. This has been recognized by many moral philosophers.[9] For example, according to Locke, we have natural rights to—among other things—life, liberty, and the ownership of property.[10] Robert Nozick put it like this:

7 Velleman, "What Happens When Someone Acts?"; Korsgaard, *Self-Constitution*.

8 Kagan, *Normative Ethics*.

9 Nozick, *Anarchy, State, and Utopia*; Thomson, *Realm of Rights*; Raz, "On the Nature of Rights."

10 Locke, *Two Treatises of Civil Government*.

"Individuals have rights and there are things no person or group may do to them (without violating their rights). So strong and far-reaching are these rights that they raise the question of what, if anything, the state and its officials may do."[11]

We possess these natural or moral rights in virtue of our natures—for example, our humanity, or our rationality, or the fact that, as sentient beings, we have interests.[12] Even in a state of nature, we humans would possess such rights. We do not have them because there is some bill of rights, or constitution, that declares that we possess them. No—their existence is independent of any such legal pronouncement or ruling. For many moral philosophers, natural rights play an important role in our understanding of moral reality; in particular, they explain wrongdoings.[13] Why was it wrong for Lee Harvey Oswald to assassinate JFK? Because JFK possessed a right to life, and by killing him, Oswald violated this right. But it would not have been wrong for Oswald to swat an annoying fly at that very same moment in 1963, causing its death, since flies do not possess a right to life.

In contrast, legal rights are artifacts of the state.[14] We possess them simply because the correct governmental body has decreed that we possess them. As a British citizen, former prime minister David Cameron possesses a legal right to reside in the United Kingdom that former president Bill Clinton, a citizen of the United States only, lacks. Cameron possesses this right of residence because the British state has decreed that part of what it is to be a British citizen is to possess such a right. In a state of nature, there would be no legal rights. In contrast with natural rights, there are either no or more limited necessary connections between legal rights and morality or wrongdoings. Natural rights and legal rights can (and have) come apart. For example, in Nazi Germany, the state stripped Jewish people of the legal recognition of some of their (natural) property rights. Although these people still possessed a moral right to this property, they no longer—according to the German state—had any legal right to it. I shall not be concerned here with the legal right to mental autonomy. However, it should be noted that a number of legal scholars and moral philosophers have already called for its recognition by the law.[15]

We should also distinguish between the negative and positive dimensions of a (natural) right.[16] Rights correlate with duties: if I have a right to *X*, then you

11 Nozick, *Anarchy, State, and Utopia*, ix.

12 Raz, "On the Nature of Rights"; Markovits, *Moral Reason*.

13 Nozick, *Anarchy, State, and Utopia*; Thomson, *The Realm of Rights*.

14 Hart, *The Concept of Law*.

15 Bublitz and Merkel, "Crimes against Minds"; Douglas and Forsberg, "Three Rationales for a Legal Right to Mental Integrity."

16 Narveson, *The Libertarian Idea*.

have a duty to abstain from preventing me from attaining *X* or, if appropriately situated, a duty to assist me in attaining *X*. The former duty corresponds to the negative component of my right to *X*, the latter duty with the positive component of my right. JFK's right to life entailed a duty on the part of all third parties to abstain from killing him. This corresponds to the negative component of his right to life. But his right to life also entailed an obligation on appropriately situated others to get him medical attention once he had been shot. This corresponds to the positive aspect of his right to life.

The right to mental autonomy, under investigation here, has positive and negative dimensions. This has already been recognized by those moral philosophers and legal scholars who have written about this right.[17] Most discussion of our right to mental autonomy has focused on its negative component. This should be apparent from its standard characterization as our right against significant, nonconsensual interference with our minds. This negative component of our right to mental autonomy entails (something like) a duty on the part of third parties to abstain from engaging in significant, nonconsensual interventions in our minds. But this right also has a positive dimension characterized by Bublitz and Merkel as the "freedom *to* self-determine one's inner realm, e.g., the content of one's thoughts, consciousness or any other mental phenomena."[18] This aspect of your right very plausibly corresponds to a duty on the part of appropriately situated others—for example, educators or mental health professionals—to assist you in mentally self-determining.

Back to our initial question: Why think that we possess a natural right to mental autonomy? The case of TMS-ing the anti-vaxxer, with which I began this paper, gives us strong reason, I believe, to hold that this is the case. Recall that in the example, I used TMS technology to nonconsensually implant a desire to get vaccinated against COVID-19 into your mind while you were asleep. Intuitively, I have wronged you in so acting. In general, wrongdoings are explained by natural rights violations.[19] Granting this, we should think that I have violated (at least) one of your natural rights in so acting.

But which right? As I noted before, I have not harmed you by inserting this desire into your mind. It is an easily satisfiable desire, one that causes you no suffering and is quickly extinguished once you go and get vaccinated against COVID-19. Consequently, we cannot say that I have violated your right against being harmed. One promising explanation of the wrongness of this action is

17 Bublitz and Merkel, "Crimes against Minds"; Douglas and Forsberg, "Three Rationales for a Legal Right to Mental Integrity."

18 Bublitz and Merkel, "Crimes against Minds," 60.

19 Nozick, *Anarchy, State, and Utopia*; Thomson, *Realm of Rights*.

that my action was wrong because it violated your right to bodily autonomy. After all, I directly influenced the functioning of your brain with the electromagnetic waves my TMS device emits. Very plausibly, this constitutes a violation of your right to bodily autonomy: your brain is clearly part of your body. When granting that the mind is (something like) the functioning of the brain, all direct manipulations of the mind are going to involve interventions in brain function. Consequently, it looks like we can explain the wrongness of my action of inserting a desire into your mind with my TMS device simply through appeal to your right to bodily autonomy.

Nevertheless, it is still natural to think that I wronged you in some way that is "over and above" the wrong I committed by interfering with the functioning of your brain. There is some residual wrongdoing, so to speak, that is left unaccounted for if we try to explain the wrongness of my action simply through appeal to this violation of your bodily autonomy. If I use my TMS device to (harmlessly) zap your bowels, such that you suddenly need to go to the toilet, then I have done something wrong. But, intuitively, I have done something *less wrong* than when I interfere with your thinking with my TMS device. When I zap your bowels, I have violated your right to bodily autonomy but not your right to mental autonomy. The extra, or residual, wrongdoing that is left over in my act of inserting a desire into your mind, once we subtract out my violation of your right to bodily autonomy, is, I contend, a separate violation of your distinct right to mental autonomy. The most complete explanation of the "full wrongness" of my action, I believe, is that my action was wrong, not simply because it violated your right to bodily autonomy, but also because it violated a distinctive natural right to mental autonomy—a right against significant, nonconsensual interference with your mind—that you possess.[20] In this way, then, we are warranted in positing a natural right to mental autonomy as part of the best explanation of the wrongness of my act of inserting a desire into your mind without your consent.

On this point, Douglas and Forsberg contrast the case of a barista who, seeing that one of her regular customers looks a little down, surreptitiously slips into his coffee a mild, fast-acting *antidepressant* that lifts his mood for several hours with the case of a barista who, seeing that one of her regular customers is a little wheezy, covertly slips into his coffee a mild, fast-acting *anti-asthmatic medication* that makes his breathing somewhat easier for several hours.[21] Intuitively, benevolently spiking someone's coffee with an antidepressant is *prima*

20 Bublitz and Merkel, "Crimes against Minds"; Douglas and Forsberg, "Three Rationales for a Legal Right to Mental Integrity."

21 Douglas and Forsberg, "Three Rationales for a Legal Right to Mental Integrity," 188.

facie more wrong than benevolently spiking it with a similarly mild anti-asthmatic medication.[22] But we cannot explain this moral difference through an appeal to the right to bodily autonomy: each intervention involves a similar degree of bodily interference. The best explanation, it seems, of it being more wrong to covertly slip someone an antidepressant than an anti-asthmatic is that people possess a right to mental autonomy over and above their right to bodily autonomy and that the antidepressant, but not the anti-asthmatic, interferes with the person's mind, violating this right to mental autonomy.

In their "Crimes against Minds: On Mental Manipulations, Harms and a Human Right to Mental Self-Determination," the *locus classicus* for all recent discussions of the right to mental autonomy, Bublitz and Merkel catalog a range of hypothetical cases that collectively constitute further strong evidence that we have a natural right to mental autonomy.[23] Their first case concerns a struggling restaurant that spikes customers' drinks with a chemical—a low dose of ghrelin that increases their feeling of being hungry but that otherwise has no discernible effects—such that they order more food, thereby increasing the restaurant's revenue. Intuitively, this kind of manipulation is wrong, and wrongdoings are explained by rights violations.[24] The most natural explanation of the wrongfulness of this action, it seems to me, is that it violated the customers' rights to mental autonomy. Other cases they describe include the use of subliminal messaging by an online store and the covert nonconsensual modulation of brain activity, leading to wild mood swings, using an implanted deep brain stimulator electrode.[25] In each such case, although there is plausibly some violation of bodily autonomy since the brain (at least) is nonconsensually influenced, there is nevertheless still a need to invoke a distinctive right to mental autonomy to fully explain the wrongness of the described actions. This constitutes further reason, I think, to posit a natural right to mental autonomy.

2. The Nature and Scope of the Right to Mental Autonomy

Let us suppose that we do indeed possess a natural right to mental autonomy—as a number of moral philosophers and legal scholars have been professing.[26] Important questions still remain. In particular, the question of the exact *scope*

22 Douglas and Forsberg, "Three Rationales for a Legal Right to Mental Integrity."

23 Bublitz and Merkel, "Crimes against Minds."

24 Nozick, *Anarchy, State, and Utopia*; Thomson, *Realm of Rights*.

25 Bublitz and Merkel, "Crimes against Minds," 58–59.

26 Bublitz and Merkel, "Crimes against Minds"; Ienca and Andorno, "Towards New Human Rights in the Age of Neuroscience and Neurotechnology"; Douglas and Forsberg, "Three Rationales for a Legal Right to Mental Integrity."

of this right still stands: What limits does the right place on the morally permissible ways of influencing someone's thinking? What ways of seeking to change someone's mind manifest appropriate respect for their right to mental autonomy? What makes some ways of influencing someone's thinking—rational argumentation, say—permissible, but other ways—pharmacological manipulation—impermissible? In the rest of this paper, I investigate this matter and develop an account. I should say in advance that my proposal is very much intended to be understood as a *working* account—not as a definitive statement, but rather as a proposal that serves as a good "first pass" that will (most likely) need to be refined in later work.

The standard characterization of the right to mental autonomy is that it is your right against significant, nonconsensual interference with your mind.[27] There is something going for this characterization: just as there are ways of influencing someone else's body that are so trivial that they do not count as violating their right to bodily autonomy—for example, waving your hands around near someone's arm such that it causes the hairs on their arm to quiver—there may, plausibly enough, be ways of nonconsensually influencing someone's mind that are so trivial they do not count as violating their right to mental autonomy.[28]

Nevertheless, this analysis is lacking in certain key respects. First, it is quite obscure what counts as a significant, nonconsensual intervention on, or interference with, someone's mind. This characterization does not really help us to partition the permissible ways of influencing someone's thinking from the impermissible ways. Second, there are plausible counterexamples. For example, there is nothing even *prima facie* wrong, or wrong-making, about changing someone's mind on some important topic by (nonconsensually) presenting them with compelling arguments—say, by suddenly and loudly proclaiming your argument on a soapbox on a bustling street such that they cannot help but hear them. You have not violated anyone's right to mental autonomy by so acting, but this looks to count as a significant, nonconsensual intervention on their mind. Consequently, it seems that the right to mental autonomy cannot be correctly analyzed simply as the right against significant, nonconsensual interference with your mind. There must be more to the right to mental autonomy than this.

In their 2014 paper "Crimes against Minds," Bublitz and Merkel offer an alternative analysis of this right. (Bublitz further discusses this account in his 2020 paper, "Why Means Matter.") They suggest that we can understand

27 Bublitz and Merkel, "Crimes against Minds"; Douglas and Forsberg, "Three Rationales for a Legal Right to Mental Integrity."

28 Douglas and Forsberg, "Three Rationales for a Legal Right to Mental Integrity."

the scope of the right to mental autonomy by first distinguishing between *direct* and *indirect* interventions on the mind. Direct interventions include changing someone's mind through the use of TMS, direct brain stimulation, or psychoactive substances. In contrast, rational persuasion counts among the indirect interventions. Bublitz and Merkel characterize this distinction in the following way:

> Direct interventions are those working directly on the brain . . . whereas indirect interventions are somehow more remote—mediated, as it were, by internal processes on the part of the addressee. Tentatively, indirect . . . interventions are those stimuli which are perceived sensually . . . and pass through the mind of the person, being processed by a host of psychological mechanisms. Thus, conscious communication in all its forms is an indirect intervention. By contrast, direct . . . interventions are stimuli reaching the brain by other routes than sensual perception. . . . Roughly one could say that *indirect interventions are inputs into the cognitive machinery our minds are adapted to process, whereas direct interventions change the cognitive machinery itself.*[29]

Bublitz and Merkel then suggest that this distinction carves at the normative joints with respect to the scope of our right to mental autonomy. Roughly speaking, direct interventions on our minds violate our right to mental autonomy; indirect ones do not. In their words: "Prima facie, indirect interventions are permissible, direct ones not."[30] A virtue of this account is that it correctly classes your act of changing someone's mind on an important topic by (non-consensually) presenting them with compelling arguments—an indirect intervention—as permissible and as not violating their right to mental autonomy. Likewise, it correctly classes the barista's action of improving her customer's mood by spiking his coffee with anti-depressants—a direct intervention on his thinking—as impermissible.

However, as the authors themselves acknowledge, this analysis of the right to mental autonomy is problematic. Most pertinently, manipulating someone's mind with subliminal messaging counts as an *indirect* intervention on their thinking. But it is still morally wrong. Consider, for example, Bublitz and Merkel's own example of subliminal influence:

> An online store shows Flash movies to customers which subliminally prime brand *C* and cause customers to evaluate *C* more positively.[31]

29 Bublitz and Merket, "Crimes against Minds," 69–70 (emphasis added).

30 Bublitz and Merkel, "Crimes against Minds," 73.

31 Bublitz and Merkel, "Crimes against Minds."

> While stimuli are not powerful enough to create completely new desires, they tip the scales of inclined customers toward *C*'s product. While overall sales remain constant, *C*'s products are increasingly bought."[32]

Here, viewers are being caused to evaluate *C* more positively by the prime they unconsciously perceive—very plausibly, via the familiarity bias.[33] Intuitively, there is something morally objectionable about seeking to influence consumers' choices in this kind of way. Further evidence for this comes from the furor over market researcher James Vicary's 1957 claim that he had caused an 18.1 percent increase in Coca-Cola sales and a 57.8 percent increase in popcorn sales by inserting single frames saying "Drink Coca-Cola" and "Eat Popcorn" into a movie. According to Vicary, these frames were presented so briefly that they could not have been consciously perceived—rather, they had their behavioral effects subliminally. Although these results turned out to be fabricated, Vicary's claim still led to a moral panic among the general public at the time, with calls to ban subliminal advertising that have persisted to the present day.[34] Granting the veracity of these moral intuitions, it follows that Bublitz and Merkel's distinction between direct and indirect interventions is not carving at the normative joints with respect to articulating the scope of our right to mental autonomy.[35] This right can be violated by indirect interventions on our thinking just as it can be by direct ones. In this way, then, we can see why the right to mental autonomy cannot be correctly analyzed simply as our right not to be subject to direct interventions on our minds.

Another example of an intervention on people's minds that is indirect but nevertheless morally wrong is brainwashing. As I write, the government of China is imprisoning many thousands of Uighur people in "transformation through education" camps, in which Uighur people are brainwashed into accepting tenets and ideals endorsed by the Chinese State and repudiating their own culture.[36] Of course, the wrongs committed here by the Chinese State are many and various. They include, among their number, violations of the right to liberty, the right to bodily autonomy, and the right to life.[37] But there is also a clear violation of the imprisoned Uighur people's right to mental autonomy: the brainwashing they undergo is an attempt by the Chinese State to change their beliefs, desires, and intentions through a nonrational process. For example, detainees are forced to

32 Bublitz and Merkel, "Crimes against Minds," 58.

33 Chartrand, Huber, Shiv, and Tanner, "Nonconscious Goals and Consumer Choice."

34 O'Barr, "'Subliminal' Advertising."

35 Bublitz and Merkel, "Crimes against Minds."

36 Haitiwaji, "Our Souls Are Dead."

37 BBC News, "Who are the Uyghurs?"

repeatedly sing songs declaring their love for the Communist Party of China and, more generally, to outwardly conform to the behavioral ideals preferred by the Chinese State—no doubt in the hope that this will lead to detainees adjusting their attitudes to fit (or rationalize) these behaviors. Irrespective of whether this brainwashing is successful, it is still an attempt to modify and control people's thinking. The conduct of the Chinese State is clearly morally wrong and is an attempt to violate the detainees' right to mental autonomy. But their interventions on the Uighur peoples' thinking are indirect, according to Bublitz and Merkel's classification.[38] This constitutes further reason to hold that the right to mental autonomy cannot be correctly analyzed simply as our right not to be subject to direct interventions on our minds.[39]

But how should we understand it? My proposal here is that the right to mental autonomy should be analyzed as the right to form attitudes in light of reasons. The permissible ways of causing someone to form attitude *A* are partitioned from the impermissible ways by the fact that they involve presenting the person in question with reasons for forming attitude *A*. If someone possesses a right to mental autonomy, then the only morally permissible way to attempt to change that person's mind (say, to cause them to believe that *p*, to desire that *q*, or to pursue end *E*) is to present them with normative reasons for so changing their mind—for example, by presenting them with decisive evidence that *p* is true, or by informing them of sufficient reasons for desiring that *q* or for pursuing end *E*. All other ways of intentionally changing that person's mind—methods that seek to alter their thinking through some (nonconsensual) *nonrational* (or non-reason-giving) process, such as neurosurgery or some sci-fi form of mind control—are classed, on this analysis, as morally impermissible.[40]

In the rest of this paper, I will be developing and defending this analysis of the right to mental autonomy as the right to form attitudes in light of reasons. But first, I will provide some clarification. What is the notion of a reason that I am working with here? By "reasons," I mean normative reasons—considerations that count for or against performing some action or in favor of forming or revising some or other attitude. I will also be understanding the scope of "reasons" to be quite wide. In addition to considerations like the fact that the stove will burn your hand counting as a reason for you to abstain from placing

38 Bublitz and Merkel, "Crimes against Minds."

39 One further example of an indirect violation of someone's right to mental autonomy would be hypnotizing someone without their consent.

40 I am here only defending the view that we adult humans have a right to mental autonomy. It is consistent with everything that I have said here that children do not possess such a right. This may explain why it is permissible to nonrationally condition or habituate children, but not adults, into holding certain attitudes—widely accepted moral judgments, for example.

your hand on it, and a sound argument for proposition *p* counting as a reason to believe that *p*, I will also be countenancing as reasons what might be considered (by some) to be some more edge cases. So, for example, the smell of baked bread is going to constitute a reason, in my use of the term, to feel hunger toward the baked bread in question. After all, from a biological or evolutionary point of view, a fitting or appropriate response to good-smelling food is to feel hunger toward it and to form the desire to eat it. Given all this, it sounds perfectly natural to my ears to say that the smell of the baked bread counts as a reason both in favor of eating the bread and in favor of forming a mental state of hunger that is directed toward the bread. Similarly, sad music is going to count as giving you a reason, on my understanding of the term, for forming certain affective states and feelings—and not just as a cause of you entering those states. Likewise, viewing a painting that expresses the sublime—such as Caspar David Friedrich's *Wanderer above the Sea of Fog*—is going to count, on my view, as a reason for you to feel awe.

A second clarification: What is it exactly to form an attitude in light of a reason? On my understanding, you form an attitude *A* in light of reason *R* just when, and because, (1) you have responded appropriately to reason *R* by forming attitude *A*, and (2) your awareness of R is causally responsible (in the right kind of way) for your forming attitude *A*. I will follow Levy in holding that what it is to respond appropriately to a reason is "to be better or worse disposed toward an action, or to raise or lower one's credence, in a way that reflects the actual force of a reason."[41] And what it is for your awareness of *R* to be causally responsible (in the right kind of way) for your forming attitude *A* is (something like) for your awareness of *R* to cause you to form attitude *A* in a way that does not involve any deviant causal chain—for example, by your awareness of *R* directly causing you to form attitude *A*, unmediated by any intervening mental events. This characterization of what it is to form an attitude in light of a reason, or to respond to a reason, should suffice for my dialectical purpose here of providing an analysis of the right to mental autonomy.

Why should we accept my account of the right to mental autonomy as the right to form attitudes in light of reasons? Ultimately, we should because it captures our moral intuitions concerning the matter: it correctly classes, I claim, the intuitively impermissible ways of influencing someone's thinking as impermissible and the intuitively permissible ways as permissible. And our moral intuitions are our ultimate data in this region of philosophical space.[42] For example, it correctly explains why your right to mental autonomy is violated

41 Levy, "Nudge, Nudge, Wink, Wink," 283.

42 Kagan, *Normative Ethics*.

when I nonconsensually insert some desire into your mind through neurosurgery or sci-fi mind control: you are not forming this desire in light of normative reasons. Rather, you are simply having this desire foisted upon you through a nonrational process. And it correctly classifies my act of convincing you of some important policy *P* by nonconsensually presenting you with compelling arguments for *P*—for example, by loudly proclaiming them on my soapbox, which you happen to overhear—as permissible. Even though my influence on your thinking here is both significant and (in some sense) nonconsensual, there is nothing even *prima facie* wrong about it. The analysis at hand explains this: I cause you to affirm policy *P* by presenting you with reasons for doing so. Consequently, I do not violate your right to mental autonomy.

What about the instances of morally wrong indirect interventions—subliminal messaging and brainwashing—that Bublitz and Merkel's account fails to correctly classify?[43] First, my analysis can, I claim, explain why the above-described subliminal advertising (by the online store) is wrong. (Recall that the online store shows Flash movies to subliminally prime brand *C*, an action that causes customers to evaluate *C* more positively.) Such messaging is an attempt to *bypass* the customer's rational faculties. It succeeds, when it does, not by presenting the subject with reasons to evaluate the product in question more positively. Rather, it succeeds, when it does, by inculcating a more positive evaluation of the product through some covert and nonrational process—in this case, through priming and the familiarity bias.[44] Consequently, it counts, according to the analysis at hand, as violating the customer's right to mental autonomy.[45]

Second, my analysis can also explain why brainwashing violates people's right to mental autonomy. When I brainwash you into believing that *p*, desiring that *q*, intending end *E*, or positively evaluating *X*, I cause you to acquire these attitudes without presenting you with normative reasons for forming them. For example, the agents of the Chinese Communist Party might cause you to evaluate the Chinese State more positively by forcing you, at one of their reeducation camps, to repeatedly sing about your love for it and otherwise engage in behavior manifesting support for its tenets and ideals. Nevertheless, these do *not* constitute normative reasons for you to evaluate the Chinese State more

43 Bublitz and Merkel, "Crimes against Minds."

44 Chartrand, Huber, Shiv, and Tanner, "Nonconscious Goals and Consumer Choice." The familiarity bias is the psychological effect where subjects are more positively disposed toward familiar stimuli—including those that have been perceived only subliminally, moments before—than unfamiliar stimuli (Park and Lessig, "Familiarity and Its Impact").

45 I will consider the objection that familiarity is actually a *reason* to prefer a product, and thus that this instance of subliminal messaging does not violate the right to mental autonomy on my analysis, below in section 3.

positively—quite the opposite, in fact! (Normative reasons to positively evaluate the Chinese State would be evidence that said state was a just state, that it did not commit human rights violations, or that it had a beneficent effect upon its citizens, etc.) Consequently, according to my analysis, such brainwashing counts as violating your right to mental autonomy.

This analysis also leaves room for an attractive explanation of why it is intuitively (even) *more wrong*, so to speak, to interfere with someone's thinking through a direct intervention—such as nonconsensual neurosurgery or TMS—than through an indirect one—such as subliminal messaging. The former interventions, unlike the latter, involve a violation of the person's right to bodily autonomy in addition to their violation of the person's right to mental autonomy. Similarly, this account also allows us to class most actual instances of brainwashing as being, all things considered, more wrong than influencing someone through subliminal messaging: such instances of brainwashing (nearly always) involve concurrent violations of other rights—such as the right to liberty and free expression in the case of the Chinese State's reeducation camps.

Aside from according with our moral intuitions, this account explains why philosophers have dubbed this right "the right to mental *autonomy*." After all, and very plausibly, we act autonomously just when, and because, we act for good reasons. Consider, for example, the difference between autonomously deciding to take drugs of your own free will (to see what it felt like, say) and being compelled to take drugs, against your own better judgment and contrary to your will, by the overwhelming force of your addiction. The former action is clearly a *more* autonomous action than the second, even though both have their sources within your mind. One natural explanation of this is that only in the former case are you acting for good reasons (by your lights, at least). You—the agent—are not really the source of your action when you are overwhelmed by some impulse from which you are both alienated and which you do not take to give you good reasons to so act.[46] This suggests the following account: you act autonomously just when, and because, you act for good reasons.[47] By symmetry, we should think that you form some attitude—the belief that p, the desire that q, the intention to pursue end E—autonomously just when, and because, you form that attitude for normative reasons. *You* are the author of some attitude formation or revision—rather than a passive receiver of that attitude—just when, and because, the attitude is formed in light of reasons. If this is correct, then we have further reason to conceive the right to mental

46 Frankfurt, "Freedom of the Will."

47 Velleman, "What Happens When Someone Acts?"; Korsgaard, "Skepticism about Practical Reason" and *Self-Constitution*.

autonomy—your right to make up your mind *for yourself*, so to speak—as the right to form attitudes in light of reasons.

3. PROBLEMATIC CASES CONSIDERED

I want to finish by considering a number of problematic cases for my analysis of the right to mental autonomy. My account makes straightforward predictions about the conditions under which an intentional action violates someone's right to mental autonomy: an intentional action *A* violates someone *S*'s right to mental autonomy just when (1) *A* causes *S* to form, or revise, some attitude, and (2) *A* does not cause *S* to form, or revise, this attitude in light of reasons for forming, or revising, this attitude. Each of the problematic cases that I now discuss—interference with perceptual states, the airing of non-reason-giving advertising, the use of benevolent nudging, and subliminal messaging—presents a challenge for this account.

3.1. Perceptual States

Do perceptual states fall under the purview of the right to mental autonomy? Suppose that I wave my TMS device over your visual cortex while you are working at your desk, causing you to experience various technicolor phosphenes in your visual field. These phosphenes are momentarily distracting but swiftly disappear and have no discernible long-term effects on your thinking or experience of the world. And you know that visual experiences are mere hallucinations. Have I wronged you by so acting? Have I violated your right to mental autonomy?

My intuitions here go like this: it seems clear that I have wronged you in *some* way or another by directly inducing visual experiences within you without your consent. You could reasonably ask me to stop, and you could seek assistance from others—including, plausibly enough, the law—to make me stop if I persisted. But it is not completely obvious to me that I have violated your right to mental *autonomy* by so acting. After all, visual perceptual experiences are not states that we can *self-determine*—except indirectly by choosing to look at this or that. On the other hand, I can violate your right to bodily autonomy by physically intervening in the functioning of your stomach or kidneys, even though you have no direct control over them. By analogy, why think that the right to mental autonomy only limns a sphere of sovereignty around those mental states that you can self-determine? On balance, then, I would say that your right to mental autonomy has been violated by my actions here.

What does my analysis of the right to mental autonomy imply about this case? Well, I have caused you to form some attitudes—the visual perceptual

states in question—without presenting you with reasons to form those states. Consequently, it seems, I have violated your right to mental autonomy, according to the analysis at hand, since I have caused you to form an attitude without presenting you with reasons to do so. One initial concern with this analysis is the question of whether perceptual states can be correctly said to be (propositional) *attitudes*. However, I will simply be assuming here that they are. After all, this is the dominant view in the literature.[48] When granting that perceptual states are indeed (propositional) attitudes, they fall firmly under the purview of the right to mental autonomy on the analysis defended here.

Here is a way of understanding the problem I am raising here: perceptual states do not seem to be subject to rational norms. It does not make sense, on the face of it, to say that a perceptual state is rational or irrational. Nor does it make sense to say that the contents of the external world arrayed in my visual field give me a *reason* to enter into such and such a visual perceptual state. Consequently, when I jump in front of you and cause you to form perceptual states representing my presence, I am causing you to form attitudes—namely, your perceptual representations of me—without giving you reasons for forming them. But this means that, according to the analysis at hand, I am wronging you and violating your right to mental autonomy simply by jumping into your visual field—or, indeed, by impinging upon your sensory experience in any way, shape, or form! But this is absurd. Clearly, it is not wrong for someone else to enter your sensory field, thereby causing you to enter certain corresponding perceptual states. Something must have gone wrong in my analysis.

But what? The proponent of my analysis of the right to mental autonomy has two options here. Either she can give up the claim that perceptual states fall under the purview of the right to mental autonomy, or she can hold that perceptual states are subject to rational norms (in some sense) and that such states are formed in light of reasons (in some sense of the word "reason") when someone enters your sensory field and causes you to enter certain appropriate corresponding perceptual states. In my view, both options are reasonable. If she pursues the former, then she must explain why it is (*pro tanto*) wrong for me to induce visual phosphenes by waving my TMS device over your visual cortex in the absence of your consent without appealing to your right to mental autonomy. This could perhaps be done through reference to your right to bodily autonomy alone or through reference to some distinct right to mental *integrity* that has a broader purview than the right to mental autonomy.

48 Byrne, "Perception and Conceptual Content." For the dissenting perspective that perceptual states are not propositional attitudes, see Crane, "Is Perception a Propositional Attitude?"

However, I prefer the second option. The proposition that perceptual states are subject to norms (of some kind or another) is a compelling one. A number of philosophers hold that beliefs are subject to epistemic norms—including norms of sensitivity to the evidence and rational requirements such as consistency—because beliefs aim at being true.[49] Your perceptual states aim at accurately representing those aspects of the external world that are currently occupying your perceptual fields. By symmetry, we could hold that perceptual states are subject to norms of accuracy: the perceptual states that you *ought* to form are those that accurately represent the contents of your perceptual fields.[50] We are then in a position to say that an accurate perceptual state that is formed "in the right kind of way"—that is, through the normal sequence of perceptual processing that transforms sensory input into perceptual states—is formed *through a rational process* (in an extended sense of the phrase). At each step of the perceptual processing, you form the perceptual state that you *ought* to form, given your sensory input and prior knowledge of the causal structure of the world. (The standard view in contemporary cognitive science is that perceptual processing is a sequence of probabilistic inferences, governed by certain epistemic norms.)[51] Your sensory input then constitutes a *reason*—in some extended sense of the word—to form the appropriate corresponding perceptual states that accurately represent it. In this way, then, we can see why it makes sense to say that perceptual states are subject to rational norms and that they are formed in light of reasons (in some extended sense of the terms).

What is the significance of this? Well, it means that the proponent of my analysis of the right to mental autonomy can maintain that my act of nonconsensually inducing visual phosphenes in you with my TMS device violates your right to mental autonomy while denying that it is violated if I merely step into your visual field. The former action now counts as causing you to form an attitude without presenting you with reasons to do so, whereas the latter does not. When I step into your visual field, I cause you to form perceptual states by presenting you with reasons for forming these perceptual attitudes. In this way,

49 Velleman, "The Possibility of Practical Reason" and "On the Aim of Belief"; Cowie, "In Defence of Instrumentalism"; Buckley, "Varieties of Epistemic Instrumentalism"; Cote-Bouchard, "Two Types of Epistemic Instrumentalism."

50 For an extended defense of the view that perceptual states are subject to rational norms, see, for example, Siegel, *Rationality of Perception*. For a defense of the view that perceptual states are epistemically evaluable, see Jenkin, "Perceptual Learning and Reasons-Responsiveness."

51 Friston, "A Theory of Cortical Responses."

then, the analysis of the right to mental autonomy under consideration here can accommodate our moral intuitions on these cases.[52]

3.2. *Advertising*

A rough-and-ready distinction can be drawn between reason-giving and non-reason-giving advertising. Let us say that an advertisement for *X* is reason-giving just when the advertisement presents the viewer with reasons for purchasing or desiring *X* or makes the case that the viewer should purchase or desire *X*. In contrast, an advertisement for *X* is non-reason-giving just when it aims to cause viewers to desire or purchase *X* without presenting reasons for desiring or purchasing *X*—for example, by exploiting the "beauty sells" effect

52 On the story I have just sketched, the perceptual states that you ought to form are those that accurately represent the contents of your perceptual fields (since this is what such states aim at representing). However, this commitment looks to generate a problem for my analysis of the right to mental autonomy. After all, during visual illusions, your visual system fails to accurately represent the contents of your visual field. You, therefore, count, under such circumstances and according to my story, as failing to form the perceptual states you ought to form. In this case, according to my analysis of the right to mental autonomy, I would be wronging you by presenting you with a visual illusion that caused you to form visual perceptual attitudes without presenting you with reasons for forming those attitudes. But this is absurd. I do nothing even *prima facie* wrong when I present you with a visual illusion that causes you to have an illusory experience. However, the proponent of my analysis has a quick fix available to her for this problem. What this case tells us, I think, is that the aim of perception is *not* to accurately represent the contents of your perceptual field but rather to accurately represent the *appearances*—where the appearances can be characterized as (something like) the way the world would (normally) look (or sound, etc.) to someone occupying your vantage point. In the absence of an illusion, you ought to form those perceptual states that accurately represent the contents of your perceptual field because those contents are—or coincide with, etc.—the appearances under such circumstances. But, in the presence of an illusion, the perceptual states that you ought to form are those that accurately represent the (illusory) appearances and not those that accurately represent the actual contents of your perceptual field. This accords, I think, with our intuitions. If I experience the trompe l'oeil illusion when viewing del Caso's notable painting *Escaping Criticism*, then it does not seem like I am making a mistake or violating a norm. Of course, my perceptual attitudes have false (propositional) contents. They are representing the world as containing a boy climbing out of a framed painting, and there is no such boy in front of me. But there is the *appearance* of such a boy. And my perceptual states are accurately representing that appearance. Since perception aims at representing the appearances, I have formed the perceptual states that I ought to have formed, even though their contents ("there is a boy in front of me climbing out of a framed painting") are false. I lack the space here to flesh out this account of appearances and the aim of perception. But it strikes me as being plausible, and thus I would be warranted in appealing to it to evade the above-presented objection to my account.

and associating *X* with beautiful people.[53] Of course, many actual adverts will be both reason-giving in some respects and non-reason-giving in others—for example, an advert that accurately represents the virtues of the product while simultaneously associating it in the mind of the viewer with attractive people.

Consider, for example, the 1978 Tab cola "Beautiful People" television advertisement. Over a montage of beautiful people drinking Tab, a song with the following lyrics is sung:

> Tab, what a beautiful drink.
> Tab, for beautiful people.
> Tab, you're beautiful to me.
> Sixteen ounces and just one calorie.[54]

Although this advert does present the viewer with *some* normative reasons for purchasing Tab cola (it allows you to drink something that tastes similar to Coca-Cola, but which contains only one calorie and is thus better for your waistline), it also (quite blatantly, in my view) attempts to inculcate within the viewer a desire for Tab cola by associating it with beautiful people. The song even asserts that the drink is "for beautiful people." Advertisers have long held that "beauty sells" and have employed attractive people as endorsers, spokespeople, or models in their adverts.[55] The empirical evidence supports this contention: the physical attractiveness of the person featured in an advert increases advertiser believability, viewers' willingness to purchase, viewers' positive attitude toward the product, and the rates of actual purchase. [56]

Now, no one can seriously believe that drinking Tab cola will turn them into a beautiful person or make it the case that attractive people will want to be in relationships with them. Nevertheless, people seem to acquire a greater desire for a product when it is associated in their minds with beautiful people. The exact psychic mechanism by which this happens is contested and a matter

53 Brumbaugh, "Physical Attractiveness and Personality in Advertising"; Yin and Pryor, "Beauty in the Age of Marketing."

54 The advertisement can be viewed at https://www.youtube.com/watch?v=IrPkWNJeHzg.

55 Brumbaugh, "Physical Attractiveness and Personality in Advertising."

56 For a discussion of increased advertiser believability, see Kamins, "An Investigation into the Match-Up Hypothesis." Regarding the effect on viewers' willingness to purchase, see Kahle and Homer, "Physical Attractiveness of the Celebrity Endorser"; Petroshius and Crocker, "An Empirical Analysis of Spokesperson Characteristics." For discussion of viewers' positive attitude toward the product, see Kahle and Homer, "Physical Attractiveness of the Celebrity Endorser." Regarding effects on the rates of actual purchase, see Caballero and Solomon, "Effects of Model Attractiveness."

of debate.[57] However, insofar as this mechanism does not involve forming attitudes in light of reasons, then the airing of such an advertisement is going to count, on my analysis, as violating the viewer's right to mental autonomy. After all, it would be to inculcate a desire for a product within the viewer without presenting her with a normative reason for desiring said product or without making the case, through rational means, that she should desire this product. According to my analysis, this would make the airing of such adverts *prima facie* wrong.

Does this result accord with our moral intuitions? Non-reason-giving advertisements are everywhere. (Purely reason-giving adverts are either rare or nonexistent.) But most of us do not regard the airing of such non-reason-giving advertisements as morally wrong. (If we did, there would presumably be more of an uproar about them!) In which case, it looks like my analysis of the right to mental autonomy—if it does indeed class the airing of non-reason-giving adverts as morally wrong—has itself gone wrong: the proposition that the airing of non-reason-giving advertisements is morally wrong appears to be inconsistent with our moral intuitions concerning the matter.

However, I think the proponent of my analysis can convincingly push back against this indictment. First, it is not obvious to me that our moral intuitions support the proposition that there is nothing wrong about companies airing non-reason-giving adverts. On the contrary, it seems to me most people do think there is *something* objectionable about such advertisements. In my experience, most people, when quizzed upon the morality of non-reason-giving advertisements, such as the Tab cola commercial, describe them as being "manipulative." For example, Bublitz and Merkel describe non-reason-giving adverts as being "manipulative influences."[58] And clearly, when we describe some action as "manipulative," we mean to communicate that it is morally objectionable in some respect. Indeed, I believe that the wrongness of such manipulative actions consists (in part, at least) in the fact that they are attempts to influence someone's thinking and behavior without presenting them with reasons for thinking or behaving in the desired ways. In other words, such manipulative actions are wrong (at least, in part) because they violate the manipulated person's right to mental autonomy.

Second, the proponent of my analysis of the right to mental autonomy may be able to accommodate the proposition that there is nothing morally objectionable about the advertising industry's use of beautiful people as an instrument of persuasion by holding that the mechanism by which the "beauty effect" in

57 Brumbaugh, "Physical Attractiveness and Personality in Advertising"; Yin and Pryor, "Beauty in the Age of Marketing."

58 Bublitz and Merkel, "Crimes against Minds," 72.

advertising influences us is a rational one. For example, according to some psychologists, it could be the case that the "beauty effect" in advertising is mediated by our implicit belief that attractive people are likely to have different personality traits to the general population—in particular, that they are more trustworthy, credible, and expert in matters that they speak about or are associated with.[59] In this case, it would be rational for us to be more convinced or persuaded by an advertisement that employs beautiful people than by one that features less physically attractive people. After all, given our background implicit beliefs, beliefs perhaps supported by the statistics of our environment, the testimony of the beautiful people about the product will seem more likely to be true than the testimonies of the less physically attractive people. And this is surely a *reason* for us to be more persuaded by the testimony of such people. In this way, then, as well, the proponent of my analysis of the right to mental autonomy can resist the charge that this account incorrectly classifies the airing of non-reason-giving adverts as a morally wrong violation of viewers' right to mental autonomy.

An alternative way in which the proponent of my analysis of the right to mental autonomy may be able to accommodate the claim that there is nothing morally objectionable about advertisers' use of the "beauty sells" effect is by holding that the viewers of these adverts count as having waived their right to mental autonomy. It seems highly plausible to me that the right to mental autonomy, like many other rights, can be waived—or example, if someone suffering with long-term depression consented to neurosurgery that cured them by directly adjusting the attitudes constitutive of, or causally responsible for, their depression. Had the neurosurgeon not received the subject's consent, their actions here would have counted as violating both the subject's right to bodily autonomy and their right to mental autonomy. However, because the subject granted their consent in this case, they count as having waived both these rights, and there is no wrongdoing. This is one way in which an individual can waive their rights—that is, explicitly. But individuals can also waive their rights in a more *implicit* way—for example, by voluntarily engaging in an activity that they know will involve some probable impact on them (an impact that would count as violating their rights if they were not voluntarily engaging in the activity in question). To take a concrete and pertinent example, if I voluntarily buy a copy of *Vogue* magazine, I should expect to see beautiful people wearing the expensive watches that the advertiser wants me to buy. Plausibly enough, I may count as having implicitly waived my right to mental autonomy with respect to the influence of these adverts on my thinking. And, if I count as having waived my right to mental autonomy in the case of purchasing a *Vogue*

59 Brumbaugh, "Physical Attractiveness and Personality in Advertising."

magazine, then presumably, I should also count as having waived it when I choose to watch a television channel that I know airs advertisements. In this case, the influence of these adverts on my thinking could not count as violating my right to mental autonomy. If this is correct, then this is a second way in which the proponent of my analysis of the right to mental autonomy can resist the charge that their account incorrectly classifies the airing of non-reason-giving adverts as morally wrong.

3.3 Benevolent Nudging

Much recent work in psychology, behavioral economics, and moral philosophy has concerned the phenomenon of nudging.[60] Roughly speaking, a nudge is a way of influencing someone's actions in a predictable way by changing aspects of their "choice architecture"—that is, the context in which they choose—without forbidding any options or changing their economic incentives.[61] One concrete example of a nudge is the selection of defaults effect: people are more likely to accept the default option when presented with a range of options.[62]

The behavioral economist Richard Thaler and legal scholar Cass Sunstein have argued that nudge effects can be deployed in public policy to promote both prudent and prosocial behavior among the general public.[63] For example, the selection of defaults effect can be utilized to increase pension contributions among employees by changing the defaults on the superannuation policies to which they sign up.[64] Thaler and Sunstein dub this use of nudges in public policy "libertarian paternalism": it is *paternalistic* because individuals are manipulated into promoting their own self-interest, but it is nevertheless *libertarian* because this practice does not close off any previously existing options that people had.

Thaler and Sunstein regard the use of nudging in public policy to promote the common good as morally permissible and desirable.[65] However, it looks like such nudging—despite being benevolent—is going to violate the nudged people's right to mental autonomy, at least on the analysis of that right defended here. After all, the fact that a candidate's name is at the top of the ballot is, on the face of it, *not* a reason to vote for them. But, in light of the ballot order effect,

60 Thaler and Sunstein, *Nudge*; Wilkinson, "Nudging and Manipulation"; Doris, *Talking to Our Selves* and "Précis of Talking to Our Selves"; Levy, "Nudge, Nudge, Wink, Wink."

61 Thaler and Sunstein, *Nudge*; Levy, "Nudge, Nudge, Wink, Wink."

62 Smith, Goldstein, and Johnson, "Choice without Awareness."

63 Thaler and Sunstein, *Nudge*.

64 Levy, "Nudge, Nudge, Wink, Wink."

65 Thaler and Sunstein, *Nudge*.

it must be a cause of at least some people's decision to vote for that candidate, whether or not they know it. This means that the intentional utilization of the ballot order effect to influence people's voting constitutes an attempt to influence people's voting preferences without giving them a normative reason to adopt that voting preference. It, therefore, counts, on the analysis at hand, as an attempt to violate their right to mental autonomy. Likewise, with respect to the intention to use the selection of defaults effects in public policy, it is, plausibly enough, a cause of people selecting the default option that is nevertheless *not* a reason for them to so act. If so, then the implementation of Thaler and Sunstein's libertarian paternalism in public policy would be (at least) *pro tanto* wrong. But this is seemingly inconsistent with the moral judgment that the practice of benevolently nudging individuals to behave in prudent and prosocial ways is permissible and commendable.

How should the proponent of my analysis of the right to mental autonomy respond to this problem? I think she has a few different options available to her. First, she can hold that the libertarian paternalistic policy of implementing benevolent nudging is actually morally wrong on the grounds that it violates the nudged individual's right to mental autonomy. Support for this stance comes from the great deal of anxiety expressed by philosophers, psychologists, economists, and nonacademic commentators about the use of nudging in public policy.[66] Second, the proponent of my analysis can hold that, although the use of nudging is a *wrong-making* feature of public policy because it violates the nudged individuals' right to mental autonomy, such a policy has various other good-making or right-making features—such as the fact that it promotes the prudent and prosocial behavior of nudged individuals—that collectively outweigh this wrong-making feature, such that public policy involving benevolent nudging is an all things considered permissible course of action for governments to engage in. Last, one could deny that nudges have their influence on people without giving reasons or through nonrational mechanisms. A number of philosophers have recently argued that this is the case. For example, Neil Levy argues that nudges constitute good *reasons* for the nudged subject to act in the ways that the nudges push them toward.[67] And Andreas Schmidt argues

66 Bovens, "The Ethics of Nudge"; Wilkinson, "Nudging and Manipulation"; Levy, "Nudge, Nudge, Wink, Wink"; Schmidt, "Getting Real on Rationality."

67 As Levy puts it: "Most actual and proposed nudges function by presenting reasons to agents. They often present higher-order evidence, and higher-order evidence is evidence. It is, of course, rational to guide our decisions and our beliefs in the light of evidence. There is no reason to think, therefore, that most nudges bypass reasoning" ("Nudge, Nudge, Wink, Wink," 297). Consider again the selection of defaults effect (the phenomenon whereby people are more likely to accept the default option when presented with a range

that nudging not only works through rational mechanisms but overall promotes the rational agency of the nudged individuals.[68] In these ways, then, the proponent of my analysis of the right to mental autonomy can resist the charge that it incorrectly classes the use of nudging in libertarian paternalistic public policy as morally impermissible.

3.4. Subliminal Messaging Again

Recall Bublitz and Merkel's example of subliminal messaging: an online store shows Flash movies to subliminally prime brand *C*, an action that causes customers to evaluate *C* more positively, likely through a mechanism such as the familiarity bias.[69] Intuitively, there is something morally objectionable about subliminally influencing customers' preferences in such a manner. And, very plausibly, this course of action is morally objectionable, or wrong, because it violates the customers' right to mental autonomy.[70] My analysis of the right to mental autonomy looks like it will be able to accommodate this observation. After all, the fact that you have seen some brand before or are familiar with it is not, on the face of it, a reason to prefer it. (Rather, the reasons to prefer some particular brand include, for example, the fact that products from that brand have been found to be satisfactory or good in prior experience, etc.)

One objection to this conclusion arises out of the thought that familiarity may actually be a reason to prefer a product. After all, the familiarity bias is—very plausibly—a useful heuristic, one upon which it is rational to rely, given the statistics of our environment. If I want to buy—say—some shampoo, it is rational for me to prefer the familiar brand because familiarity correlates with wide usage, and wide usage indicates that something is a satisfactory product. Given this, it is reasonable to think that the familiarity of a product really is a reason to prefer it. (Advertising might be thought to sever the correlation between familiarity and wide usage to some degree—but that does not mean that relying upon familiarity is not rational or that the familiarity of a product is not a reason to prefer it since the familiarity of a product would still be

---

of options). Very plausibly, Levy suggests, the fact that some certain option has been selected to be the default option is a *recommendation* of that option. And recommendations are reasons to choose some option (at least when they are given by a reliable source). Consequently, Levy concludes, the nudge at work in the selection of defaults effect—the fact that some option is the default—is a reason to pick the default option.

68 Schmidt, "Getting Real on Rationality."

69 On subliminal messaging, see Bublitz and Merkel, "Crimes against Minds." On the familiarity bias, see Chartrand, Huber, Shiv, and Tanner, "Nonconscious Goals and Consumer Choice."

70 Bublitz and Merkel, "Crimes against Minds."

(some) evidence of its wide usage. Furthermore, the fact that a company has the resources to advertise is evidence of its financial success, and that constitutes (some) evidence that they make satisfactory products.) Now, when granting that the familiarity of a product is indeed a reason to prefer it, the fact that the subliminal priming by the online store operates via the familiarity bias means that the primed customers' newfound preference for brand *C* is an attitude that has been formed in light of a reason. Consequently, this subliminal influence on their thinking does not count as violating their right to mental autonomy on the analysis of that right defended here since their preference was formed in light of a reason. But this conflicts with our moral intuitions about the case, which suggest that this piece of subliminal influence involved some wrongful rights violation—most plausibly, a violation of the customers' right to mental autonomy. After all, the business behind the online store is tampering with your preferences without your even being aware of them so acting!

What is the significance of this? Well, it suggests that there is something lacking with the analysis of the right to mental autonomy developed here. After all, according to that account, your right to mental autonomy can only be violated if you are (intentionally) caused by some third party to form an attitude in a way that does *not* involve you forming that attitude in light of a normative reason. But, in the case of subliminal influence at hand, it looks like the customers are having their right to mental autonomy violated even though they are being (intentionally) caused by some third party to form an attitude in light of a normative reason. But this means that an attitude can be formed in light of reasons but still be formed in a way that constitutes a violation of the subject's right to mental autonomy. And this is contrary to the entailments of the account developed here—which has it that someone forming an attitude in light of a reason is sufficient for their *not* having their right to mental autonomy violated. If this is correct, then the account of the right to mental autonomy promoted here can, at best, constitute a partial analysis of that right—one that articulates a merely necessary condition on the permissible ways of influencing someone's thinking, not a necessary and sufficient condition.

In what remains of this paper, I will augment my analysis of the right to mental autonomy such that it can accommodate the problematic case at hand and other structurally similar cases. My working hypothesis in this paper has been that your right to mental autonomy is your right to form attitudes in light of reasons. My augmented version of this account is that your right to mental autonomy is your right to form attitudes in light of *overt* reasons. The permissible ways of influencing someone's thinking are partitioned from the impermissible ways by the fact that they involve causing someone to form an attitude in light of an overt reason.

What do I mean by an *overt* reason? Let us say that an overt reason is a reason that would cause you to form an attitude in light of it, if it did, through a route that did not circumnavigate your awareness or consciousness. But what does it take for something to circumnavigate consciousness? What precisely does this mean? We can fruitfully characterize what it is for something to enter, or circumnavigate, consciousness through appeal to some concrete examples. Subliminal influences—such as the before-mentioned subliminal primes—are a paradigm case of phenomena that have their effect on you while circumnavigating consciousness. Consequently, subliminal influences, even if they are reasons to form the attitudes that they cause and have their effect on the mind through a rational process, are nevertheless not *overt* reasons to form these attitudes. Attempts to influence someone's thinking through subliminal influences therefore count as violating their right to mental autonomy, on my augmented analysis, and thus constitute *pro tanto* wrongs. This accords with our moral intuitions about these cases.

What is an example of a consideration that does not circumnavigate consciousness and is thus apt to constitute an overt reason? Consider, for example, your act of informing me over the phone that you have dyed your hair black. Suppose I then form the belief that your hair is now dyed black in light of this testimony. My experience of your testimony is conscious. Your testimony, therefore, does not count as circumnavigating my awareness. It thus counts as an example of an *overt reason* for me to form the belief that your hair is now dyed black. Consequently, my causing you to form the belief that I have dyed my hair black by my telling you over the phone that this has happened does not constitute a violation of your right to mental autonomy (irrespective of whether I am actually telling the truth). This accords with our moral intuitions about the matter.

Your right to mental autonomy, then, on this augmented version of my account, is your right to form attitudes in light of overt reasons—that is, in light of reasons that have their influence on your thinking without circumnavigating your consciousness or awareness. Let us call this addition to my analysis the "daylight condition" or the "transparency condition." This augmented version of my analysis possesses all the benefits of the original version—it classes all of the above-cataloged cases of intuitively impermissible influences on someone's thinking as being (*pro tanto*) impermissible, and the before-cataloged cases of intuitively permissible influences as being permissible—while also correctly classifying cases of subliminal influences as being (*pro tanto*) impermissible (regardless of whether or not these subliminal influences constitute reasons). It therefore looks to be accommodating our moral intuitions better than my

first pass at an analysis of the right to mental autonomy. This constitutes a strong reason to prefer it.[71]

One important upshot of this account—that I unfortunately lack the space here to properly unpack—concerns its significance for debates over the ethics of nudging. As we saw before, some have criticized the use of benevolent "libertarian paternalistic" nudging by governments on the grounds that it violates agents' autonomy.[72] And others have criticized the use of nudging by Big Tech surveillance capitalists (to prompt their users into interacting more with their products or platforms) on the exact same grounds.[73] An influential rebuff to these critiques comes from those who have argued that nudges are reasons, or that they operate through rational mechanisms and actually overall promote the rational agency and autonomy of the nudged individuals.[74] By the lights of my first analysis of the right to mental autonomy, the use of nudges will not violate nudged people's right to mental autonomy if nudges are reasons. However, my augmented account may have the resources to imply that the use of nudging will violate individuals' right to mental autonomy even when granting that nudges are reasons. After all, nudges are standardly covert. (Agents are typically not aware that they are being nudged.) The question now arises as to whether nudges influence us via mechanisms that circumnavigate our awareness of consciousness in the above-described way. If they do, then the intentional use of nudging to influence people—benevolently or otherwise—will constitute an (attempted) violation of their right to mental autonomy, even if nudges are reasons. This may be a way of resurrecting the above ethical critique of the use of nudging by the government and Big Tech companies, etc., from Levy and Schmidt's rejoinder that nudges, very plausibly, constitute reasons and operate through rational mechanisms. Unfortunately, I lack the space here to further develop or evaluate this line of thought.

4. Conclusion

I have argued that we have a right to mental autonomy and that this right is correctly analyzed as our right to form attitudes in light of (overt) reasons. Once understood this way, we can see why this right protects us against all (nonconsensual) "nonrational" interference with our thinking—including

71 Kagan, *Normative Ethics*.

72 Bovens, "The Ethics of *Nudge*"; Wilkinson, "Nudging and Manipulation."

73 Zuboff, *The Age of Surveillance Capitalism*.

74 For argument that nudges are reasons, see Levy, "Nudge, Nudge, Wink, Wink"; see also Schmidt, "Getting Real on Ratinality."

nonconsensual neurosurgery, pharmacological manipulations, sci-fi mind control, subliminal messaging, and non-reason-giving advertising or nudging. Rather, the only fully permissible ways to seek to influence someone's thinking—those ways that involve no violation of the right to mental autonomy—are through methods that seek to engage their rational faculties without bypassing their awareness. This result, I claimed, accords with our moral intuitions concerning the matter.

Trinity College Dublin
william.je.ratoff@gmail.com

REFERENCES

BBC News. "Who Are the Uyghurs and Why Is China Being Accused of Genocide?" May 24, 2022. https://www.bbc.com/news/world-asia-china-22278037.

Bovens, Luc. "The Ethics of *Nudge*." In *Preference Change: Approaches from Philosophy, Economics, and Psychology*, edited by Till Grüne-Yanoff and Sven Ove Hansson, 207–19. Berlin: Springer Science and Business Media, 2009.

Brumbaugh, Anne M. "Physical Attractiveness and Personality in Advertising: More Than Just a Pretty Face?" *Advances in Consumer Research* 20, no. 1 (1993): 159–64.

Bublitz, Jan Christoph. "Why Means Matter: Legally Relevant Differences between Direct and Indirect Interventions into Other Minds." In *Neurointerventions and the Law: Regulating Human Mental Capacity*, edited by Nicole A. Vincent, Thomas Nadelhoffer, and Allan McCay, 49–88. New York: Oxford University Press, 2020.

Bublitz, Jan Christoph, and Reinhard Merkel. "Crimes against Minds: On Mental Manipulations, Harms and a Human Right to Mental Self-Determination." *Criminal Law and Philosophy* 8, no. 1 (January 2014): 51–77.

Buckley, Daniel. "Varieties of Epistemic Instrumentalism." *Synthese* 198, no. 10 (October 2021): 9293–313.

Byrne, Alex. "Perception and Conceptual Content." In *Contemporary Debates in Epistemology*, edited by Ernest Sosa and Matthias Steup, 231–50. Oxford: Blackwell, 2005.

Caballero, Marjorie J., and Paul J. Solomon. "Effects of Model Attractiveness on Sales Response." *Journal of Advertising* 13, no. 1 (1984): 17–23.

Chartrand, Tanya L, Joel Huber, Baba Shiv, and Robin J. Tanner. "Nonconscious Goals and Consumer Choice." *Journal of Consumer Research* 35, no. 2

(August 2008): 189–201.
Côté-Bouchard, Charles. "Two Types of Epistemic Instrumentalism." *Synthese* 198, no. 6 (June 2021): 5455–75.
Cowie, Christopher. "In Defence of Instrumentalism about Epistemic Normativity." *Synthese* 191, no. 16 (November 2014): 4003–17.
Crane, Tim. "Is Perception a Propositional Attitude?" *Philosophical Quarterly* 59, no. 236 (July 2009): 452–69.
Doris, John M. "Précis of *Talking to Our Selves: Reflection, Ignorance, and Agency*." *Behavioral and Brain Sciences* 41 (2018): e36.
———. *Talking to Our Selves: Reflection, Ignorance, and Agency.* Oxford: Oxford University Press, 2015.
Douglas, Thomas, and Lisa Forsberg. "Three Rationales for a Legal Right to Mental Integrity." In *Neurolaw: Advances in Neuroscience, Justice and Security*, edited by Sjors Ligthart, Dave van Toor, Tjis Kooijmans, Thomas Douglas, and Gerben Meynen, 179–201. London: Palgrave Macmillan, 2021.
Frankfurt, Harry G. "Freedom of the Will and the Concept of a Person." *Journal of Philosophy* 68, no. 1 (January 1971): 5–20.
Friston, Karl. "A Theory of Cortical Responses." *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 360, no. 1456 (2005): 815–36.
Haitiwaji, Gulbahar, and Rozenn Morgat. "Our Souls Are Dead: How I Survived a Chinese 'Re-education' Camp for Uyghurs." *Guardian,* January 12, 2021. https://www.theguardian.com/world/2021/jan/12/uighur-xinjiang-re-education-camp-china-gulbahar-haitiwaji.
Hart, H. L. A. *The Concept of Law*. Oxford: Clarendon Press, 1994.
Ienca, Marcello, and Roberto Andorno. "Towards New Human Rights in the Age of Neuroscience and Neurotechnology." *Life Sciences, Society, and Policy* 13, no. 5 (2017).
Jenkin, Zoe. "Perceptual Learning and Reasons-Responsiveness." *Noûs* 57, no. 2 (June 2023): 481–508.
Kagan, Shelly. *Normative Ethics*. New York: Routledge, 1998.
Kahle, Lynn R., and Pamela M. Homer. "Physical Attractiveness of the Celebrity Endorser: A Social Adaptation Perspective." *Journal of Consumer Research* 11, no. 4 (March 1985): 954–61.
Kamins, Michael A. "An Investigation into the 'Match-Up' Hypothesis in Celebrity Advertising: When Beauty May Be Only Skin Deep." *Journal of Advertising* 19, no. 1 (1990): 4–13.
Korsgaard, Christine M. *Self-Constitution: Agency, Identity, Integrity*. Oxford: Oxford University Press, 2009.
———. "Skepticism about Practical Reason." *Journal of Philosophy* 83, no. 1

(January 1986): 5–25.
Levy, Neil. "Nudge, Nudge, Wink, Wink: Nudging is Giving Reasons." *Ergo* 6, no. 10 (2019): 281–302.
Locke, John. *Two Treatises of Government*. London: Awnsham Churchill, 1689.
Markovits, Julia. *Moral Reason*. Oxford: Oxford University Press, 2014.
Narveson, Jan. *The Libertarian Idea*. Peterborough, Ontario: Broadview, 2001.
Nozick, Robert. *Anarchy, State, and Utopia*. New York: Basic Books, 1974.
O'Barr, William M. "'Subliminal' Advertising." *Advertising and Society Review* 6, no. 4 (2005).
Park, C. Whan, and V. Parker Lessig. "Familiarity and Its Impact on Consumer Decision Biases and Heuristics." *Journal of Consumer Research* 8, no. 2 (September 1981): 223–31.
Petroshius, Susan M., and Kenneth E. Crocker. "An Empirical Analysis of Spokesperson Characteristics on Advertisement and Product Evaluations." *Journal of the Academy of Marketing Science* 17, no. 3 (June 1989): 217–25.
Raz, Joseph. "On the Nature of Rights." *Mind* 93, no. 370 (April 1984): 194–214.
Schmidt, Andreas T. "Getting Real on Rationality: Behavioral Science, Nudging, and Public Policy." *Ethics* 129, no. 4 (July 2019): 511–43.
Siegel, Susanna. *The Rationality of Perception*. Oxford: Oxford University Press, 2017.
Smith, N. Craig, Daniel G. Goldstein, and Eric Johnson. "Choice without Awareness: Ethical and Policy Implications of Defaults." *Journal of Public Policy and Marketing* 32, no. 2 (2013): 159–72.
Thaler, R. H., and Cass R. Sunstein. *Nudge: Improving Decisions about Health, Wealth, and Happiness.* New Haven: Yale University Press, 2008.
Thomson, Judith Jarvis. *The Realm of Rights*. Cambridge, MA: Harvard University Press, 1990.
Velleman, J. David. "On the Aim of Belief." In *The Possibility of Practical Reason*, 244–81. Oxford: Oxford University Press, 2000.
———. "The Possibility of Practical Reason." *Ethics* 106, no. 4 (July 1996): 694–726.
———. "What Happens When Someone Acts?" *Mind* 101, no. 403 (July 1992): 461–81.
Wilkinson, T. M. "Nudging and Manipulation." *Political Studies* 61, no. 2 (June 2013): 341–55.
Yin, Bingqing, and Susie Pryor. "Beauty in the Age of Marketing." *Review of Business and Finance* Case *Studies* 3, no. 1 (2012): 119–32.
Zuboff, Shoshana. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. New York: Public Affairs, 2019.

Journal of Ethics and Social Philosophy
VOL. 27, NO. 2 · APRIL 2024
https://doi.org/10.26556/jesp.v27i2.3060

THE POINT OF BLAMING AI SYSTEMS

Hannah Altehenger and Leonhard Menges

One key feature of both our present age and the decades to come is that we face the increasing arrival of powerful AI in many important domains of our lives. Many authors have argued that this raises new and deep ethical challenges.[1] One of the philosophically most interesting is, as Christian List has recently put it, that "we may have to adjust some of our conventional anthropocentric approaches to morality."[2] Or in other words, the arrival of powerful AI suggests "that our moral theories and regulatory frameworks should be 'future-proofed'"—i.e., reassessed in the face of these developments.[3]

One core part of our regulatory frameworks that can be found almost universally across human societies is our practice of praise and blame.[4] Praising others for (what we perceive to be) commendable behavior and blaming them for (what we perceive as) transgressions are key forms that our "regulatory interactions" can take.

Hence, one important part of "future-proofing" our extant regulatory frameworks in the face of the increasing arrival of powerful AI is to ask whether it makes sense to *extend* these practices and, in particular, our practice of *blame* to these systems.[5] This is the question that we shall focus on in this paper.

Our main claim is that, contrary to what one might initially think, this question should be answered in the affirmative—i.e., we shall argue that it can make sense to blame AI systems. More specifically, we shall defend the claim that we have a *pro tanto* reason to extend our blaming practices to these systems.

To support this claim, we shall proceed as follows: in section 1, we will present in more detail the claim that the increasing presence of AI systems creates a need for future-proofing our regulatory practices. We contend that

1 For overviews, see Noorman, "Computing and Moral Responsibility"; Müller, "Ethics of Artificial Intelligence and Robotics."

2 List, "Group Agency and Artificial Intelligence," 1215.

3 List, "Group Agency and Artificial Intelligence," 1240.

4 See Sommers, *Relative Justice*.

5 Like many other philosophical works on "regulatory practices," we shall focus on blame rather than praise.

future-proofing blame is one key element in such an endeavor that List himself has overlooked, and we also clarify how our paper relates to the so-called responsibility gap debate, which has recently received much attention in AI ethics.[6] In the main part of the paper (section 2), we first discuss how to proceed to answer the question of whether it makes sense to extend our blaming practices to AI systems. We propose that this issue shall be settled by focusing on the *functions* that these practices fulfill. We then argue that our blaming practices can fulfill several valuable functions when targeting AI systems, which suggests that we have at least a *pro tanto* reason to extend those practices to these systems. Before concluding, we will discuss how the issue of whether it makes sense to blame AI systems relates to the issue of whether AI systems can be blameworthy (section 3).

1. PRELIMINARIES

The claim that the increasing arrival of AI gives rise to deep ethical challenges is a commonplace. Things become more interesting, though, once we ask why *exactly* the ethical challenges raised by AI systems seem to be of a more fundamental nature than, say, the challenges raised by the increasing reliance on "traditional" machines since the Industrial Revolution. Here is what we take to be the most convincing answer to these questions: unlike the machines that arrived on the scene during the Industrial Revolution, we now face the increasing arrival of systems that have the ability to (i) operate relatively autonomously in largely uncontrolled environments and (ii) make "high-stakes" decisions.[7] List illustrates this point in the following passage:

> If a system has only limited capacities, such as a robotic floor cleaner or a pre-programmed factory robot, or if its use has no serious spill-over effects beyond a restricted environment, as in the case of an automated train in a tunnel, then it does not give rise to qualitatively novel risks, compared to earlier technologies. . . . By contrast, if an AI system operates relatively freely in a largely uncontrolled environment, as in the case of a driverless car or a fully autonomous drone, or if it can make high-stakes decisions on its own, as in the case of some medical, financial, and military systems, then the societal implications are qualitatively novel.

6 See, e.g., Matthias, "The Responsibility Gap"; Sparrow, "Killer Robots"; Himmelreich, "Responsibility for Killer Robots"; Köhler, "Instrumental Robots"; Nyholm, *Humans and Robots*, ch. 3; Danaher, "Tragic Choices."

7 List, "Group Agency and Artificial Intelligence," 1218; see also, e.g., Müller, "Ethics of Artificial Intelligence and Robotics," sec. 1.2; Nyholm, *Humans and Robots*, ch. 2.

> *We are then dealing with artefacts as genuine decision-makers, perhaps for the first time in human history.*[8]

If the development of novel AI systems were restricted to sophisticated vending machines or systems that can autonomously assemble IKEA furniture, then few of us would feel that the ethical challenges these systems raise were qualitatively novel. But the development of AI systems also includes entities like driverless cars, autonomous air vehicles, medical helper AI systems, diagnostic devices, and financial trading systems. Unlike the former, these systems all operate in "high-stakes contexts," where the occurrence of some amount of serious harm seems inevitable. However, due to their increasingly autonomous mode of functioning, it will be increasingly difficult, if not impossible, to hold some human being responsible for that harm.

According to List, all of this "suggests that our moral theories and regulatory frameworks should be 'future-proofed'"—i.e., reassessed in the face of these developments.[9] List also provides a sketch of how AI systems could be held responsible for the harm they cause:

> The proposed form of AI responsibility may, in turn, have to be underwritten by certain assets, financial guarantees, and/or insurance, so that, in the event of a harm, the system or its legal representatives can be made to pay appropriate fines and compensation.[10]

The passage just quoted arguably captures *some* of our practices of holding each other responsible. However, imposing fines and demanding compensation for perceived transgressions clearly does not exhaust these practices. Another crucial practice that seems to dominate our everyday moral interactions and that List's account of holding AI responsible omits is *blame*. Hence, future-proofing our responsibility practices in a *comprehensive* way would also require reassessing our blaming practices, and, more specifically, asking whether it makes sense to extend these practices to AI systems. It is this task that our paper focuses on.

However, before moving on to this task, two clarifications are in order. First, we need to clarify what kind of AI systems we are interested in. Second, we need to explain how our main concern relates to the so-called responsibility gap debate.

8 List, "Group Agency and Artificial Intelligence," 1218 (emphasis added).

9 List, "Group Agency and Artificial Intelligence," 1240.

10 List, "Group Agency and Artificial Intelligence," 1230.

Regarding the first issue, we are merely interested in those AI systems that qualify as intentional agents in a minimal sense of the term. Following List, we shall assume that minimal intentional agency requires

1. "Representational states (which encode an entity's 'beliefs' about how things are),"
2. "Motivational states (which encode its 'desires' or 'goals' as to how it would like things to be)," and, finally,
3. "A capacity to interact with its environment on the basis of these states so as to 'act' in pursuit of its desires or goals in line with its beliefs."[11]

We shall furthermore assume that many already-existing and even more near-future AI systems meet the conditions for minimal intentional agency.[12]

Some may object that no further discussion is needed, once this assumption is in place: (minimal) intentional agency, the objection goes, is sufficient, both for blameworthiness and for its making sense to be the target of blame.[13]

We have two replies to this objection. One says that there are many entities which fulfill the above conditions for minimal intentional agency but which are such that, intuitively, it seems to be an open question whether they fulfill the conditions for blameworthiness or whether blaming them makes sense. Toddlers, people with severe cognitive disabilities, psychopaths, as well as many nonhuman animals qualify as minimal intentional agents (given the above understanding of minimal intentional agency). Intuitively, however, it seems at least to be an open question whether they satisfy the conditions for blameworthiness and whether blaming them makes sense.

Our second reply is that the distinction between minimal intentional agency on the one hand and the kind of agency that is necessary for blameworthiness or for its making sense to be the target of blame on the other is not only intuitive; it is also one that is commonly made in different philosophical debates.

11 List, "Group Agency and Artificial Intelligence," 1219.

12 Let us forestall a possible misunderstanding: in presupposing that many already-existing and even more near-future AI systems have representational states and motivational states, it may seem that we have made a highly contested assumption—namely, that many current and even more near-future AI systems "have minds." But this way of putting the matter is misleading. To be sure, the claim that many existing and near-future AI systems have belief- and desire-like states seems to entail that they have minds *in a minimal sense*. However, this should not be confused with the claim that such systems can have full-fledged, human-level minds, complete with phenomenally conscious states, the capacity for self-consciousness, verbal abilities, emotions, and a rich network of diverse propositional attitudes. That many or, indeed, any existing and near-future AI systems have minds of this kind is *not* what we are presupposing. Many thanks to Peter Schulte for helpful advice on this point.

13 Many thanks to an anonymous reviewer for urging us to address this objection.

Authors who are skeptical about blameworthiness or the justifiability of blame, for example, are, typically, not skeptical about (minimal) intentional agency. Consider Derk Pereboom's skepticism about a specific kind of blameworthiness—what he calls blameworthiness in the "basic desert sense."[14] Pereboom argues that both luck and determinism undermine the sort of agency that is necessary for this kind of blameworthiness. But he does not argue that these factors undermine (minimal) intentional agency. Similarly, many authors in AI ethics in general and the responsibility gap debate in particular share our assumption that the relevant AI systems, i.e., those systems that are claimed to generate responsibility gaps, are intentional or, as it is also sometimes put, "autonomous" agents in a minimal sense of these terms.[15] Those authors assume or argue that AI systems are agents in some minimal sense and contend that it is, nonetheless, inappropriate or even impossible to blame them when they cause unjustified harm.

The considerations offered in the preceding should be enough to show that the above objection fails: even if one assumes that an entity satisfies the conditions for minimal intentional agency, it is still an interesting, open question whether it satisfies the conditions for blameworthiness or whether blaming it makes sense.

Let us turn next to our second clarification—namely, how our paper relates to the responsibility gap debate. We shall be primarily concerned with the issue of whether it makes sense *to blame* AI systems rather than with the issue of whether AI systems can be *blameworthy*. We would like to emphasize that these are distinct questions. For it could turn out that AI systems can be blameworthy, but it does not make sense to blame them, and it could also turn out that it makes sense to blame AI systems even if they cannot be blameworthy.[16] Many authors in the responsibility gap debate ask who, if anyone, can be *blameworthy (responsible)* if an AI system causes some unjustified harm.[17] The focus of our paper will thus be different from theirs. However, some authors within this debate are (also) concerned with the question of whether we can *blame* AI

14 Pereboom, *Free Will, Agency, and Meaning in Life*.

15 See, e.g., Sparrow, "Killer Robots," 65, 74; Danaher, "Robots, Law, and the Retribution Gap," 301; Nyholm, "Attributing Agency to Automated Systems," 1207–9; Burri, "What Is the Moral Problem with Killer Robots?" 165–66; Himmelreich, "Responsibility for Killer Robots," 734; Köhler, "Instrumental Robots," 3124; Königs, "Artificial Intelligence and Responsibility Gaps," 36.

16 We shall expand on the sense of "making sense" that is at issue here in the next section. Moreover, we will take up the issue of AI *blameworthiness* again in section 3.

17 See, e.g., Matthias, "The Responsibility Gap"; Sparrow, "Killer Robots"; Himmelreich, "Responsibility for Killer Robots"; Köhler, "Instrumental Robots"; Nyholm, *Humans and Robots*, ch. 3; Kiener, "Can We Bridge AI's Responsibility Gap at Will?"

systems.[18] In particular, these theorists have argued that blaming AI systems is *not possible*. An argument to this conclusion says, roughly, that blaming is a form of harming and that it is impossible to harm AI systems.[19] We will discuss this particular line of thinking in section 2.1. In general, though, the remainder of this paper should make clear that we disagree with the claim that it is impossible to blame AI systems, and the considerations that we shall offer in the next section can be read as an argument against this view.

2. HOW BLAMING AI SYSTEMS MAKES SENSE

To a first approximation, *blame* can be characterized as "a reaction to something of negative normative significance about someone or their behavior."[20] There are many controversies surrounding the exact nature of blame.[21] However, for the purposes of this paper, it will be best to stay neutral on this issue. Together with many theorists working on blame, we shall assume that manifestations of blame can be quite diverse. Among other things, they can take the form of openly expressed anger, unexpressed feelings of resentment, or even seemingly dispassionate acts of relationship modification (e.g., calmly unfriending someone on one's social media account).[22]

With this minimal understanding of blame in place, let us ask next how we should proceed in order to settle the issue of whether it makes sense to extend our blaming practices to AI systems. We propose that the best answer to this question is to focus on blame's *functions*. Or, somewhat more precisely, proceeding from the assumption (to be substantiated in a moment) that our blaming practices have several valuable functions, we put forward the following suggestion: to decide whether it makes sense to extend our blaming practices to AI systems, we should ask whether these practices can still fulfill enough of their valuable functions when targeting AI systems.

Our suggestion relies on two background assumptions which, however, seem very plausible (as we shall argue next). The first is as follows:

18 Many thanks to an anonymous reviewer for bringing this point to our attention.

19 See Solum, "Legal Personhood for Artificial Intelligences," 1245–46; Sparrow, "Killer Robots"; Danaher, "Robots, Law and the Retribution Gap."

20 Tognazzini and Coates, "Blame."

21 For overviews, see Coates and Tognazzini, "Nature and Ethics of Blame" and "Contours of Blame"; Tognazzini and Coates, "Blame"; Smith, "Blame and Holding Responsible"; Menges, "Blaming."

22 See, e.g., Smith, "Blame and Holding Responsible," sec. 2.

1. Our blaming practices fulfill several valuable functions.[23]

As mentioned previously, there is much controversy about the exact nature of blame. However, most theorists seem to agree that blame has certain valuable functions or, as it is more commonly expressed, "has a point."[24] We shall elaborate on what these functions are in the remainder of this section. For now, we merely want to stress that the assumption that our blaming practices fulfill certain valuable functions seems to be widely shared among theorists working on blame.[25] (Note that if blame possessed no valuable functions, it would be hard to understand why so many philosophers try to show that blaming people can be appropriate even if determinism is true—if it "had no point," then everybody should be happy to get rid of it.)

Our second background assumption can be put as follows:

2. If our blaming practices would still fulfill their valuable functions in targeting entities of type *x* (or, at least, enough of these functions for them to still "have a point"), then we have a *pro tanto* reason to extend these practices to entities of type *x*.

Claim 2 seems very intuitive, at least assuming that one does not read into it something stronger than it says. Claim 2 does *not* say that we ought, all things considered, to extend our blaming practices to entities of type *x* if, in targeting entities of type *x*, our blaming practices would fulfill (enough of their) valuable functions.[26] Nor does it say that we would have sufficient reason to do so. Instead, claim 2 makes a much more modest claim—namely that, in this case, we would have a *pro tanto* reason to extend these practices to entities of type *x* (which then may or may not be outweighed by other reasons against such an extension).

23 To clarify, we use the term "function" in a minimal sense of "what a thing does," and, consequently, the term "valuable functions" in the sense of "the positive effects a thing has." Or, to put the same point in a slightly different and somewhat colloquial manner: what we are interested in when we talk about the "valuable functions" of our blaming practices are the "cool things that blame does for us." We are grateful to Sebastian Köhler for urging us to be clearer on this point and for suggesting that we express this point in this manner.

24 See Watson, "Responsibility and the Limits of Evil," 230. See also Macnamara, "Blame, Communication, and Morally Responsible Agency," 219; Fricker, "What's the Point of Blame?"; Wang, "Communication Argument."

25 Note that the assumption that blame fulfills certain (valuable) functions is independent from the claim that blame can ultimately only be defined in terms of its functions (this is, roughly, the view of McKenna, "Directed Blame and Conversation"; Fricker, "What's the Point of Blame?"; and Shoemaker and Vargas, "Moral Torch Fishing"). One can accept the former assumption, while rejecting the latter.

26 Here and in the following we use the expression "enough of their valuable functions" as a shorthand for "enough valuable functions for our blaming practices to still 'have a point.'"

In the following, we shall argue that our blaming practices would fulfill several valuable functions when targeting AI systems (and clearly enough of their valuable functions to still "have a point") and that we, therefore, have at least a *pro tanto* reason to extend them to these systems.

2.1. Retribution

It may seem natural to claim that one valuable function of our blaming practices is *retribution*, i.e., that one valuable feature of these practices is that they help ensure that the guilty "get what they deserve."

Could appealing to this function support the claim that our blaming practices would fulfill valuable functions in targeting AI systems? We are skeptical about this for two reasons.[27]

First, we are skeptical about the idea that the retribution function is a *valuable* function. Our skepticism is motivated by a general antiretributivist stance, i.e., we would reject the idea *that there is something (noninstrumentally) good in a guilty party's being harmed*, which is at the very core of retributivist thinking.[28]

Second, there is reason to doubt that the retributive function could still *be fulfilled* if the blamee was an AI system.[29] After all, in order for this function to be fulfilled, it is necessary that a blaming response can in some way be *harmful* for the target, since, as was just mentioned, the idea that there is something good about a guilty party's *being harmed* is at the very core of retributivist thinking. Now, there is no difficulty seeing how a blaming response can be harmful if the target is a human being: few of us like to be blamed by others. Indeed, it often *feels quite uncomfortable*, if not *somewhat painful* to be the recipient of blame. But it is much more difficult to see how blame could harm AI systems. There is a complicated debate about the nature of harm, but it seems plausible that for something to be harmful, it must at least do one of the following: cause bad (painful) experiences, frustrate desire, set back some interest, or diminish an agent's quality of life. First, however, it is difficult to see how blaming responses should lead AI systems to have painful experiences since these systems plausibly lack phenomenal consciousness (at least those that are currently around and that will be around in the near future).[30] Second, while we are very sympathetic to the assumption that AI systems can have desires, it is difficult to see how blame, as a general matter of fact, should frustrate these desires: while it does

27 A view that may be somewhat similar to ours is expressed by Gogoshin ("Robot Responsibility and Moral Community," 9).

28 For an overview, see Walen, "Retributive Justice."

29 See also Sparrow, "Killer Robots," 71–73; Danaher, "Robots, Law, and the Retribution Gap."

30 For an argument in support of this claim, see, e.g., the reasoning put forward by List, "Group Agency and Artificial Intelligence," 1237–38.

seem plausible that the vast majority of human beings has some desire(s) which are frustrated by instances of blame, making the same assumption about AI systems would seem to require a fair amount of undue anthropomorphizing.[31] Third, it is far from clear what it means to say that AI systems have interests or a quality of life. In view of all this, it is considerably difficult to see how our blaming responses would still retain their harmful character in targeting AI systems and, consequently, how they could still fulfill their retributive function.[32]

It would be too hasty to conclude from this, though, that we have no reason to extend our blaming practices to AI systems. This is because, as the remainder of the paper will show, prospects look much brighter once we turn to further (valuable) functions of these practices.

2.2. Modification of Behavior

While the retributive function is essentially backward-looking, there is a further important function of blame that is essentially forward-looking—namely, modifying the future behavior of the blamee.[33]

In order for blame to fulfill its behavior-modification function when targeting an AI system, the latter would obviously have to possess some kind of feedback mechanism. More specifically, the system would have to be able to recognize instances of blame as such and to process them in a way that would eventually lead to behavior modification. In principle, this may happen in two ways: the first way is "classic reprogramming." Imagine that, once an AI system has "registered" a number of blaming responses directed at it, it sends a corresponding signal, which then leads to reprogramming, i.e., a human supervisor assesses these responses and, if judged appropriate, makes some fitting alterations to the system's priorities. The second way is autonomous machine learning. Imagine that after a training phase with a sufficiently large "blame database," an AI system uses further instances of blame directed at it to itself update its database with desirable responses. We are not the first to maintain that autonomous machine learning may one day lead to "blame-sensitive" AI. In particular, Dane Gogoshin and Daniel Tigard have recently contended that

31 In fact, we defend this view in our unpublished manuscript "How Robots Can Be Blameworthy" (coauthored with Peter Schulte).

32 The reasoning that we have just offered is admittedly sketchy. Hence, we do not claim to have shown that it is impossible that blame's retributive function can be fulfilled when the blamee is an AI system. The point we wish to make is a weaker one: at least for those AI systems that are currently around and that will be around in the foreseeable future, it seems much more plausible to assume that this function cannot be fulfilled than to assume that it can.

33 See, e.g., McGeer, "Civilizing Blame," sec. 2.3.

relevant reinforcement learning mechanisms may allow for the construction of AI systems that can modify their behavior in reaction to our blaming responses.[34]

There are obviously some pros and cons to both approaches and some significant technical challenges to overcome in order to implement them. However, we would like to stress, in line with the aforementioned treatments of the matter, that there do not seem to be any in-principle obstacles here. Registering instances of blame and treating them as a source of feedback ultimately just amounts to a form of learning. Hence, on the plausible assumption that learning in AI systems is possible and that further substantial progress will be made in that domain in the coming decades, it seems plausible that, at some future point at least, AI systems can be construed that can use our blame responses as a source for learning. And once this point will be reached, there do not seem to be any obstacles to the fulfillment of blame's behavior-modification function.

Interestingly, there are even respects in which the fulfillment of this function may be *easier* if the blamee is an AI system rather than a human being: first, unlike in the case of human beings, the fulfillment of blame's behavior-modification function cannot be thwarted by episodes of *akrasia*. Once a relevant episode of learning has been completed, the system will adapt its overt behavior accordingly. Second, humans sometimes respond to being blamed in destructive ways, such as counter-blaming or playing the "blame game," seeking fault elsewhere, and so on.[35] A well-programmed AI can avoid these responses.

Suppose, though, that our assessment in this section was overly optimistic and that, contrary to what we have just claimed, it is unlikely that blame can fulfill its behavior-modification function when targeting AI systems (because no or only very few AI systems will ever possess the relevant learning mechanisms). Would this mean that extending our blaming practices to AI systems would be pointless? In the remaining sections, we will argue that this would not follow. As we will show, our blaming practices have several additional valuable functions, some of which can be fulfilled surprisingly well when the blamee is an AI system.

2.3. *Conversation*

As several theorists have stressed, blame seems to possess another important function which may be somewhat less obvious than the retribution and

34 Gogoshin, "Robots as Ideal Moral Agents per the Moral Responsibility System" and "Robot Responsibility"; Tigard, "Artificial Moral Responsibility." Both Gogoshin and Tigard in turn draw on Wallach and Allan's work on artificial moral cognition in *Moral Machines*.

35 See, e.g., Pettigrove, "Meekness and 'Moral' Anger"; Pereboom, *Wrongdoing and the Moral Emotions*, ch. 1.

behavior-modification function. This is the function of initiating or sustaining conversations about the negative normative or evaluative status of what happened—henceforth referred to as "normative conversations."[36] This function can be fulfilled by open statements, but also by less explicit forms of communication (e.g., a raised eyebrow can also start a normative conversation).

A normative conversation initiated or sustained by an instance of blame can be valuable in many respects. It can give the targets of blame reasons to act differently in the future and help them to further develop their ability to respond to relevant reasons.[37] It provides an opportunity for targets of blame to explain or even justify what they did, to learn about how we perceive their conduct, and to ask for forgiveness.[38] These are important processes because we need a peaceful way to deal with the "normative ruptures" in our social webs. For instance, when we directly blame a friend for telling a mean joke about us, we start a conversation with her about what she did. We communicate that we found her behavior unacceptable and, thereby, start an exchange of our views about the reasons and values that are at issue. Ideally, she will ask for forgiveness and, thereby, try to restore our friendship.

Can blame fulfill the function of initiating or sustaining a conversation about the negative normative or evaluative status of what happened when the blamee is an AI system? Regarding current AI systems, this seems implausible. A key worry regarding these systems is that not even their designers are able to understand why they come to a certain conclusion and not to a different one.[39] In that case, having a normative conversation is impossible. We cannot converse with someone about the normative status of what they did who is unable to explain, much less justify, what they did.[40]

This situation may change in the future. A lot of energy is currently being put into theorizing about and engineering so-called transparent or explainable

36 See, e.g., Watson, "Responsibility and the Limits of Evil"; McKenna, "Directed Blame"; McGeer, "Civilizing Blame"; Macnamara, "Reactive Attitudes as Communicative Entitites"; Mason, *Ways to be Blameworthy*, ch. 5; Wang, "Communication Argument." For a similar point, see also Tigard, "Technological Answerability."

37 See, e.g., Vargas, *Building Better Beings*; McGeer, "Scaffolding Agency."

38 See, e.g., McKenna, "Directed Blame and Conversation"; Fricker, "What's the Point of Blame?"

39 See, e.g., Müller, "Ethics of Artificial Intelligence and Robotics," sec. 2.3.

40 Some may object that recent successes of large language models like ChatGPT show that normative conversations between humans and AI systems are already happening. First, however, these systems, too, cannot explain or justify how they came to their decisions. Second, it seems unclear whether they can ask for forgiveness and be forgiven. However, insofar as these are key aspects of normative conversations, there is reason to doubt that such conversations between humans and current AI systems are already possible.

AI (XAI).[41] In a nutshell, the idea is to build AI systems that allow the users, engineers, regulators, and so on to understand how and why the system comes to a certain decision or proposal.[42] Now, an XAI system in this sense is not yet a system with which one can have the same kind of normative conversation that we know from our direct interactions with human wrongdoers. That the system can make us understand why and how it comes to a decision does not yet guarantee that it understands us when we challenge its decisions, that it learns from our blame, that it asks for forgiveness, and so on. Perhaps such a fully "conversable" AI system can be engineered.[43] But independently of this, we would like to offer the following novel line of reasoning: even if the prospect of conversable AI does not turn out to be realistic and even if there will never be a fully transparent AI system, there would still be an important sense in which blame can fulfill its function of initiating and sustaining a normative conversation when the blamee is an AI system.

Our starting point is the observation that, in everyday life, we often initiate or sustain a conversation about the negative normative or evaluative status of what people did who can neither explain nor justify their conduct—for example, when we discuss our histories. In our communities, it is important for us to converse with each other about the wrongdoings of, for example, American slaveholders or German Nazis, despite the fact that the transgressors—given that they are no longer living—are unable to explain or justify their behavior or to ask for forgiveness.[44] The value of these conversations cannot be that it helps the transgressors develop their rational abilities, change their behavior, or understand what we think about what they did. Rather, the value of these conversations lies in *helping us today*. That is, these conversations help us to

41 Floridi et al., "AI4People"; Bathaee, "The Artificial Intelligence Black Box and the Failure of Intent and Causation"; Langer et al., "What Do We Want from Explainable Artificial Intelligence (XAI)?"; Baum et al., "From Responsibility to Reason-Giving Explainable Artificial Intelligence." For a critical discussion of the need for XAI in the medical sector, see London, "Artificial Intelligence and Black-Box Medical Decisions." In this context, see also Daniel Tigard's recent suggestion that we should design (what he calls) technologically answerable systems, i.e., systems that have the ability to provide their users with answers as to why a certain behavioral output occurred ("Technological Answerability").

42 One way to achieve this is to equip the AI system with an "ethical black box" analogous to a flight data recorder that records its decision-making process. See, e.g., Winfield and Jirotka, "Case for an Ethical Black Box."

43 On the issue of "conversable" AI systems, see also List, "Group Agency and Artificial Intelligence," 1228–32.

44 A parallel argument could be run for human agents whose psychological makeup is such that playing a constructive part in a normative conversation is very difficult (if not impossible) for them—e.g., agents with narcissistic personality disorder.

develop our capacity to respond to relevant moral reasons, to not do what these transgressors did, and to understand how the world perceives their conduct. These are important issues. To converse about the normative or evaluative status of what certain transgressors did thus plays important roles even if these transgressors cannot be part of the conversation.

The same can be true when the blamee is an AI system. Even if we cannot converse with a self-driving car that prioritizes driving its customers home quickly over protecting the safety of pedestrians, we can converse with each other about the normative or evaluative status of what the car does. This can play important roles in developing our normative reasoning abilities, changing future conduct, and sharing how we perceive the normative and evaluative world.

Thus, regardless of whether XAI will ever be fully realized and even if AI systems never achieve the status of conversable entities, there still is a sense in which blame can fulfill its valuable function of initiating and sustaining normative conversations when the blamee is an AI system.

2.4. Protest

As several theorists have argued, another important function of blame is to enable a specific form of *moral protest*.[45] The core idea here is that, by blaming another party, we can "stand up for [ourselves]" (or others) and "put something important on record"—namely, roughly speaking, that the way the other party has treated us (or the third person we are standing up for) was not okay.[46] Or, as Angela Smith has put it, one key aim (or function) of our blaming responses is to "*register* the fact that the person wronged did not deserve such treatment" and "to prompt moral recognition and acknowledgment of this fact on the part of the wrongdoer and/or others in the moral community."[47]

The latter qualification is important since it highlights the fact that the protest function of blame can be fulfilled even if it is unlikely, or even impossible, to gain moral recognition from the transgressor herself and, we may add, even if the transgressor is unlikely, or even unable, to modify her behavior in response to our blame. Indeed, according to Matt Talbert, "such protest is *meant largely for the protester and for his fellow sufferers,*" and its "intelligibility depends [not] on whether anyone will be converted to a better moral point of view."[48] By protesting, we make it clear to ourselves and those around us that we are standing

45 See, e.g., Talbert, "Moral Competence"; Smith, "Moral Blame and Moral Protest"; Pereboom, *Wrongdoing and the Moral Emotions*, ch. 2.

46 Talbert, "Moral Competence," 106.

47 Smith, "Moral Blame and Moral Protest," 43.

48 Talbert, "Moral Competence," 107 (emphasis added).

up for something. It is not necessary that the party whose conduct we protest does or can understand our protest, or respond to it, or reform their behavior in reaction to it. We can protest the behavior of a cruel dictator who will never learn about our protest just as we can protest against what the American slaveholders or German Nazis did even if they are long dead. Or, as we may also put it, the protest function of blame is more about the protesters and those who learn about the protest than about the party whose conduct we protest.

In view of this, it seems very plausible that the protest function of blame could be fulfilled if the blamee is an AI system. For illustration, let us take up the case of the self-driving car again, which prioritizes driving customers home quickly over protecting pedestrians' safety. Such a hierarchy of goals is objectionable, and the behavior that expresses it can thus be an appropriate target of protest. As pedestrians, it makes complete sense to stand up for our safety and make clear that the goal structure that manifests itself in the car's conduct is unacceptable. We, thereby, show to ourselves and those around us that our safety matters to us. Whether or not the car can understand our protest, or modify its behavior in reaction to our protest, is irrelevant for whether it makes sense to protest.

The protest function thus seems to be a clear example of an important function that our blaming practices can still fulfill when the blamee is an AI system.

2.5. Signaling

The same holds true for what has recently been argued to be another important function of blame—namely, *to signal one's commitment to certain norms and values*, or, more specifically, to signal that one is "a member of a particular moral tribe, someone who cares about a set of norms and their breaches, someone who is disposed to police the norms, and more."[49]

For illustration, imagine you witness your colleague telling a racist joke about another colleague.[50] In responding to this with blame (e.g., by telling the joke teller angrily that their joke is inappropriate and deeply hurtful to the victim), one is sending the signal that one is committed to the norm that racist behavior is not okay. Importantly, one is not merely sending this signal to the blamee, but also to bystanders as well as to the victim. To the latter, one is also sending a signal of solidarity ("I know that what *x* is saying is wrong and I've got your back!"). Finally, one is sending information about one's "agential qualities," i.e., roughly speaking, about one's character or regard for others. Thus, a single

49 Shoemaker and Vargas, "Moral Torch Fishing," 587.

50 The following is inspired by Shoemaker and Vargas's discussion of the case of Sarah ("Moral Torch Fishing," 589–90).

blaming response may send "many different signals far and wide" and hence fulfill its signaling function *through many different channels*.[51]

The latter point is important because it suggests that there can be instances of blame which are, again, more about the blamer and those who witness the blaming response than about the blamee. This point is also highlighted by David Shoemaker and Manuel Vargas:

> Given its multichannel nature, in some cases blame's signal may even *exclude the blamed agent altogether*. This is a significant and underappreciated point, for it makes clear just how distinct blame may be from harsh treatment, sanctions, and punishment of the blamed agent. In such cases of "gossipy" blaming, the blamed agent is oftentimes beside the point. Yet the moral signal can remain crucial for the reputation of the blamer and an important data point for social cooperation.[52]

To briefly expand on the last point, note that blaming responses are often *quasi-automatic* reactions to perceived breaches of norms and, in view of their quasi-automaticity, difficult to fake. Hence, there is a high likelihood that observers of a blaming response will be able to gather accurate information from it, making such responses indeed "an important data point for social cooperation."[53]

If we apply the considerations detailed in the preceding to the question we are interested in—namely, whether the signaling function can be fulfilled when the blamee is an AI system—we arrive at the same affirmative answer as we did in the case of the protest function and the conversation function—and for parallel reasons. For illustration, take, again, our example of the self-driving car. When we blame the car for prioritizing driving its customers quickly to their destination over protecting the safety of pedestrians, we signal that we are committed to certain moral norms (e.g., about the importance of not putting other people's lives at risk for trivial reasons). This in turn allows others who observe our response to gather valuable information about our normative stance toward certain types of traffic behavior, about how we would behave in traffic, and, more generally, about certain general agential qualities we possess. For instance, our caring about the safety of pedestrians shows that we possess some amount of regard for our fellow human beings (at least if we additionally assume that the car's conduct presents no immediate danger to ourselves). And just as before, the signaling function can be fulfilled in this

51 Shoemaker and Vargas, "Moral Torch Fishing," 590.

52 Shoemaker and Vargas, "Moral Torch Fishing," 590.

53 Shoemaker and Vargas, "Moral Torch Fishing," 590.

case, *even if* we assume that the target itself does not understand our signaling nor modify its behavior in response to it. This is because the signaling function, just like the protest and conversation function, can be more about the blamer and those who witness the blaming response than about the blamee and, due to its "multi-channel nature," can be fulfilled even if the channel from blamer to blamee is "closed."[54]

The signaling function is thus another example of an important function of blame that could be fulfilled when the blamee is an AI system. On a final note, we believe that this function might even become increasingly important to us (i) the more AI systems become part of our daily social interactions and (ii) the more such systems perform activities that we could also perform ourselves (such as driving cars, waiting tables, taking care of the elderly, etc.). After all, assuming that we will increasingly face situations in which AI systems display problematic conduct in the course of performing actions that *we* could also perform, the following further assumption seems plausible, too: we will increasingly feel the need to signal our commitment to certain norms and values in order to reassure each other that we belong to the same "moral tribe" and to signal our solidarity with potential victims.[55]

2.6. Relationship Management

Tim Scanlon has argued that blame should be understood in terms of relationship modification. According to him, to blame is, roughly, to register impairments in relationships—for example, between friends—and to modify one's attitudes accordingly.[56] In this paper, we remain agnostic about how, exactly, to spell out the nature of blame (see the beginning of section 2). However, it seems plausible to us that Scanlon has identified a further valuable function of blame: by blaming people, we can manage our relationships with them. In what

54 Some readers may still feel uncomfortable with the idea that our blaming practices can be more about the blamer and those who witness an instance of blame than about the blamee. Here is a further consideration in support of this point: even when we focus exclusively on instances of blame where all parties involved are human beings, so-called dyadic cases of blame, where the victim of a transgression overtly blames the transgressor face to face, "are actually not all that frequent" (Shoemaker and Vargas, "Moral Torch Fishing," 590). While they certainly occur, they seem to be far outnumbered by nondyadic cases and, more specifically, cases in which *we blame a transgressor to others in the absence of the transgressor*.

55 To illustrate this point with a concrete example, take the (imagined) case of a waiter robot that prioritizes serving customers with white skin over customers with a different skin color. On witnessing this, many of us would presumably feel the need to signal our commitment to the norm that racist behavior is not OK, as well as our solidarity with potential victims.

56 See Scanlon, *Moral Dimensions*, 128–29.

follows, we will argue that, somewhat surprisingly perhaps, this function can be fulfilled to an important extent when the blamee is an AI system.

Let us begin with Scanlon's account of relationships that we will presuppose in the following. His view starts with paradigmatic intimate relationships like friendship. But it is also meant to make sense of less intimate relationships, for example between colleagues, and even of people's relationships with countries, companies, and other entities, as we will spell out in more detail below. The core idea is that relationships consist in attitudes and dispositions that the parties have toward each other.[57] For our purposes, we can think of representational states about, for example, what to expect from one another and motivational states about how to act toward each other. Take the relationship between colleagues as an example. The relationship-specific standards tell us what we, as colleagues, can be expected to believe and desire in our roles as colleagues. These standards also tell us what an entity needs to be able to be a party in a relationship. In particular, Scanlon argues that being able to make decisions and to regularly and nonaccidentally conform to the standards that govern a relationship is sufficient for being able to be a party in the relevant kind of relationship.[58]

Very briefly, our main argument is this: many AI systems can make decisions in the sense of interacting with their environments based on their representational and motivational states (see section 1 above). Moreover, they can nonaccidentally conform to certain standards. Therefore, they can be parties in some of the relationships Scanlon is concerned with. They can also breach these standards and, thus, we need ways to register these breaches and to revise our relationships accordingly. Blaming these systems can fulfill this important function. This is the skeleton of our view. Let us now flesh it out.

Consider, first, an asymmetrical, nonclose relationship between humans. In Kazuo Ishiguro's novel *The Remains of the Day,* the butler Stevens reflects on the issue of what makes a great butler. Especially important is the duty "to devote the utmost care in the devising of the staff plan."[59] Imagine that the new employer, Mr. Farraday, expects from Stevens utmost care, realizes Stevens's "slovenliness at the stage of drawing up the staff plan," and responds by placing this responsibility on another employee.[60] Thereby, Mr. Farraday would revise their relationship as a response to Stevens's not having the attitudes he expects from his butler. A response of this kind is important in a nonideal world because we need ways to revise our professional relationships in accordance

57 See Scanlon, *Moral Dimensions*, 131.

58 See Scanlon, *Moral Dimensions*, 161–62, 165.

59 Ishiguro, *The Remains of the Day*, 5.

60 Ishiguro, *The Remains of the Day*, 5.

with whether others exercise the care we can reasonably expect from them. Human responses to AI systems can play very similar roles. Imagine that Stevens is replaced by an AI system. The users train it such that when devising a good staff plan comes into conflict with other jobs, say, searching the internet for deals, devising the staff plan is prioritized. Imagine that this works well for a long time, but then the system autonomously prioritizes searching for deals, which results in faulty staff plans and "many quarrels, false accusations, unnecessary dismissals."[61] The users' response would be very similar to the one we imagined from Mr. Farraday: they would register that an expectation regarding the program's priorities has been breached. They would revise their attitudes to it by deciding to not rely on the system anymore and express this by, for example, ordering a new one. It is important for us to be able to respond in this way. If some entity does not have the priorities we can reasonably expect it to have, then we need to be able to change our attitudes toward it. Thus, blaming AI systems in this way fulfills a valuable function.

Some may reply that Stevens is a human being, but an AI system is not, which is, they may say, a crucial difference for whether revising relationships makes sense. We think that *being human* is not an important feature for the relevant kind of relationship management. To see this, consider, second, relationships between individual humans and nonhuman entities, such as collective agents. Scanlon, for instance, discusses the case of a ferry accident with many casualties. He argues that we sometimes "have grounds to suspend our trust of the ferry company (say, by revoking its license to operate ferries)."[62] He explains that this "presupposes trust as the . . . default relationship against [which] a given relationship is measured."[63] Therefore, suspending our trust is a response to the company's impairing the default relationship and hence a form of blame on Scanlon's account. For another case, consider nongovernmental organizations (NGOs) and their donors. They are parties in a relationship that is partly constituted by the NGOs' expectation to be financially supported and the donors' expectation that the money is used in accordance with certain values. Sometimes NGOs fail on this. A Greenpeace activist injured two spectators of a Euro 2020 soccer game and risked harming many more when parachuting into the Munich Olympic Stadium to protest diesel and petrol cars.[64] Plausibly, the donations of donors were not used in adequate ways in this case. For a donor, it would have been appropriate to revise their relationship with Greenpeace, for example, by

61 Ishiguro, *The Remains of the Day*, 5.

62 Scanlon, *Moral Dimensions*, 163.

63 Scanlon, *Moral Dimensions*, 164.

64 *Guardian*, "Greenpeace Apologises for Injuries Caused by Parachuting Protester at Euro 2020."

sending critical emails or donating less for a certain period. Such responses would play the important role of reshaping the relationship that the NGO has impaired. AI systems can, in the relevant ways, be like NGOs. Imagine an AI system that calculates how to use donations in the most efficient way to support human well-being and decides to invest in a certain program, but this turns out to be a very inefficient way to achieve the goal. Then, it would be appropriate for the users to revise their reliance on the system, to give negative feedback, and to look for a better alternative. This response is very similar to the donors' blaming Greenpeace in the parachuting case, and it fulfills the same important functions. Thus, blaming AI systems can be an important way to manage our relationships with nonhuman agents (just like blaming NGOs can).[65]

Some may reply that companies and NGOs, in contrast to AI systems, are constituted by human beings and that this makes an important difference for whether revising relationships with them makes sense. Again, we think that *being constituted by humans* is not a relevant factor here. To see this, consider, third, relationships between humans and their pets. Scanlon argues that for many humans the point of having pets is to have close relationships with them.[66] This relationship includes the expectation that the other party will not harm you or, depending on the kind of pet, that it does what you order it to do. If our pets do not live up to these expectations, it makes sense to revise our attitudes and relationships, for example, by modifying our desire to spend time and play with them. However, the same, we would argue, holds for some near-future or even current AI systems, like care, toy, or sex robots. For some people, one important point of having them is to have a relationship with them.[67] Such a relationship is governed by, for example, the standard not to harm the owners, and, in some cases, the standard that the robots do what the owners order them to do. If the systems breach these standards, their owners can appropriately revise their attitudes toward them, for example, by modifying their desire to spend time with them.

To sum up, many of us have important relationships with employees, companies, NGOs, or pets. These asymmetrical relationships differ in many respects from paradigmatic intimate relationships like close friendship or romantic love. However, what they share with the latter is that they are governed by standards that the parties involved in the relationships can (fail to) live up to. If the other

65 For a defense of the view that there are important parallels between the "regulatory interactions" we can have with collective agents on the one hand and with AI systems on the other, see also List, "Group Agency and Artificial Intelligence."

66 See Scanlon, *Moral Dimensions*, 166.

67 For examples, see, e.g., Nyholm, *Humans and Robots*, 105–9; see Ishiguro, *Klara and the Sun*, for a vivid fictional example.

party breaches the standard and, thereby, impairs the relationship, we can register this and revise our attitudes accordingly. This form of blame enables us to manage our relationships with these entities, which is important in the nonideal world we live in. The same, we have argued, holds true for AI systems. We can have asymmetrical relationships with them that are governed by standards that these systems can (fail to) live up to. If they breach these standards, we can understand this as impairing the relationship we can have with them. It is important for us to be able to manage these relationships. Thus, blaming AI systems within relationships of these kinds plays a valuable role.[68]

2.7. Taking Stock

In the preceding, we took a closer look at the various valuable functions of our blaming practices and discussed which of these functions could still be fulfilled when the blamee is an AI system. We began with a negative claim: the retribution function can plausibly no longer be fulfilled. However, as we furthermore argued, it is also doubtful whether this function is valuable. Regarding the behavior-modification function, we contended that there are no in-principle obstacles to its fulfillment, but that the degree to which this function could be fulfilled would ultimately depend on whether AI systems will be equipped with the relevant learning mechanisms. When we turned to the conversation, protest, and signaling function of blame, such empirical contingencies became less important. These functions, we argued, could still be fulfilled surprisingly well (even if, e.g., AI systems never reach the status of "conversable entities"). The same held true for the relationship-modification function. We have thus arrived at the conclusion that our blaming practices could fulfill several valuable functions when targeting AI systems. If correct, this result would ensure that they would still "have a point" and give us a *pro tanto* reason to extend them to these systems (see the beginning of section 2 above).

3. BLAMING AI AND AI BLAMEWORTHINESS

The result of the last section, however, may not seem enough to make such an extension fully appropriate. This is because, intuitively, it is *fully* appropriate to blame an entity for its conduct only if that entity is *blameworthy*, i.e., morally

68 Some authors, inspired by Peter Strawson's "Freedom and Resentment," claim that another important function of blame is to enable close, personal, symmetrical relationships: without blame responses like resentment, the idea is, there would be no such thing as real friendship or love (see, e.g., Shabo, "Where Love and Resentment Meet"). However, we are skeptical about whether this Strawsonian picture is correct (see, e.g., Milam, "Reactive Attitudes and Personal Relationships") and hence will not pursue this line of thought any further.

responsible for that conduct.[69] Hence, it seems that in order to show that it can be fully appropriate to blame AI systems, one would also have to show that AI systems can be morally responsible agents.

List, who suggests that certain forms of holding responsible other than blame should be extended to AI systems (see section 1), also discusses the issue of AI responsibility. His general stance on this issue is quite optimistic:

> While there are significant technical challenges here, conceptually, there is no reason why an AI system could not qualify as a moral agent and, in addition, satisfy the knowledge and control conditions I have stated. Even if existing AI-systems do not yet meet these requirements, there is no reason to think that having an electronic or otherwise engineered hardware is an in-principle barrier to their satisfaction.[70]

Thus, according to List, there are no in-principle obstacles to (future) AI systems fulfilling the conditions for blameworthiness.[71] Assuming List's optimistic stance on this point is correct, this would enable us to arrive at the following conclusion: we have reason to assume that it will be fully appropriate to extend our blaming practices to some future AI systems since (i) we have reason to assume that some future AI systems will be blameworthy for their conduct and (ii) our blaming practices would still fulfill several valuable functions in targeting AI systems (as was argued previously).

However, not everyone will share this optimistic stance on the point of AI blameworthiness.[72] Unfortunately, this is an issue too big to be settled within the scope of this paper.[73] So let us suppose that there are in-principle obstacles

69 To clarify, we presuppose that there are different senses in which it can be appropriate to blame a target. When we say that blaming a certain target is *fully* appropriate, we mean that blaming the target is appropriate in *all* (relevant) senses, i.e., that blaming the target would not merely be *all-things-considered permissible*, but also *fitting* and *deserved*. An anonymous referee urged us to address the important issue of whether the practice of blaming children may be an everyday counterexample to our claim that, intuitively, only blameworthy agents are fully appropriate targets of blame. Here is a brief sketch of how we would respond to this: the practice of blaming children may show that it can sometimes be *all-things-considered permissible* to blame those who are not fully blameworthy (perhaps because it may sometimes have good consequences to blame children). But we do not think that the practice of blaming children shows that it can sometimes be *fully appropriate* (fitting, deserved, etc.) to blame those who are not (fully) blameworthy.

70 List, "Group Agency and Artificial Intelligence," 1229.

71 See also List, "Group Agency and Artificial Intelligence," 1227–31.

72 See, e.g., Hakli and Mäkelä, "Moral Responsibility of Robots and Hybrid Agents."

73 For more on this topic, see also our unpublished manuscript "How Robots Can Be Blameworthy" (coauthored with Peter Schulte).

to AI systems fulfilling the conditions for blameworthiness. It may then seem to follow that our above reasoning would at best be of merely theoretical interest. However, this conclusion may be premature.

One common way to frame discussions about blameworthiness is in *moral* terms. The general idea is that the "worthiness" in "blameworthiness" should be understood in terms of fairness, justice, or desert.[74] Despite important differences, these views share the following core assumption: if an agent fails to fulfill the conditions for blameworthiness, then it would be, in some sense, *morally* inappropriate to blame her (e.g., unjust, unfair, or undeserved) since blame, and, in particular, "open blame," is (at least somewhat) harmful for the blamee. However, as we have argued before (section 2.1), blame seems to *lose* its harmful character when the blamee is an AI system. Now, suppose that we are right about this. Then, it seems to follow that one key motive for avoiding "blame without blameworthiness"—namely, its being morally inappropriate in the way just articulated—no longer seems to apply when the blamee is an AI system.[75]

This, in turn, enables us to arrive at the following result: even if we combine our above reasoning with the assumption that no future AI system will fulfill the conditions for blameworthiness, we might still have good reason to extend our blaming practices to these systems. This is because one key type of moral concern for avoiding "blame without blameworthiness" no longer seems to apply when the blamee is an AI system. And this consideration, combined with the consideration that blame could still fulfill several valuable functions when targeting AI systems, might seem enough for an extension to these systems to be justified.

Against this, though, one might object that blaming a nonblameworthy AI system might still be problematic, especially if there is a blameworthy agent in the vicinity.[76] In particular, one might worry that it may deflect attention away from the real culprit (e.g., the designer or the company) and enable them to get off the hook too easily.

We agree that this is a valid worry. In reply, let us make three points. First, according to the account we have defended, the fact that blaming AI systems would fulfill several valuable functions merely gives us a *pro tanto* reason to

74 See views on fairness (e.g., Wallace, *Responsibility and the Moral Sentiments*), justice (e.g., G. Strawson, "Impossibility of Moral Responsibility"), or desert (e.g., McKenna, "Basically Deserved Blame and Its Value").

75 To clarify, we do *not* want to claim that AI systems lack moral status (or lack moral rights). Our point is a much weaker one: unlike in the case of human beings, a certain prominent class of moral concerns about displaying blaming responses toward nonblameworthy entities seems to become irrelevant when the blamee is an AI system.

76 Many thanks to an anonymous reviewer for raising this important objection.

blame these systems. This reason may very well be outweighed by considerations of the kind just articulated. Thus, our account is perfectly compatible with the claim that we should sometimes only blame the designer or the company, even if it would also make sense to blame the AI system.

Second, sometimes there will be no other agent (either individual or collective) who is blameworthy if an AI system causes (unjustified) harm. In fact, the assumption that we should expect such cases to arise is one key driving force for discussions about responsibility gaps (see section 1). In these cases, blaming a nonblameworthy AI system would not have the problematic consequences mentioned above.

We would maintain, though, that sometimes there will be another agent who is blameworthy, and it will also be true that blaming the nonblameworthy AI system will have some undesirable consequences, but we may still have sufficient reason to blame the AI system. For instance, sometimes it may be very important to respond directly, i.e., in the given situation, to harmful behavior displayed by an AI system, but the real culprit may not be available. For illustration, think, once more, of the signaling function of blame (section 2.5). We can imagine cases in which we have strong reason to send a signal of solidarity to the victim, and it may be that we can only achieve this by responding directly (and in a negative manner) to the AI system that caused the harm in that situation. In sum, we think that there may also be cases in which we will have sufficient reason to blame a nonblameworthy AI system even if this could, in a sense, be said to amount to an act of "misfired" blame and even if doing so had the undesirable consequences described above.

4. CONCLUSION

A common and important part of our everyday moral lives is to blame ourselves and others for bad conduct. The arrival of powerful AI systems that operate autonomously in high-stakes contexts raises the question of whether it makes sense to target these systems with blame when they make bad decisions. We have argued for the admittedly surprising claim that it indeed makes sense to include these systems in our blaming practices since many of the important functions that are fulfilled by blaming humans can also be served by blaming AI systems. We concluded that this gives us good *pro tanto* reason to extend our blaming practices to AI systems.

It does not follow from this that we are obliged to include AI systems in our blaming practices or that there are no important differences between blaming humans and blaming AI systems. Still, the conclusion is important. For even if the arrival of powerful AI systems should require that we reshape some of our

moral theories and regulatory practices, our blaming practices do not need a fundamental revision and are in this sense "future proofed": we can hold onto them and have good reason to include more players on the field.[77]

University of Konstanz
hannah.altehenger@uni-konstanz.de

University of Salzburg
leonhard.menges@plus.ac.at

REFERENCES

Bathaee, Yavar. "The Artificial Intelligence Black Box and the Failure of Intent and Causation." *Harvard Journal of Law and Technology* 31, no. 2 (Spring 2018): 889–938.

Baum, Kevin, Susanne Mantel, Eva Schmidt, and Timo Speith. "From Responsibility to Reason-Giving Explainable Artificial Intelligence." *Philosophy and Technology* 35, no. 12 (March 2022). https://doi.org/10.1007/s13347-022-00510-w.

Burri, Susanne. "What Is the Moral Problem with Killer Robots?" In *Who Should Die? The Ethics of Killing in War*, edited by Bradley Jay Strawser, Ryan Jenkins, and Michael Robillard, 163–85. Oxford: Oxford University Press, 2018.

Coates, D. Justin, and Neal A. Tognazzini, eds. *Blame: Its Nature and Norms*. New York: Oxford University Press, 2013.

———. "The Contours of Blame." In Coates and Tognazzini, *Blame*, 3–26.

———. "The Nature and Ethics of Blame." *Philosophy Compass* 7, no. 3 (March 2012): 197–207.

77 The paper was accepted by *JESP* on April 27, 2023. We would like to thank two anonymous referees for *JESP* for engaging deeply with this paper and for providing two sets of very helpful comments, which greatly improved it. Furthermore, we would like to thank the participants of an online workshop on practical philosophy in January 2023, Susanne Burri, Max Kiener, Sebastian Köhler, Peter Königs, and Sven Nyholm, for highly constructive oral and written feedback. We are also grateful to Leonie Eichhorn and Shawn Wang for very helpful written comments, to Peter Schulte for very helpful advice, and to Dorothea Debus, Damiano Ranzenigo, Fabian Stöhr, as well as to the students of Leonhard Menges's Advanced Seminar in Practical Philosophy in December 2022 for very helpful discussions. Finally, we would like to thank Claire Davis for proofreading. Leonhard Menges's work on the paper was supported by the Austrian Science Fund (FWF) P34851-G and is part of the Sense of Responsibility Worth Worrying About research project.

Danaher, John. "Robots, Law and the Retribution Gap." *Ethics and Information Technology* 18, no. 4 (December 2016): 299–309.

———. "Tragic Choices and the Virtue of Techno-Responsibility Gaps." *Philosophy and Technology* 35, no. 26 (June 2022). https://doi.org/10.1007/s13347-022-00519-1.

Floridi, Luciano, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, et al. "AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations." *Minds and Machines* 28, no. 4 (December 2018): 689–707.

Fricker, Miranda. "What's the Point of Blame? A Paradigm Based Explanation." *Noûs* 50, no. 1 (March 2016): 165–83.

Gogoshin, Dane Leigh. "Robot Responsibility and Moral Community." *Frontiers in Robotics and AI* 8, no. 768092 (November 2021). https://www.frontiersin.org/articles/10.3389/frobt.2021.768092.

———. "Robots as Ideal Moral Agents per the Moral Responsibility System." In *Culturally Sustainable Social Robotics: Proceedings of Robophilosophy 2020*, edited by Marco Norskov, Johanna Seibt, and Oliver Santiago Quick, 525–34. Vol. 335 of *Frontiers in Artificial Intelligence and Applications*. Amsterdam: IOS Press, 2020.

Guardian. "Greenpeace Apologises for Injuries Caused by Parachuting Protester at Euro 2020." June 16, 2021. https://www.theguardian.com/football/2021/jun/15/greanpeace-protester-avoids-accident-after-parachuting-into-germany-v-france.

Hakli, Raul, and Pekka Mäkelä. "Moral Responsibility of Robots and Hybrid Agents." *Monist* 102, no. 2 (April 2019): 259–75.

Himmelreich, Johannes. "Responsibility for Killer Robots." *Ethical Theory and Moral Practice* 22, no. 3 (June 2019): 731–47.

Ishiguro, Kazuo. *Klara and the Sun*. London: Faber and Faber, 2022.

———. *The Remains of the Day*. London: Faber and Faber, 1989.

Kiener, Maximilian. "Can We Bridge AI's Responsibility Gap at Will?" *Ethical Theory and Moral Practice* 25, no. 4 (September 2022): 575–93.

Köhler, Sebastian. "Instrumental Robots." *Science and Engineering Ethics* 26, no. 6 (December 2020): 3121–41.

Königs, Peter. "Artificial Intelligence and Responsibility Gaps: What Is the Problem?" *Ethics and Information Technology* 24, no. 36 (September 2022). https://doi.org/10.1007/s10676-022-09643-0.

Langer, Markus, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. "What Do We Want from Explainable Artificial Intelligence (XAI)? A Stakeholder Perspective on XAI

and a Conceptual Model Guiding Interdisciplinary XAI Research." *Artificial Intelligence* 296, no. 103473 (July 2021). https://doi.org/10.1016/j.artint.2021.103473.

List, Christian. "Group Agency and Artificial Intelligence." *Philosophy and Technology* 34, no. 4 (December 2021): 1213–42.

London, Alex John. "Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability." *Hastings Center Report* 49, no. 1 (January 2019): 15–21.

Macnamara, Coleen. "Blame, Communication, and Morally Responsible Agency." In *The Nature of Moral Responsibility: New Essays*, edited by Randolph Clarke, Michael McKenna, and Angela M. Smith, 211–36. New York: Oxford University Press, 2015.

———. "Reactive Attitudes as Communicative Entities." *Philosophy and Phenomenological Research* 90, no. 3 (May 2015): 546–69.

Mason, Elinor. *Ways to Be Blameworthy: Rightness, Wrongness, and Responsibility*. New York: Oxford University Press, 2019.

Matthias, Andreas. "The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata." *Ethics and Information Technology* 6, no. 3 (September 2004): 175–83.

McGeer, Victoria. "Civilizing Blame." In Coates and Tognazzini, *Blame*, 162–88.

———. "Scaffolding Agency: A Proleptic Account of the Reactive Attitudes." *European Journal of Philosophy* 27, no. 2 (June 2019): 301–23.

McKenna, Michael. "Basically Deserved Blame and Its Value." *Journal of Ethics and Social Philosophy* 15, no. 3 (August 2019): 255–82.

———. "Directed Blame and Conversation." In Coates and Tognazzini, *Blame*, 119–40.

Menges, Leonhard. "Blaming." In *The Routledge Handbook of Philosophy of Responsibility*, edited by Maximilian Kiener, 315–25. New York: Routledge, 2023.

Milam, Per-Erik. "Reactive Attitudes and Personal Relationships." *Canadian Journal of Philosophy* 46, no. 1 (February 2016): 102–22.

Müller, Vincent C. "Ethics of Artificial Intelligence and Robotics." *Stanford Encyclopedia of Philosophy* (Winter 2020). https://plato.stanford.edu/archives/win2020/entries/ethics-ai/.

Noorman, Merel. "Computing and Moral Responsibility." *Stanford Encyclopedia of Philosophy* (Spring 2020). https://plato.stanford.edu/archives/spr2020/entries/computing-responsibility/.

Nyholm, Sven. "Attributing Agency to Automated Systems: Reflections on Human–Robot Collaborations and Responsibility-Loci." *Science and Engineering Ethics* 24, no. 4 (August 2018): 1201–19.

———. *Humans and Robots: Ethics, Agency, and Anthropomorphism*. Lanham, MD: Rowman and Littlefield, 2020.

Pereboom, Derk. *Free Will, Agency, and Meaning in Life*. New York: Oxford University Press, 2014.

———. *Wrongdoing and the Moral Emotions*. New York: Oxford University Press, 2021.

Pettigrove, Glen. "Meekness and 'Moral' Anger." *Ethics* 122, no. 2 (January 2012): 341–70.

Scanlon, T. M. *Moral Dimensions: Permissibility, Meaning, Blame*. Cambridge, MA: Harvard University Press, 2008.

Shabo, Seth. "Where Love and Resentment Meet: Strawson's Intrapersonal Defense of Compatibilism." *Philosophical Review* 121, no. 1 (January 2012): 95–124.

Shoemaker, David, and Manuel Vargas. "Moral Torch Fishing: A Signaling Theory of Blame." *Noûs* 55, no. 3 (September 2021): 581–602.

Smith, Angela M. "Blame and Holding Responsible." In *Oxford Handbook of Moral Responsibility*, edited by Dana Kay Nelkin and Derk Pereboom, 269–86. Oxford: Oxford University Press, 2022.

———. "Moral Blame and Moral Protest." In Coates and Tognazzini, *Blame*, 27–48.

Solum, Lawrence. "Legal Personhood for Artificial Intelligences." *North Carolina Law Review* 70, no. 4 (1992): 1231–87.

Sommers, Tamler. *Relative Justice: Cultural Diversity, Free WIll, and Moral Responsibility*. Princeton, NJ: Princeton University Press, 2012.

Sparrow, Robert. "Killer Robots." *Journal of Applied Philosophy* 24, no. 1 (February 2007): 62–77.

Strawson, Galen. "The Impossibility of Moral Responsibility." In *Free Will*, edited by Gary Watson, 212–28. New York: Oxford University Press, 2003.

Strawson, Peter F. "Freedom and Resentment." In *Free Will*, edited by Gary Watson, 72–93. New York: Oxford University Press, 2003.

Talbert, Matthew. "Moral Competence, Moral Blame, and Protest." *Journal of Ethics* 16, no. 1 (March 2012): 89–109.

Tigard, Daniel W. "Artificial Moral Responsibility: How We Can and Cannot Hold Machines Responsible." *Cambridge Quarterly of Healthcare Ethics* 30, no. 3 (July 2021): 435–47.

———. "Technological Answerability and the Severance Problem: Staying Connected by Demanding Answers." *Science and Engineering Ethics* 27, no. 5 (October 2021): 59.

Tognazzini, Neal, and D. Justin Coates. "Blame." *Stanford Encyclopedia of Philosophy* (Fall 2018). https://plato.stanford.edu/archives/fall2018/entries/

blame/.
Vargas, Manuel. *Building Better Beings: A Theory of Moral Responsibility*. New York: Oxford University Press, 2013.
Walen, Alec. "Retributive Justice." *Stanford Encyclopedia of Philosophy* (Summer 2021). https://plato.stanford.edu/archives/sum2021/entries/justice-retributive/.
Wallace, R. Jay. *Responsibility and the Moral Sentiments*. Cambridge, MA: Harvard University Press, 1994.
Wallach, Wendell, and Colin Allen. *Moral Machines: Teaching Robots Right from Wrong*. New York: Oxford University Press, 2009.
Wang, Shawn Tinghao. "The Communication Argument and the Pluralist Challenge." *Canadian Journal of Philosophy* 51, no. 5 (July 2021): 384–99.
Watson, Gary. "Responsibility and the Limits of Evil: Variations on a Strawsonian Theme." In *Agency and Answerability: Selected Essays*, 219–59. New York: Oxford University Press, 2004.
Winfield, Alan F. T., and Marina Jirotka. "The Case for an Ethical Black Box." In *Towards Autonomous Robotic Systems*, edited by Yang Gao, Saber Fallah, Yaochu Jin, and Constantina Lekakou, 262–73. Cham, Switzerland: Springer International Publishing, 2017.

Journal of Ethics and Social Philosophy
VOL. 27, NO. 2 · APRIL 2024
https://doi.org/10.26556/jesp.v27i2.2470

WHAT RELATIONAL EGALITARIANS SHOULD (NOT) BELIEVE

Andreas Bengtson and Lauritz Aastrup Munch

According to relational egalitarianism, justice requires that people relate as equals. On a common view, *X* and *Y* relate as equals if, and only if, they (1) *regard* each other as equals; and (2) *treat* each other as equals. Here are some passages that express this view:

> [According to relational egalitarianism,] *X* and *Y* relate as equals if, and only if: (1) *X* and *Y* treat one another as equals; (2) *X* and *Y* regard one another as equals.[1]

> [Relational egalitarianism] identifies a social ideal, the ideal of a society in which people regard and treat one another as equals.[2]

> [Relational egalitarianism] prescribes equal treatment as well as equal regard.... Indeed, believing that others are unable or unlikely to make the right decision or manage a particular situation in a way that will effectively advance their good implies a failure to regard them as equals in a relevant sense, namely in terms of their moral agency.[3]

We will refer to this view—the view that for *X* and *Y* to relate as equals, they must regard and treat each other as equals—as *the two-part view*.[4] Consider a case that paradigmatically falls within the scope of this view:

1 Lippert-Rasmussen, *Relational Egalitarianism*, 117.

2 Miller, "Equality and Justice," 224.

3 Hojlund, "What Should Relational Egalitarians Believe?" 56

4 Voigt argues that Anderson and Scheffler (in addition to Miller) also (at least) endorse the two-part view ("Relational Equality and the Expressive Dimension of State Action"). Regarding Anderson, she says:

> For Anderson, all three criteria—equal treatment, equal regard, and expressive concerns—seem to be requirements of relational equality. She considers hierarchies of esteem—"whereby those on the top elicit honor and admiration, while those below are stigmatized and held in contempt as objects of ridicule, loathing, or disgust" (Anderson 2008: 263)—as inimical to social equality, suggesting that citizens' attitudes towards one another clearly fall within the remit of relational equality. Her account also requires that citizens meet certain standards of

> *Racist*: Albert is a White racist. He believes that Black people are morally inferior to White people. Stumbling upon a Black person, Bertram, on the street, Albert regards him as inferior (thinking to himself, "this person is inferior to me") and treats him accordingly by shouting racial slurs at him.

Clearly, Albert fails to *regard* and *treat* Bertram as an equal, i.e., he violates 1 and 2. Albert's belief that Bertram is morally inferior is sufficient for Albert not regarding Bertram as an equal. And his action—shouting demeaning racial slurs—is sufficient for Albert not treating Bertram as an equal. This is as it should be: clearly, relational egalitarianism should find the relationship between Albert and Bertram objectionable *qua* being an inegalitarian relationship. Consider next:

> *Akratic Racist*: Connor is a White racist. He believes that Black people are morally inferior to White people. However, after reading a complicated treatise on how one ought to treat morally inferior people, Connor finds it an insurmountable mental task to derive practical guidance on how he ought to treat others from his belief in the inferiority of Black people. For this reason, his belief never manifests itself in how he acts, let alone in his deliberations about how to act. Stumbling upon a Black person, Derek, on the street, Connor *treats* him as his equal by walking past him.[5]

> conduct when interacting with one another: "To stand as an equal before others in discussion means that one is entitled to participate, that others recognize an obligation to listen respectfully and respond to one's arguments, that no one need bow and scrape before others or represent themselves as inferior to others as a condition of having their claim heard" (Anderson, 199: 313). ("Relational Equality and the Expressive Dimension of State Action," 439–40)

Voigt does not, however, say why she takes Scheffler to endorse the two-part view. It may be because Scheffler argues that to relate as equals, the parties must satisfy the *egalitarian deliberative constraint*:

> If you and I have an egalitarian relationship, then I have a standing disposition to treat your strong interests as playing just as significant a role as mine in constraining our decisions and influencing what we will do. And you have a reciprocal disposition with regard to my interests. In addition, both of us normally act on these dispositions. This means that each of our equally important interests constrains our joint decisions to the same extent. ("The Practice of Equality," 25)

However, as we will see, attitudes and dispositions come apart; see also Lippert-Rasmussen, *Relational Egalitarianism*, 201–5. We thank two anonymous reviewers for pushing us to further clarify this.

5 The akratic racist is thus different from a *strategic* racist, i.e., a racist who regards Black people as morally inferior but treats them as equals to avoid criticism by others (cf.

Some may find Akratic Racist psychologically suspicious. They may find that racist beliefs (regard) and racist treatment cannot come apart in this way; the belief will manifest itself somehow in how Connor treats Black people, e.g., in microaggressions.[6] However, this concern is not available to the relational egalitarian since the two-part view already assumes that beliefs and treatment can be separated—otherwise, there would be no reason to mention both in laying out what it takes to relate as equals. And since we are scrutinizing the two-part view, we follow relational egalitarians in assuming that they can be separated. Thus, Connor has what we may refer to as a "free-floating belief" about the moral inferiority of some people. Although Connor *treats* Derek as his equal—and thus satisfies 2—he fails, due to this free-floating belief, to *regard* Derek as his equal—and thus fails 1.[7] For this reason, Connor and Derek fail to relate as equals. And since justice requires that people relate as equals, Connor commits an injustice.

Here is a problem with this line of reasoning that this paper is dedicated to spelling out. Although the two-part view suggests that Connor's belief instantiates an injustice, it turns out that nothing in the stock of arguments found in the literature on relational egalitarianism supposed to flesh out what it means to relate as equals justifies saying that Connor instantiates an injustice *qua* failing to regard Derek as a moral equal.[8] Or so we argue in section 1 below. Another

---

Lippert-Rasmussen, *Relational Egalitarianism*, 72). There is a relevant difference between the two in the sense that the racist belief does enter the deliberation of the strategic racist—that is not the case for the akratic racist. We return to the strategic racist later.

6 For those who are skeptical, note that for our purposes, any case in which a person regards somebody as morally inferior but where these attitudes do not affect how this person *treats* others (perhaps even due to mere luck) will do. Alternatively, imagine that the government announces that they will be deploying a mental scanner that will reveal to everyone when a person relies on the belief that somebody is morally inferior in their deliberation. Presumably, some of those who believe like Connor will be deterred from relying on this consideration in their deliberations, even though they possess the belief.

7 Or so we shall assume. Connor's behavior to walk past Derek on the street does not strike us as objectionable.

8 We write "instantiate an injustice," but an anonymous reviewer asks if it follows from this that we should judge that Connor is blameworthy. Not necessarily. Suppose that Connor had the belief that Black people are morally inferior because he was taught so in public institutions. In that case, it is clearly regrettable that Connor has this belief, but he may not be blameworthy, and it may not be fitting to condemn him. This points to the wider question of what relational egalitarians who support the two-part view should say about the causes of "bad" beliefs. Perhaps the public institution that fosters the belief that Black people are inferior to White people should be condemned. In fostering this belief, it treats Black people as inferiors. But this also raises the question: Does the wrongness turn on whether the public institution also communicates the view that White people nevertheless ought to treat Black people as equals? Presumably, if one supports the two-part view, this

way of stating the same point: the regard requirement—i.e., 1—of the two-part view appears unjustified. We consider in section 2 whether relational egalitarians can avoid this result by dismissing Akratic Racist as a relevant test case.

A natural question then arises: Despite this shortcoming, can we—on behalf of relational egalitarians—nevertheless come up with an argument that enables us to justify why Connor instantiates an injustice and thereby vindicate the (regard part of the) two-part view? Although we have a less decisive answer to this question, we argue in section 3 that the prospects here are not excellent, and that any forthcoming solution will be deeply controversial. In any case, relational egalitarians who are attracted to the two-part view face a challenge.

We conclude by pointing to three ways in which relational egalitarians could modify their theory to deal with our challenge. They could (i) adopt a treat-only view of relating as equals, (ii) come up with a novel argument that justifies the regard requirement, or (iii) weaken their commitment to the regard requirement.

1. RELATING AS EQUALS AND FREE-FLOATING BELIEFS

We will start by discussing several arguments proposed by relational egalitarians on what it takes to relate as equals and why we should relate as equals. We will argue that none of the arguments entail that "free-floating beliefs," such as the belief entertained by Connor, are objectionable. In other words, we will argue that relational egalitarians have failed to establish that justice demands that we *regard* other people as our equals. But before we do so, a bit of background on relational egalitarianism may be helpful.

Relational egalitarians sometimes present their view in contrast to distributive theories of justice.[9] According to relational egalitarians, distributive theories of justice fail to focus on that which ultimately matters from the point of view of justice. What ultimately matters, justice-wise, is not that people have the same amount of resources (or welfare, opportunity for welfare, or equal access to advantage, for that matter). After all, racism may be prevalent in a society in

is better in the sense that they encourage White people to at least satisfy the treatment component (as opposed to violating both the regard and the treatment components). We thank an anonymous reviewer for raising this issue.

9 See, e.g., Anderson, "What Is the Point of Equality?"; Scheffler, "What Is Egalitarianism?" That the relational egalitarian project may be understood "negatively" in this sense should be distinguished from another sense in which relational egalitarianism may be understood as a "negative project"—namely, in the sense that it finds it easier to describe what relational equality is not, versus what it is. See, e.g., Wolff, "Social Equality and Social Inequality." We thank an anonymous reviewer for this clarification.

which everyone has the same amount of resources. Instead, justice ultimately requires that people stand in relations of equality to each other. And to stand in such relations—as explained in the introduction—requires that people regard and treat each other as equals. We will expand on this basic understanding of relational egalitarianism in what follows as we investigate whether relational egalitarians can explain what is objectionable about Akratic Racist.

1.1. Scheffler's Egalitarian Deliberative Constraint

Let us start with Samuel Scheffler's egalitarian deliberative constraint. Taking as his starting point close interpersonal relationships—such as a marriage—Scheffler argues that relating as equals requires that the parties to the relationship satisfy:

> *The Egalitarian Deliberative Constraint* (EDC): If you and I have an egalitarian relationship, then I have a standing disposition to treat your strong interests as playing just as significant a role as mine in constraining our decisions and influencing what we will do. And you have a reciprocal disposition with regard to my interests. In addition, both of us normally act on these dispositions. This means that each of our equally important interests constrains our joint decisions to the same extent.[10]

This appears to be a plausible, necessary requirement of what it takes to relate as equals. Consider a marriage in which the husband's interests always trump his wife's interests in collective matters. Clearly, we would not say that this is a relationship among equals. Indeed, the husband seems to stand as a superior in relation to his wife. Can the EDC explain why free-floating beliefs are objectionable?

We may initially believe that it can. The EDC specifies that for me to relate as an equal to you, I must have a standing disposition to treat your strong interests as being as important as my strong interests. But if I regard you as inferior to me, it seems that I do not have a standing disposition to treat your strong interests in accordance with this. In many cases, this would also be true. For many people, how they (are disposed to) treat others is determined in large part by how they regard others. But the relationship is contingent. Merely because, in many actual instances, attitudes and outward behavior align, it does not follow that they are necessarily so aligned. And this is exactly the case with the akratic racist, Connor. That he regards some people as inferior does not

10 Scheffler, "The Practice of Equality," 25. Scheffler understands interests to include needs, values, and preferences (26). See also Cohen, *Finding Oneself in the Other*, 196; Viehoff, "Power and Equality," 353.

manifest itself in how he acts nor in his deliberations about how to act.[11] So it is not the case that Connor has a standing disposition to treat the strong interests of Black people as less important than the strong interests of White people. When interacting with Black people, Connor, because he is akratic, treats their strong interests as playing just as significant a role as his in their collective affairs. This is to say, although Connor regards Black people as inferior to White people, he does not violate the EDC. The EDC cannot explain why free-floating beliefs are objectionable. Notice that this explanatory shortcoming does not detract from the plausibility of Scheffler's deliberative constraint as a component of relational egalitarianism. A disposition to treat others as equals is clearly something that relational egalitarians may find valuable. One reason for this is that egalitarian dispositions are often conducive to people, in fact, treating one another as equals. Our claim is the narrower one that this line of reasoning does not help us diagnose why Connor instantiates an injustice.

1.2. Deontic Relational Egalitarianism

According to Kasper Lippert-Rasmussen, the most plausible reason for why we must relate as equals is that, "as a matter of fact, we are one another's moral equals and in relating as equals we honour that fact."[12] If *X* treats *Y* in a racist manner, *X* treats *Y* as his moral inferior, thereby dishonoring the fact that *Y* is his moral equal. Lippert-Rasmussen ultimately grounds the requirement that people must relate as equals in fairness.[13] So, for our purposes, the question is whether it is unfair that Connor, the akratic racist, regards Black people as inferior (given that it does not affect how he treats, nor how he deliberates about how to treat, Black people).

Why would it be unfair for Connor to have such free-floating beliefs? It cannot be because he thereby treats Black people as inferior since, after all, his belief does not in any way affect how he treats them. Neither can it be because he is disposed to treat Black people as inferior since, once again, his belief does not affect his dispositions—they are independent. The problem with appealing to fairness is that, in a sense, it is assuming what needs to be proven. Something

11 Cf. "One might think that to regard someone as an equal is to be disposed to treat her as an equal. But that isn't so. I can say, 'I regard him as an equal, but I'm too selfish (or biased) to treat him as one.'" (Cohen, *Finding Oneself in the Other*, 197).

12 Lippert Rasmussen, *Relational Egalitarianism*, 170. He says that Anderson ("What Is the Point of Equality?" 313) and Schemmel ("Distributive and Relational Equality," 366) support this argument—or at least an argument along those lines—for why we must relate as equals (*Relational Egalitarianism*, 171).

13 Lippert-Rasmussen, *Relational Egalitarianism*, 172.

may be unfair in either an absolute or a comparative sense.[14] But in either of these senses, that something can only be judged unfair if it has, prior to that, been established that persons have a claim to that something in the first place. But that is exactly what we are discussing. So relational egalitarians cannot merely say that it is unfair that Connor has this free-floating belief about Black people; they must also provide an argument for *why* it is unfair. And it is hard to see what that argument might be, given that it cannot appeal to Connor's deliberation or treatment.

Perhaps Lippert-Rasmussen's argument is pointing us in the right direction in emphasizing that when Connor regards Black people as inferior, he fails to live in accordance with the fact that they are moral equals. But then that does not have to do with fairness. It simply has to do with the fact that people should live in truth, and Connor fails to live in truth. Perhaps relational egalitarians may appeal to this argument when trying to explain why free-floating beliefs are objectionable.

1.3. Telic Relational Egalitarianism

In fact, some relational egalitarians have hinted at an argument along those lines for why it is bad that people relate as unequals. Scheffler argues that "inegalitarian societies [which are inegalitarian in the sense that relationships are inegalitarian] compromise human flourishing; they limit personal freedom, corrupt human relationships, undermine self-respect, and inhibit truthful living."[15] When Scheffler says that inegalitarian relationships inhibit truthful living, it seems that he may have in mind the argument hinted at above. Even if he does not, we may create an argument that is Schefflerian in spirit.[16] Such an argument may be formalized as follows:

P1. Justice requires that people live in truth.
P2. If people are moral equals, they live in truth only if they relate as equals.
P3. People are moral equals.
C. Justice requires that people relate as equals.

This argument can explain why Connor's free-floating belief is objectionable. When Connor regards Black people as morally inferior, he fails to relate to Black people as his equals, and since they are his equals, he fails to live in truth.

14 Broome, "Fairness," 94–95; Estlund, *Democratic Authority*, 69.

15 Scheffler, "Choice, Circumstance, and the Value of Equality," 19.

16 For our purposes, it is not important whether Scheffler supports this argument. What is important is whether this argument, as laid out, can explain why free-floating beliefs are objectionable.

But justice requires that he lives in truth. So Connor's free-floating belief is objectionable according to this relational egalitarian argument.

The problem with this argument is that P1 seems false. It may be good that people live in truth, but from this it does not follow that justice requires that people live in truth. It may be good that people exercise every day, but it is clearly not a requirement of justice that people exercise every day. Similarly, many people have insufficient knowledge, and therefore false beliefs about complicated matters—e.g., nuclear physics—and even though it may be good for people to not have false beliefs about nuclear physics, it is clearly not a justice requirement that people not have false beliefs about nuclear physics. Thus, an underlying assumption of the argument seems to be that because it is good to live in truth, justice requires that we live in truth. But that inference is not valid. This means that we would need an additional argument for why justice requires that people live in truth. And this raises a second problem: what could that argument be? Relational egalitarians have surely not come up with such an argument—and it is hard to imagine what that argument might be. Thus, it seems that the living-in-truth argument is not a promising argument for why free-floating beliefs are unjust.

One may object that we can present a stronger version of the living-in-truth argument, which may offer better support for objecting to the free-floating belief. We could, for instance, weaken P1 such that it says: justice requires that people live in truth *when it comes to truths that have relevance to justice*. We will refer to this as P1*. In support of P1*, suppose that Connor must believe that Black people are inferior to White people to justify his privilege in society and justify the disadvantages that Black people experience.[17] If Connor believed otherwise, he would likely experience cognitive dissonance and be motivated to change his behavior (or beliefs). In this case, living in truth (where those truths have relevance to justice) is necessary for Connor to do his duty of justice.[18]

We have the following two responses.[19] First, this objection assumes that there is a tight connection between beliefs and actions because individuals will find it uncomfortable if their beliefs and actions do not align and will therefore push for consistency. In that sense, this objection assumes that regard and treatment will not come apart in practice. But as we noted in the beginning, this move is not viable to a relational egalitarian who supports the two-part

17 See, e.g., Mills, "White Ignorance."

18 We thank an anonymous reviewer for developing and raising this objection.

19 But some of what we say in the next section in response to a central objection to Akratic Racist may also apply here.

view since that view already assumes that beliefs and treatment can be separated in some sense. Also, there are clearly examples where beliefs in moral inferiority and egalitarian treatment stably co-occur. As mentioned earlier, a strategic racist is one who regards Black people as morally inferior but treats them as equals to avoid criticism by others. For the strategic racist, there is not a tight connection between his belief that Black people are inferior and his treatment of Black people as equals precisely because he is strategic: he does not want to show his belief to others for fear of social sanctions.[20] In that case, justice, treatment-wise, may be achieved even if one has the belief that some are inferior; living in truth is not necessary for him to do his duty of justice (unless we assume that regard is also a requirement of justice, but that is the question we are trying to settle). Second, even if the argument could work, it does not clearly establish what is constitutive for equal relations in the first place (whether that is regard, treatment, some combination, or something else). Instead, it might be said to speak to the question of what we should do to realize an egalitarian society, assuming we already know what an egalitarian society is (i.e., egalitarian beliefs may be strongly conducive to realizing the ideal of relating as equals, and for this reason we should instill egalitarian beliefs in people). But we are interested in the question of what it should mean for a relation to be equal on relational egalitarianism in the first place.

Relational egalitarians have also pointed to other reasons for why inegalitarian relationships are bad (and egalitarian relationships are good). Indeed, Scheffler points to some of them: that inegalitarian relations limit personal freedom, corrupt human relationships, and undermine self-respect.[21] Clearly, the mere

20 The more general point here is, of course, that we cannot pair any specific belief (such as a belief in the moral inferiority of Black people) with a specific disposition or action since dispositions and acts tend to be complex functions of one's entire set of beliefs and desires.

21 See also Scanlon, *The Difficulty of Tolerance*, 204, 212. Relational egalitarians point to a couple of additional reasons: that inegalitarian relationships are bad because they lead to less protection of the inferior's interests than an egalitarian relationship would (Anderson, "Expanding the Egalitarian Toolbox," 145–46); that inegalitarian relationships are bad because they lead to feelings of superiority in the superior (Anderson, "Equality," 50; Fourie, "What Is Social Equality?" 119–21; Scheffler, "Choice, Circumstance, and the Value of Equality," 19); and that inegalitarian relationships lead to servility and deferential behavior (O'Neill, "What Should Egalitarians Believe?" 126; cf. Pettit, *Republicanism*, 87). The arguments we provide in the text also explain why these reasons cannot explain why free-floating beliefs are objectionable. A final additional reason that some relational egalitarians point to is the impersonal badness of unequal relations (most notably O'Neill, "What Should Egalitarians Believe?"). There are several problems with this suggestion for the purposes of explaining Akratic Racist. First, why is the case of Akratic Racist impersonally bad? That requires a further argument. Second, building a theory of justice upon impersonal badness provides a thin foundation—and one with which many will disagree.

fact that Connor has this free-floating belief does not limit anyone's personal freedom. After all, this belief is never manifested in how he acts. Neither does Connor's belief corrupt human relationships.[22] Connor treats Black people as he would have treated them if he had the belief that they are equal to him. It is hard to see how that could corrupt human relationships. And finally, it does not undermine the self-respect of Black people since Black people will never find out that he regards them as inferior (they cannot infer from how he treats them that he regards them as inferior since his behavior is not different from how it would have been had he regarded them as equal). Thus, the reasons proposed by relational egalitarians as to why inegalitarian relationships are bad (and egalitarian relationships are good) cannot explain why free-floating beliefs are unjust.

1.4. Ross and the Level Playing Field

A new, interesting relational egalitarian argument has been proposed by Lewis Ross in the context of explaining why demographic profiling is undesirable.[23] According to Ross, relational equality requires "a level playing field with respect to earning the esteem of your fellow citizens."[24] Securing a level playing field requires that citizens have particular attitudes, particularly a "default attitude of indifference" as to whether a person possesses certain characteristics that are worthy of high (or low) esteem. Among these characteristics are intelligence, virtue, and vice. We must not think of a person as deserving more or less esteem than somebody else until the person has distinguished themselves in some way.[25] Indifference must be the default. These cognitive components are important, Ross argues, because they facilitate the autonomy of citizens: "it enables them to self-author how they are received by their fellow citizens, rather than to have the reception of their behavior coloured by prior assumptions."[26] In this way, Ross provides a convincing argument for why demographic profiling is objectionable. In cases of demographic profiling, the profiler does not take indifference as the

---

Third, relational egalitarians usually argue that it speaks in favor of their theory of justice that it is in line with the concerns of real-life egalitarians (e.g., Anderson, "What Is the Point of Equality?"; Schemmel, *Justice and Egalitarian Relations*; for discussion of this, see Lippert-Rasmussen, *Relational Egalitarianism*, 174–77). But clearly impersonal badness is not the (primary) concern of real-life egalitarians.

22 One may object here that we seem to assume that relationships are all about treatment. But some may say that what we believe about each other is constitutive of our relationships. We consider this view below.

23 Ross, "Profiling, Neutrality, and Social Equality."

24 Ross, "Profiling, Neutrality, and Social Equality," 815.

25 Ross, "Profiling, Neutrality, and Social Equality," 816.

26 Ross, "Profiling, Neutrality, and Social Equality," 816.

default, and he thereby hinders that the profilee can be autonomous in the sense of self-authoring how she is received by her fellow citizens. May this argument also explain why Connor's free-floating belief is objectionable?

We may think that it does. Connor does not provide a level playing field between Black people and White people. Since he regards Black people as inferiors, it takes more for a Black person to earn Connor's esteem than in the case of a White person. And in that sense, Connor grants the Black person worse opportunities to be a self-author than he grants to a White person. But if we look closer at Connor's case, we see that this is, in fact, not true. Connor's belief that Blacks are inferior to Whites does not manifest itself in how he acts, let alone in his deliberations about how to act. Thus, when Connor encounters Black people on the street, he deliberates and acts *as if* he is indifferent (even though he is not indifferent). He deliberates and acts as if there is a level playing field between Blacks and Whites. When they meet Connor on the street, Black people have the same opportunity to self-author as White people do. A Black person cannot convincingly say to Connor, "You gave me worse opportunities for earning your esteem than you gave to the White guy over there!" Another way of illustrating this is by contrasting Connor to another person, Erika, who does *not* regard Black people as inferior. When encountering Black people on the street, there is no difference in how Connor and Erika deliberate and act. So if Erika does not violate the requirement of providing a level playing field, then neither does Connor.[27]

Thus, Ross's argument cannot explain why free-floating beliefs are objectionable. Ross might be happy with that. After all, many people do not have free-floating beliefs in the way that Connor does. So his argument can explain why most, if not all, actual cases of demographic profiling are wrong. But that is not what we are after in this paper. We are exploring whether the arguments proposed by relational egalitarians for why we should relate as equals can explain why free-floating beliefs are objectionable—and ultimately, whether relational egalitarian justice requires that we regard each other as equals. Thus, we must continue our investigation.

1.5. Schemmel's Expressivist Argument

Finally, we turn to Christian Schemmel's expressivist relational egalitarian argument.[28] Although Schemmel is an institutionalist relational egalitarian—in the

27 One may object that there should also be a level playing field in Connor's mind, but there is not since he regards Black people as inferior. Without a further argument for why this must be the case, even when the belief is not materialized in any sense in Connor's interactions with Black people, this objection simply begs the question.

28 Schemmel, "Distributive and Relational Equality." See also Schemmel, *Justice and Egalitarian Relations*, ch. 2.

sense that the scope of relational egalitarian justice is limited to how the state treats its citizens—his argument can be extended to cover relations between citizens. And since his argument is interesting and original—and highly important in the literature on relational egalitarianism—it is worth investigating whether Schemmel's argument, once extended, can explain why free-floating beliefs are objectionable on relational egalitarianism.

In putting forth his argument, Schemmel starts from an example which he borrows from Thomas Pogge. We are to imagine five different scenarios in which a group of innocent persons is deprived of an important vitamin due to the arrangement of social institutions. The scenarios are as follows:

1. The shortfall is *officially mandated*, paradigmatically by the law.
2. The shortfall results from *legally authorized* conduct of private subjects.
3. Social institutions *foreseeably and avoidably engender* (but do not specifically require or authorize) the shortfall through the conduct they stimulate.
4. The shortfall arises from private conduct that is *legally prohibited but barely deterred.*
5. The shortfall arises from social institutions *avoidably leaving unmitigated the effects of a natural defect*.[29]

In the five scenarios, the vitamin deficiency and the number of deprived people are exactly the same. This means that if we find the five scenarios unequally unjust, we cannot appeal to the distributions to explain why that is the case. Schemmel uses this to argue that "the attitudes of social and political institutions towards people expressed in the way such institutions treat them are relevant to justice."[30] Whereas what is expressed in 1, Schemmel argues, is outright hostility toward the deprived group because the state aims to bring about the deprivation, 5 expresses neglect in that the state fails to offer treatment of the genetic defect.[31] In the five scenarios, different judgments of moral worth are thus expressed. Whereas the people in all of the scenarios are treated unjustly, 1 expresses that the moral worth of the disadvantaged is much lower than the moral worth of the people in 5. This is why 1 is more unjust than 5. So, what state (in)actions express is important to relational egalitarian justice, according to Schemmel. We can extend Schemmel's argument by claiming that what is expressed in how people treat each other—when this has nothing to do with the state—also matters to relational egalitarian justice. To answer whether this extended version of the

29 Schemmel, "Distributive and Relational Equality," 127.

30 Schemmel, "Distributive and Relational Equality," 133.

31 Schemmel, "Distributive and Relational Equality," 134.

argument may explain why free-floating beliefs are objectionable on relational egalitarianism, we must know how we determine what an act expresses.

According to Schemmel, "the meaning an action has is not just a matter of what the agent in question meant to express with her action, but also of how those who are subject to the action may reasonably understand it."[32] Thus, we must answer two questions to determine what an act expresses: Why did the person act as they did—i.e., what was their motivation? How was the action understood by those affected? Let us return to our akratic racist, Connor. Suppose he meets a Black person on the street and just walks past him without saying "Hi." Since the fact that Connor regards Black people as inferior never enters his deliberation, the reason why he did not say "Hi" is not because he believes the Black person is of inferior moral worth. Instead, it may be because Connor believes that the norms in the given society proscribe saying "Hi" to people you do not know when you meet them on the streets; or it may be because Connor was immersed in his own thoughts and simply forgot to say "Hi." How may Connor's (in)action be understood by the Black person? The Black person may find it appropriate, given that the norms in society proscribe saying "Hi" to people you do not know when you meet them on the streets; or the Black person may feel insulted by the fact that Connor did not say "Hi." But if that is the case, the same would happen in case another person, Rosa, who does not regard Black people as inferior, did not say hi when she met the Black person on the street. In other words, there is no relevant difference between Connor and Rosa in this case. But then the problem cannot be Connor's attitude. Thus, either Connor's inaction does not express anything inappropriate to the Black person or it does. If the former, there is no problem from the point of view of relational egalitarianism. If the latter, the problem is that the analysis becomes overinclusive—the (in)action would also express something inappropriate if the person did not have Connor's belief. So, Connor's belief cannot be the problem. The upshot is that Schemmel's expressivist argument, once properly extended, cannot explain why free-floating beliefs are objectionable on relational egalitarianism.

2. An Objection to Akratic Racist

Before turning to discuss, in the next section, whether relational egalitarians can vindicate the regard requirement by looking outside debates on relational egalitarianism, we want to consider an objection to our example of the akratic racist Connor. The objection may be put forward as a dilemma: either the Connor case describes only a moment in time, in which case it fails to be a relevant relational

32 Schemmel, "Distributive and Relational Equality," 138.

inequality, or the Connor case extends over a longer period of time, in which case the example becomes unbelievable. With regard to the first horn, suppose that "Connor meeting Derek on the street and regarding him as an unequal but treating him as an equal" is meant to describe only this moment in time: that at this particular moment, there is a relational inequality between Connor and Derek. But if it is just a relational inequality at this particular moment in time, relational egalitarians may simply say that it is not an inequality that their theory is meant to capture; they do not care about time-slice relational inequalities. With regard to the second horn, suppose that Connor continues to be confused about how to make his beliefs about inferiority match his behavior and ends up always treating Black people as equals. If so, the example becomes unbelievable. Surely someone who genuinely believed that Black people were inferior would be *motivated* to make their behavior match their beliefs, even if for some period of time the beliefs and behavior did not.[33]

We will start by addressing the first horn. Relational egalitarians in fact do, and should, care about time-slice relational inequalities. We may distinguish two conceptions of relational egalitarianism:

> *Whole Lives Relational Egalitarianism*: Justice requires that, from the perspective of their lives as a whole, people relate socially to one another as equals.
>
> *Time-Relative Relational Egalitarianism*: Justice requires that, at any given moment, people relate socially to one another as equals.[34]

33 We thank an anonymous reviewer for raising this objection. The reviewer further points out that the notion of motivation—that a person would be motivated to make their behavior match their beliefs—could be one of the important ways in which "regarding as equals" could have value independently from "treating as equals." It will create a disposition in akratic racists that the relational egalitarian could say is problematic even if it does not yet affect treatment. This points to another way of understanding the regard component: that regarding someone as *X* is motivation (a disposition) to treat them as such. Note, first, that our arguments in this paper actually leave the disposition view untouched as we also pointed out in our discussion of Scheffler's deliberative constraint. Second, the following problem may arise for relational egalitarians if they adopt this view. As we mentioned earlier, a *strategic* racist is one who regards Black people as morally inferior but treats them as equals to avoid criticism by others. For the strategic racist, his dispositions come apart from his beliefs: he is disposed to treat Black people as equals even though he believes them to be inferior. This means that the proposed suggestion cannot capture the strategic racist as unjust. But we suspect that relational egalitarians are not satisfied, justice-wise, with a society in which some people are strategic racists and treat other people as equals only because they want to look good in the eyes of others.

34 See Lippert-Rasmussen, "Is It Unjust that Elderly People Suffer from Poorer Health Than Young People?" 154.

These views differ. This can be illustrated through *changing places* cases: "Imagine a feudal society with two castes that swap position every twenty years. The first caste dominates the second for twenty years, then the second dominates the first for the subsequent twenty years, and so on. At the end of their lives, the two castes will have exerted equal amounts of control over each other."[35] According to whole lives relational egalitarianism, the relations in the feudal society are not unjust since over their lives as a whole, people relate as equals. This is not the case according to time-relative relational egalitarianism since at no time slice do the two castes relate as equals. Juliana Bidadanure argues that relational egalitarians should accept (at least) the time-relative view. "What is problematic in our examples [including the feudal case]," Bidadanure argues, "is precisely that these societies may not be communities of relational equals *at any point*. Phases of domination, marginalization, or segregation cannot be thought to cancel out diachronically."[36] In other words, relational egalitarians should object to time-slice relational inequalities precisely because such inequalities cannot be compensated at a later point in time. Thus, that Connor regards Derek as a moral inferior because he is Black at a given time slice is, and should be, objectionable according to relational egalitarians. This means that the first horn of the dilemma can be escaped; and this is sufficient to escape the dilemma raised against the case of the akratic racist Connor. However, to not solely rely on this response, we would like to show that we can escape the second horn as well.

The second horn, remember, says that if the Connor case is considered over time, it becomes unrealistic because surely someone who genuinely believed that Black people were inferior would be *motivated* to make their behavior match their beliefs. We have three responses. First, we agree that it is probably psychologically unlikely for many to behave like Connor. But there might be reasons why Connor does not make his behavior match his beliefs. Perhaps he simply does not realize that he is inconsistent in this way. And we take it that it is not uncommon at all for people to be inconsistent in the sense that their behavior does not match their beliefs, not even over time. Think, for instance, of the vast literature on cognitive biases.[37] Second, consider a variant of the Connor case in which Connor simply suspends judgment on the question of whether Black people are equal to White people. He decides to act cautiously

35 Bidadanure, "Making Sense of Age-Group Justice," 241. Perhaps it is possible to specify relational inequalities that span such a short amount of time that they should be deemed morally insignificant (say, inequalities that exist for mere seconds). But even if so, it is hard to imagine that cases such as Akratic Racist would *necessarily* fall below this threshold.

36 Bidadanure, "Making Sense of Age-Group Justice," 246.

37 See, e.g., Kahneman, *Thinking, Fast and Slow.*

in light of his uncertainty and, therefore, treat Black people as equals, even though he has not settled the question for himself. Such suspension seems to be possible over time. Importantly, Connor the suspender also fails to regard Black people as equal. Third, the case of Connor the akratic racist is chosen for methodological reasons: because we want to separate the regard component from the treat component. This is necessary to investigate which role, if any, the regard component plays in relational egalitarianism. For this reason, we need a case that is somewhat psychologically unrealistic. If it was a typical psychological case in which beliefs and behavior were aligned, we might conflate our judgment of the one with the judgment of the other. Thus, for our purposes in this paper, it is not a problem—indeed, quite the contrary—that Connor the akratic racist is not typical, psychologically speaking.

3. BEYOND RELATIONAL EGALITARIAN RESOURCES

We have argued that nothing in the stock of currently available arguments given by relational egalitarians provides us with the resources to explain why it is unjust that the akratic racist fails to live up to the requirement that we regard relevant others as equals. This is problematic since it leaves one part of the two-part view insufficiently motivated. And the result is even worse if one has the pre-theoretical intuition that we *should* condemn the akratic racist on grounds of justice since this leaves us with a misfit between what is suggested by our best available theoretical arguments and our deeply held convictions.

Some may say at this point: perhaps the arguments we need will be forthcoming or come from outside the literature on relational egalitarianism. In this section, we entertain the latter possibility by exploring whether resources from the literature on the topic of what is now commonly referred to as *the doxastic wronging thesis* can vindicate the regard requirement of the two-part view.

Recently, some have been moved by the thought that morality places demands on what we may believe is the case about others. According to the doxastic wronging thesis that captures this idea, it is possible to morally wrong someone merely in virtue of the contents of one's beliefs.[38]

We should thus investigate if some of the arguments supposed to vindicate the doxastic wronging thesis can be used to explain why Connor affronts relational egalitarian justice by failing to regard Derek as a moral equal. A point is worth bearing in mind, however. The literature on the doxastic wronging thesis is still young, and, for all we know, the best possible version and defense of

38 See, e.g., Basu, "What We Epistemically Owe to Each Other," "Radical Moral Encroachment," and "A Tale of Two Doctrines"; Bolinger, "Varieties of Moral Encroachment"; Basu and Schroeder, "Doxastic Wronging"; Schroeder, "When Beliefs Wrong."

the thesis may not yet have been put forward. Thus, some of what we suggest here may be subject to revision as the theory develops in the future. In any case, however, we are not directly interested in whether doxastic wronging exists but rather whether the arguments supposed to show why it is true can be used to show that Connor instantiates a relational egalitarian injustice. Even if doxastic wrongings are possible, this does not by itself tell us if and how relational egalitarians should figure in this fact in their theorizing about the requirements of justice.[39]

Consider then how Rima Basu, arguably the most prominent proponent of the doxastic wronging thesis, explains why doxastic wrongs are possible:

> We are, each of us, in virtue of being social beings, vulnerable, and we depend upon others for our self-esteem and self-respect. Respect and esteem, however, are not mere matters of how we're treated in word or deed, *but also a matter of how we're treated in thought*. The implication of this (quite minimal) Kantian and Strawsonian picture is that people should figure in both our theoretical and practical reasoning in a way that is different from objects. We care how we feature in the thoughts of other people and we want to be regarded in their thoughts in *the right way*; that is, *doxastic wrongs are failures to regard people in the right way*. . . . The point I wish to emphasize here is that we have *both* a moral and a doxastic responsibility of holding one another. It matters how we hold others in our thought. The beliefs we have, after all, are constitutive of our relationships.[40]

Another proponent of the doxastic wronging thesis, Mark Schroeder, explains the possibility of such wrongs as follows:

39 As we shall suggest below, relational egalitarians may have reason to hope that the doxastic wronging thesis *cannot* be vindicated. The reason for this is that if there exists a substantive part of morality that bears on justice, and this part of morality lies outside the scope of what could be captured by an account of justice as "relating as equals," then this may provide one reason to reject the relational account of justice.

Here we set aside the thesis known as moral encroachment, which is sometimes used to motivate the doxastic wronging thesis. The reason for this is that, even if moral encroachment is true, we need an explanation of why encroachment-related failures of believing based on insufficient evidence are, say, a moral wrong or an injustice in the sense that should concern relational egalitarians. Moreover, the encroachment thesis is typically motivated by way of (i) the stakes from acting on a false belief or (ii) the stakes of forming a morally problematic belief. But risk of subsequent action is ruled out for the akratic racist, and the second view presupposes an independent account of what makes some beliefs morally problematic.

40 Basu, "A Tale of Two Doctrines," 110.

> This leads me to think that in order to fully capture the ways in which beliefs can wrong, there must be some moral costs that beliefs carry in and of themselves, independently of their consequences or risked consequences. And this would be true, if our interpersonal relationships are in part constituted by our beliefs about one another. Insofar as our beliefs help to constitute our relationships, the effects of our beliefs on our relationships are not mediated by the effects of our beliefs on our actions or other behaviors. But it is in fact *plausible* that our interpersonal relationships are in part so constituted. It is plausible that the marriage is directly damaged when the jealous wife suspects her innocent husband of cheating, and if the daughter going into engineering feels betrayed by her father, upon learning of his belief, I would be hard pressed to tell her that she is wrong.[41]

Finally, Berislav Marušić and Stephen White motivate the possibility of doxastic wrongings:

> Doxastic wronging occurs when someone, through her beliefs and other doxastic responses (drawing conclusions, withholding judgment, etc.), falls short of another person's legitimate expectation to be regarded in certain ways—in particular, to figure in the other's reasoning in certain ways.[42]

We shall now discuss a representative sample of the thoughts invoked in these passages.

First, both Basu and Marušić and White seem to suggest that morality requires that we *figure in others' reasoning in certain ways*. It may not be wholly clear what they have in mind here, but one possibility is that they mean that beliefs typically come with a set of functional correlates (for instance, being disposed to rely on them in further reasoning) and that a belief may be problematic because committing to its truth thereby shoves such (objectionable) dispositions into subsequent reasoning. While this may be true, it does not help us in the present context. The reason is that Connor is committed to a belief that, deviant as he may be, plays no role in his deliberation due to akrasia. So even if Basu and others are correct about the existence of a morality-derived reasoning requirement, the thought cuts no ice against Connor.[43] In response,

41 Schroeder, "When Beliefs Wrong," 121.

42 Marušić and White, "How Can Beliefs Wrong?" 110.

43 For the thought that it is how beliefs dispose for reasoning or perception that marks out the wrong-making feature, compare Basu and Schroeder: "The racist is paradigmatically disposed to be influenced by her perceptions of race in the beliefs that she forms about

one may say that *if* Connor has the belief that Derek is morally inferior, then this is a symptom of the fact that Connor previously deliberated in an objectionable way, since only objectionable deliberation could bring about a belief with this content. In this view, while the belief itself is not the problem, the belief is direct evidence that problematic reasoning took place. In response to this, we are not persuaded that one can necessarily infer anything from a given belief about the deliberation that led to its formation. Perhaps Connor engaged seriously with philosophical argument beforehand and found the most compelling arguments for basic moral equality lacking. So, we think this move is unpersuasive. In sum, the point about reasoning should not obviously lead relational egalitarians to commit to a non-derivative concern for regard in the form of beliefs about others' moral worth.

Another argument figuring in both Schroeder's and Basu's views seems to be the claim that beliefs are *partially constitutive of interpersonal relationships*. This is an intriguing claim, but as we shall show below, interestingly problematic for several reasons, given the dialectical context that concerns us.

On Schroeder's view, for instance, the thesis of doxastic wronging can be motivated by intuitively problematic beliefs found in friendships and other intimate relations.[44] However, as David Enoch has persuasively shown in his discussion of political paternalism, if we are interested in a kind of interpersonal relationship that reaches much beyond intimate relationships, then we cannot obviously rely on this. He explains this in the context of Stroud's work on epistemic partiality in friendship:

> The friendship case is a case of partiality, as Stroud emphasizes (even in her title). It is grounded in the nature and value of a special, close, and non-universal relationship—that between you and specific others, others who are special to you. Whatever plausibility thoughts of the epistemic relevance of the moral norms have here it owes to these features of the friendship case. But these features are not shared by the case of political paternalism. There, the moral norms that are supposed to govern the belief (in the projected irrationality or akrasia of some others, say) are not partial, they are universal, and to call the relation between one and one's fellow citizens a close relationship would be a huge stretch (and a dangerous one too). Perhaps, in other words, there

another person—more easily persuaded that someone is dangerous, for example, if they are perceived as Black. Racist beliefs are naturally taken not just to be morally problematic, but specifically to wrong their subjects" ("Doxastic Wronging," 183).

44 Schroeder, "When Beliefs Wrong." See also Schroeder, *Reasons First*, ch. 9; and Stroud, "Epistemic Partiality in Friendship."

> is some plausibility to the thought that "Friendship requires epistemic irrationality."[45] The thought that politics requires epistemic irrationality is almost beyond belief.[46]

Although Enoch's focus is different from ours, his reasoning is applicable for our purposes. This is so because we, too, are interested in the content of an ideal that (we take it) is meant to apply between fellow citizens and not only within close relationships.[47] Thus, in the present context, Enoch enables us to appreciate that we cannot infer from the thought that there can be doxastic wrongs in close, interpersonal relationships that have several distinctive and typically morally salient features (such as shared history, partiality, and so on) that there could be doxastic wrongs grounded in the thin relationships between co-citizens.[48] Since we can plausibly imagine that two people may be complete strangers to one another and yet plausibly have a duty to relate as moral equals when they interact—and since relational egalitarians, in fact, argue that this is the case—it is, if Enoch is correct, hard to see how relational egalitarians can vindicate the regard requirement from such premises.[49]

Next, recall that some proponents of the doxastic wronging thesis attempt to ground the thesis in the claim that beliefs (about others) are *partially constitutive of our relations to others*. If correct, this certainly sounds like a kind of consideration that relational egalitarians should be receptive toward. Presumably, relational egalitarians should be concerned with moral wrongdoing that, specifically, occurs at sites that are constitutive of our social relationships. And if thoughts amount to one way of "treating others," as Basu seems to maintain, it seems that we have all we need to justify the regard requirement of the two-part

45 Stroud, "Epistemic Partiality in Friendship," 518.

46 Enoch, "What's Wrong with Paternalism," 33. See also Enoch and Spectre, "There Is No Such Thing as Doxastic Wronging."

47 Furthermore, we are interested in an ideal of relational justice that is meant to function as a *political* ideal. See, e.g., Anderson, "What Is the Point of Equality?"; Kolodny, "Rule over None II"; Scheffler, "The Practice of Equality"; Viehoff, "Democratic Equality and Political Authority."

48 Compare Viehoff, "Power and Equality."

49 See Anderson, "What Is the Point of Equality?"; Scheffler, "The Practice of Equality." Admittedly, some relational egalitarians have argued that we should look to relationships like friendship to distill what it means to relate as equals. However, not even these scholars would concede that equal relations in the sense relevant to relational egalitarianism should perfectly mirror the thick relationships of, say, friendship. Thus, even on such accounts the inference appears premature. See Chan, "Equality, Friendship, and Politics," for discussion of this point.

view. On this proto-argument, it turns out, we "treat" others as morally inferior when we *regard* them as morally inferior.[50]

In a sense, we cannot rule out that Basu is correct here. Perhaps we should think of our thoughts as a way of "treating" other people in the same way as some relational egalitarians suggest that we are "treating" others when we affect them causally—as Lippert-Rasmussen presents what he dubs the causal condition that is meant to hone in on what the "treatment"-requirement might amount to (notice that this is not a definition of what it means to treat others in the sense relevant for "treating as equals," although it is clearly suggestive of what kinds of activities that should count as "treatment"): "*X* and *Y* treat each other as equals only if *X* and *Y* can affect their respective situations in a relevant way."[51]

What we can do instead is to show why this argument, *as it currently stands,* does not vindicate the regard requirement of the two-part view. The first problem is that the argument begs the question in the context of vindicating the regard requirement. Basu seems to suggest that we should infer the doxastic wronging thesis from the premise that thoughts constitute a part of

50 Interestingly, reliance on Basu's reasoning here suggests that the two-part view should still be revised because, suitably interpreted, the best interpretation of the moral significance of "regarding others as equals" identifies it as a form of *treatment*.

51 Lippert-Rasmussen, *Relational Egalitarianism*, 73. Relational egalitarians have not said much about what it means to *treat* others. But Kolodny may have a similar understanding in mind when he says (although, to be fair, he may simply refer to what it takes to *relate* in the first place):

> Suppose that, in a state of nature, several people collaborate in producing some means. Then some of them run off with an unfair share of the fruits of their labors, never to encounter the others again. There is a disparity of means (snared rabbits, say) and a disparity that results from a failure of equal concern for people's independent claims to them (given equal contributions, the rabbits should have been split equally). Nevertheless, because the thieves and their victims do not continue to live together, because the disparity is not, as it were, woven into the fabric of ongoing social relations, there is no structure of hierarchy or subordination between them. (Kolodny, "Rule over None II," 293).

Because the thieves run away from the victims, and never see them again, they will lack the opportunity to treat each other as equals because they cannot affect each other's respective situations. A similar understanding may be expressed by Anderson when she says: "To stand as an equal before others in discussion means that one is entitled to participate, that others recognize an obligation to listen respectfully and respond to one's arguments, that no one need bow and scrape before others or represent themselves as inferior to others as a condition of having their claim heard" (Anderson, "What Is the Point of Equality?" 313).

Of course, one could also take a narrower (or broader for that matter) view of what it means to *treat* others. Although this issue of what it means to treat others is important, we can set it aside since we are interested in the regard component (exemplified by the Akratic Racist case). We thank an anonymous reviewer for discussion on this.

interpersonal relationships ("It matters how we hold others in our thought. The beliefs we have, after all, are constitutive of our relationships").[52] This way of reasoning may be perfectly fine for Basu's purposes, but notice that what we are after is a vindication of the claim that thoughts—or regard—is of relevance in specifying the ideal of "relating as equals." Stated differently, the premise in Basu's argument is the conclusion for which we are looking for a justification. So, this argument for the doxastic wronging thesis (which we, to reiterate, have no qualms with) is question begging for the purposes of vindicating the regard requirement. One cannot justify the conclusion that thoughts—or regard—amount to a form of relating to others by claiming that thoughts—or regard—amount to a way of relating to others.

In response to this, one may say that we ought to revise our notion of what it means to "relate to others" and accept the view that beliefs amount to a way of relating to others, since Basu's remarks appear plausible. Perhaps, but notice now that relational egalitarians have independent reasons for not wanting to make that move since it may well *prove too much.*

How so? It is worth pointing out that relational egalitarians are not only committed to some account of what it means *to relate as equals*. They are also committed to a view of what it means to *relate as such* since it is only between people who are relevantly socially related that the ideal applies. The common understanding by relational egalitarians of what it means to be relevantly socially related is that "*X* and *Y* are socially related [if and] only if (i) *X* is socially related to *Y* and *Y* is socially related to *X*, (ii) *X* can causally affect *Y* and *Y* can causally affect *X*," and (iii) *X* and *Y* can adjust their conduct in light of each other's conduct and communicate.[53]

But if Basu is correct—that (some) beliefs amount to ways of treating others—then relational egalitarians cannot say this and simultaneously say—as they often want to, and do—that the ideal of justice as relating as equals is inapplicable between agents that cannot affect each other causally; after all, contemporary people may have beliefs about people in, say, the eleventh century, and such beliefs could presumably be wrongful on a vindication of the thesis of doxastic wronging. This combination of commitments could thus create an inconsistency for relational egalitarians. Our twinned argument, then, can be summarized as a dilemma of sorts. On the first horn, relational egalitarians accept the doxastic wronging-based vindication of the regard requirement but must give up on their commitment to the usual scope of their ideal, e.g.,

52 Basu, "A Tale of Two Doctrines," 110.

53 Lippert-Rasmussen, *Relational Egalitarianism*, 126, 128. See also Anderson, "What Is the Point of Equality?" 131.

that "relations" between contemporary people and Inca peasants do not fall within the scope of relational egalitarianism.[54] On the second horn, relational egalitarians abstain from invoking the doxastic wronging line of defense, but, as we have shown above, then they cannot explain why the regard requirement is a requirement of justice as relating as equals.

Let us summarize the findings of this section. We have discussed whether relational egalitarians can vindicate the regard requirement of the two-part view, and thereby show that Akratic Racist instantiates an injustice, by turning to the recent literature on the doxastic wronging thesis. While we cannot completely rule out that this strategy will work, we made three arguments against it. First, we suggested, with inspiration from Enoch, that the kind of relationships invoked to explain the possibility of doxastic wronging is disanalogous to many of the relationships that we should expect the ideal of relating as equals to (also) cover. So more must be said. Second, we pointed out that grounding the possibility of doxastic wronging in the claim that beliefs (about others) constitute interpersonal relationships is a question-begging move in the present context. Finally, we argued that even if the regard requirement could be vindicated via the doxastic wronging thesis, it may prove too much and force relational egalitarians to substantially revise their view of what it means to be socially related.

4. Concluding Remarks

We have argued that relational egalitarians face a challenge in terms of justifying the view that the ideal of justice as relating as equals requires that people regard each other as equals. We have shown that no currently available argument will do the job, and that it is hard to see which form an argument that would fit the bill should take. This, we take it, is bad news for anyone who subscribes to what we termed *the two-part view*.

How should relational egalitarians respond to this? One option would be to abandon the regard requirement and endorse a narrower version of the two-part view according to which justice requires solely that people treat each other as equals. We can call this strategy a *hard revision*. Alternatively, relational egalitarians may look for another argument that can vindicate the regard requirement. Call this strategy *meeting the justificatory burden*. Although we cannot rule out this solution, we have noted significant skepticism about the prospects of this strategy. Finally, relational egalitarians could try to weaken the regard requirement and thereby give it a form that is more easily justified but still captures some of the commitments that motivated the regard requirement in the

54 O'Neill, "What Should Egalitarians Believe?"

first place. Call this strategy a *soft revision*. We leave it to relational egalitarians to propose such revisions.

A further implication of our argument is that relational egalitarians lack the resources to condemn people like the akratic racist Connor that we have exploited as an expository device throughout the paper—at least on grounds of relational egalitarian justice. Is this a problem? We are not sure, and we can remain neutral here. If some relational egalitarians have the intuition that Connor affronts justice, there is even more push toward revising relational egalitarianism to meet this challenge. Others could say that Connor manifests a flawed moral character or that he is indeed engaged in wrongdoing (since wrongdoing does not look like a sufficient condition for a relational egalitarian injustice). But it may also be that some relational egalitarians, when confronted with Connor, see him as a mere illustration of a deeper point: that justice does not require, constitutively, that we regard each other as equals.[55]

Aarhus University
University of Groningen
theandreasbengtson@gmail.com

Aarhus University
lauritzmunch@gmail.com

REFERENCES

Anderson, Elizabeth. "Equality." In *Oxford Handbook of Political Philosophy,* edited by David Estlund, 40–57. New York: Oxford University Press, 2012.

———. "Expanding the Egalitarian Toolbox: Equality and Bureaucracy." *Aristotelian Society Supplementary Volume* 82, no. 1 (June 2008): 139–60.

———. "What Is the Point of Equality?" *Ethics* 109, no. 2 (January 1999): 287–337.

Basu, Rima. "Radical Moral Encroachment: The Moral Stakes of Racist Beliefs." *Philosophical Issues* 29, no. 1 (October 2019): 9–23.

———. "A Tale of Two Doctrines: Moral Encroachment and Doxastic Wronging." In *Applied Epistemology,* edited by Jennifer Lackey, 99–118. Oxford:

55 A previous version of this paper was presented at the Grundlegung session at the Faculty of Philosophy, University of Groningen. We thank the audience for helpful comments. We are particularly grateful to two anonymous reviewers for really helpful written comments. For funding, Bengtson would like to thank Independent Research Fund Denmark (1027-00002B) and the Danish National Research Foundation (DNRF144), and Munch would like to thank the Carlsberg Foundation (CF20-0257).

Oxford University Press, 2021.
———. "What We Epistemically Owe to Each Other." *Philosophical Studies* 176, no. 4 (April 2019): 915–31.
Basu, Rima, and Mark Schroeder. "Doxastic Wronging." In *Pragmatic Encroachment in Epistemology,* edited by Brian Kim and Matthew McGrath, 181–205. New York: Routledge, 2019.
Bidadanure, Juliana. "Making Sense of Age-Group Justice: A Time for Relational Equality?" *Politics, Philosophy and Economics* 15, no. 3 (August 2016): 234–60.
Bolinger, Renée Jorgensen. "Varieties of Moral Encroachment." *Philosophical Perspectives* 34, no. 1 (December 2020): 5–26.
Broome, John. "Fairness." *Proceedings of the Aristotelian Society* 91, no. 1 (June 1991): 87–102.
Chan, Joseph. "Equality, Friendship, and Politics." *Proceedings of the Aristotelian Society* 121, no. 3 (October 2021): 275–98.
Cohen, G. A. *Finding Oneself in the Other.* Edited by Michael Otsuka. Princeton, NJ: Princeton University Press, 2013.
Enoch, David. "What's Wrong with Paternalism: Autonomy, Belief, and Action." *Proceedings of the Aristotelian Society* 116, no. 1 (April 2016): 21–48.
Enoch, David, and Levi Spectre. "There Is No Such Thing as Doxastic Wrongdoing." *Philosophical Perspectives* (forthcoming).
Estlund, David. *Democratic Authority: A Philosophical Framework*. Princeton, NJ: Princeton University Press, 2008.
Fourie, Carina. "What Is Social Equality? An Analysis of Status Equality as a Strongly Egalitarian Ideal." *Res Publica* 18, no. 2 (May 2012): 107–26.
Hojlund, Anne-Sofie Greisen. "What Should Relational Egalitarians Believe?" *Politics, Philosophy and Economics* 21, no. 1 (February 2022): 55–74.
Kahneman, Daniel. *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux, 2011.
Kolodny, Niko. "Rule over None II: Social Equality and the Justification of Democracy." *Philosophy and Public Affairs* 42, no. 4 (Fall 2014): 287–336.
Lippert-Rasmussen, Kasper. "Is It Unjust that Elderly People Suffer from Poorer Health Than Young People? Distributive and Relational Egalitarianism on Age-Based Health Inequalities." *Politics, Philosophy and Economics* 18, no. 2 (May 2019): 145–64.
———. *Relational Egalitarianism: Living as Equals*. Cambridge: Cambridge University Press, 2018.
Marušić, Berislav, and Stephen White. "How Can Beliefs Wrong?—A Strawsonian Epistemology." *Philosophical Topics* 46, no. 1 (Spring 2018): 97–114.
Miller, David. "Equality and Justice." *Ratio* 10, no. 3 (1997): 222–37.
Mills, Charles W. "White Ignorance." In *Race and Epistemologies of Ignorance,*

edited by Shannon Sullivan and Nancy Tuana, 11–38. Albany, NY: State University of New York Press, 2007.

O'Neill, Martin. "What Should Egalitarians Believe?" *Philosophy and Public Affairs* 36, no. 2 (Spring 2008): 119–56.

Pettit, Philip. *Republicanism: A Theory of Freedom and Government*. Oxford: Clarendon Press, 1997.

Ross, Lewis. "Profiling, Neutrality, and Social Equality." *Australasian Journal of Philosophy* 100, no. 4 (2022): 800–24.

Scanlon, T. M. *The Difficulty of Tolerance: Essays in Political Philosophy*. Cambridge: Cambridge University Press, 2003.

Scheffler, Samuel. "Choice, Circumstance, and the Value of Equality." *Politics, Philosophy and Economics* 4, no. 1 (February 2005): 5–28.

———. "The Practice of Equality." In *Social Equality: On What It Means to Be Equals,* edited by Carina Fourie, Fabian Schuppert, and Ivo Walliman-Helmer, 21–44. Oxford: Oxford University Press, 2015.

———. "What Is Egalitarianism?" *Philosophy and Public Affairs* 31, no. 1 (January 2003): 5–39.

Schemmel, Christian. "Distributive and Relational Equality." *Politics, Philosophy and Economics* 11, no. 2 (May 2012): 123–48.

———. *Justice and Egalitarian Relations*. New York: Oxford University Press, 2021.

Schroeder, Mark. *Reasons First*. Oxford: Oxford University Press, 2021.

———. "When Beliefs Wrong." *Philosophical Topics* 46, no. 1 (Spring 2018): 115–27.

Stroud, Sarah. "Epistemic Partiality in Friendship." *Ethics* 116, no. 3 (April 2006): 498–524.

Tomlin, Patrick. "What Is the Point of Egalitarian Social Relationships?" In *Distributive Justice and Access to Advantage: G.A. Cohen's Egalitarianism,* edited by Alexander Kaufman, 151–79. Cambridge: Cambridge University Press, 2014.

Viehoff, Daniel. "Democratic Equality and Political Authority." *Philosophy and Public Affairs* 42, no. 4 (Fall 2014): 337–75.

———. "Power and Equality." In *Oxford Studies in Political Philosophy*, vol. 5, edited by David Sobel, Peter Vallentyne, and Steven Wall, 3–38. Oxford: Oxford University Press, 2019.

Voigt, Kristin. "Relational Equality and the Expressive Dimension of State Action." *Social Theory and Practice* 44, no. 3 (July 2018): 437–67.

Wolff, Jonathan. "Social Equality and Social Inequality." In *Social Equality: On What It Means to Be Equals,* edited by Carina Fourie, Fabian Schuppert, and Ivo Walliman-Helmer, 209–25. Oxford: Oxford University Press, 2015.

Journal of Ethics and Social Philosophy
VOL. 27, NO. 2 · APRIL 2024
https://doi.org/10.26556/jesp.v27i2.2679

RAWLS ON JUST SAVINGS AND ECONOMIC GROWTH

Marcos Picchio

JOHN RAWLS'S discussions of justice between generations have all been brief and in passing. This is perhaps due to the difficulty of the issue, which he claims, "subjects ethical theory to severe if not impossible tests."[1] What attention Rawls does devote to justice between generations is limited to his discussion of a *just savings principle*, which he considers to be part of *justice as fairness*, i.e., his conception of domestic justice. Additionally, the scope of Rawls's discussion of justice between generations is restricted to economic matters; he is primarily concerned with addressing the question of what the rate of savings for capital investment in a just society should be. This question is fundamentally tied to issues concerning economic growth and how high the material standard of life in a just society needs to be.[2] This is a restriction I adopt in the present discussion of justice between generations, which is not to suggest that the scope of intergenerational justice is restricted to only these concerns.[3]

After an overview of the motivation for the just savings principle and its relation to the difference principle, the first task of this article is to address a controversial aspect of Rawls's brief treatment of the question of justice between generations: how the parties in the original position could be motivated to save

1 Rawls, *A Theory of Justice*, 1st ed., 284, and *A Theory of Justice*, rev. ed., 251.

2 Like Rawls, I am following standard macroeconomic theory in presuming that a society's material standard of living depends on its ability to produce goods and services. Productivity depends on both physical and human capital in addition to natural resources and technological know-how. With saving and investment, society increases its capital stock and, in turn, its productive capacity, thereby leading to economic growth and a higher material standard of living. Investment in capital is not limited to physical capital, such as machinery and factories, but also includes human capital; this may be done by way of investment in health care and education.

3 There is perhaps the more pressing question of natural resource conservation. D. Clayton Hubin is the first to point out this deficiency in Rawls's treatment of justice between generations ("Justice and Future Generations"). For recent discussion of the topic from a liberal framework, see Mazor, "Liberal Justice, Future People, and Natural Resource Conversation." I also set aside the theoretical obstacle that the nonidentity problem poses for discussions of intergenerational justice. See Parfit, *Reasons and Persons*, ch. 16.

for future generations. My focus is on the explanation found in Rawls's later work. Rawls suggests here that the correct savings principle is the principle that any generation would have wanted preceding generations to have followed.[4] By expanding upon this explanation, I respond to the objection that this approach disregards the perspective of the first generation. My intention is to show that this objection ceases to be a concern when a proper account of the parties' reasoning is developed. This explanation stays true to modeling the parties as economically rational agents. However, what is notable about the explanation I defend is that it relies on the parties adopting *maximax*—not *maximin*—as a decision rule for rational choice. Though this may come as a surprise, I maintain that this conclusion is consistent with Rawls's justificatory framework.[5] My ultimate aim, however, is not a vindication of the just savings principle. What I wish to do is defend Rawls's justificatory approach to the problem of justice between generations and, in the process, expand upon one of its biggest deficiencies: the lack of other intergenerational savings principles for the parties in the original position to consider. Once other principles are introduced and the reasoning of the parties is elaborated upon, I argue that a different savings principle would be selected. Rawls would undoubtedly reject my proposed savings principle because it requires continual economic growth over generations—a conclusion he is explicitly trying to avoid in his theory of justice.

1. Intergenerational Savings and the Difference Principle

Rawls's earliest and most comprehensive work on justice between generations occurs in section 44 of *A Theory of Justice*.[6] This is where Rawls first introduces the concept of a just savings principle. Unlike the two principles of domestic justice, Rawls never gives a determinate formulation of the just savings principle. Rawls clarifies in *Justice as Fairness: A Restatement* that a savings principle can be seen as a savings schedule, i.e., "a rule stating a fraction of social product to be saved at any given level of wealth."[7] Defining precisely what these rates should be is no task for philosophy, and like Rawls, I will leave this consideration underspecified. For this reason, it is better to understand the various

4 Rawls, *Political Liberalism*, 159–60, and *Justice as Fairness,* 273–75. Though the contents of both texts are similar, especially regarding the discussion of the problem of savings, they do not contain identical language. I rely more on *Justice as Fairness* than on *Political Liberalism* since it contains the definitive presentation of Rawls's views.

5 For an overview of Rawls's three main justificatory frameworks, see Scanlon, "Rawls on Justification."

6 Rawls, *A Theory of Justice*, 1st ed., sec. 44.

7 Rawls, *Justice as Fairness*, 160n38.

savings principles I will discuss below as *families* of savings schedules that share a common structure.

An important point to bear in mind is that Rawls does not consider intergenerational justice to be its own subject separate from that of domestic justice.[8] Further, the just savings principle is not to be understood as an additional principle of domestic justice but rather part of the complete formulation of the difference principle (which itself is part of Rawls's second principle of justice). It is also worth mentioning that in its final formulation in *Theory*, the difference principle requires that "social and economic inequalities be arranged so that they are to the greatest benefit of the least advantaged, *consistent with the just savings principle.*"[9] Curiously, the reformulation of the two principles of justice in *Justice as Fairness* does not mention the just savings principle.[10]

1.1. Clarifications to the Difference Principle

Before turning to the contents of the just savings principle, it is necessary to first focus on an important clarification (or revision) made to the difference principle that is relevant to the topic at hand. In *Justice as Fairness,* Rawls stresses that a "feature of the difference principle is that it does not require continual economic growth over generations to maximize upward indefinitely the expectations of the least advantaged."[11] This clarification reflects a concern with the possibility that the difference principle could be interpreted as requiring a high level of societal production; this would be done to make the least advantaged group as well-off as feasibly possible. The problem with requiring such a high level is that it would be inconsistent with the basic liberties—such as the right of occupational choice—ensured by the lexical priority of the first principle of justice. For Rawls, the "general level of wealth in society, including the well-being of the least advantaged, depends on people's decisions as to how to lead their lives. The priority of liberty means that we cannot be forced to engage in work that is highly productive in terms of material goods."[12] Furthermore, a society may collectively prefer to not be highly productive by scaling back on industrialization or simply opting to not work so hard; this would make the material standard of living for all members of society lower than it could have otherwise been.

8 In *Justice as Fairness,* Rawls writes: "Altogether then we have three levels of justice, moving from in-side outward: first, local justice (principles applying directly to institutions and associations); second, domestic justice (principles applying to the basic structure of society); and finally, global justice (principles applying to international law)" (11).

9 Rawls, *A Theory of Justice*, 1st ed., 302, and *A Theory of Justice*, rev. ed., 266 (emphasis added).

10 Rawls, *Justice as Fairness*, 42–43.

11 Rawls, *Justice as Fairness*, 63.

12 Rawls, *Justice as Fairness*, 64.

According to Rawls's clarified account, the difference principle only requires expansions in inequality to be mutually advantageous—namely, the more advantaged can only do better if it also benefits the least advantaged. Hence, what "the difference principle requires, then, is that however great the general level of wealth—whether high or low—the existing inequalities are to fulfill the condition of benefiting others as well as ourselves."[13] This is different from requiring the *maximization* of the prospects of the least advantaged, as some have previously thought.[14] Maximization would imply high productivity, and as we will see below, Rawls insists that a just society does not require a high material standard of living.

To illustrate the point, consider three distributions of income and wealth that would result from varying economic policies (the numbers represent the general levels among the least advantaged and most advantaged groups, respectively): D_1 (3, 3), D_2 (4, 6), and D_3 (5, 12). Suppose D_3 is a distribution only possible due to very high levels of social productivity. A misreading of the difference principle suggests that the policy that results in D_3 is the only acceptable policy since it maximizes the income and wealth levels of the least advantaged. Yet such a policy may be widely regarded as unpopular by members of a just society. Once clarified, the difference principle allows for D_2 (and arguably D_1 as well). This is because the proper reading of the difference principle permits expansions in economic inequality insofar as they are to the greatest benefit of the least advantaged subject to the constraint imposed by the priority of liberty. What is crucial to note is that the difference principle does not require a just society to make the move from D_1 to D_2 or from D_2 to D_3 if its members are reluctant to do so.[15]

13 Rawls, *Justice as Fairness*, 64.

14 To maintain a maximizing reading of the difference principle while addressing this worry, Samuel Freeman suggests that the difference principle does not require maximization of income and wealth but still requires the maximization of primary goods for the least advantaged. To illustrate this point, he envisions a scenario in which a society chooses to democratize the workplace by giving workers "more control over their working conditions and the means of production, and ownership interests in real capital" (*Rawls,* 113). This may lead to lower production levels and, in turn, lower levels of income and wealth; however, the least advantaged members would enjoy a higher index of other primary goods such as "opportunities for powers and positions of office and bases of self-respect" (*Rawls,* 113). In turn, the prospects of the least advantaged would be maximized.

15 In *Theory,* Rawls does mention that "while the difference principle is, strictly speaking, a maximizing principle, there is a significant distinction between the cases that fall short of the best arrangement" (*A Theory of Justice*, 1st ed., 79, and *A Theory of Justice*, rev. ed., 68). This statement should not be interpreted as *requiring* maximization but only that it is an ideal state of affairs. Rawls's distinction between a *thoroughly just scheme* and a *perfectly just scheme* (*A Theory of Justice*, 1st ed., 78–79, and *A Theory of Justice*, rev. ed., 68) is relevant

1.2. Motivations for the Just Savings Principle

There are three issues that motivate Rawls's discussion of the just savings principle. The first is the appeal to a conception of society as a system of fair cooperation over time from one generation to the next—a central organizing idea in Rawls's theory of justice. Rawls writes: "Since society is a system of cooperation between generations over time, a principle for savings is required."[16] The second issue is that of weighing the interests of the present generation against those of future generations. Determining how high the social minimum should be set and how well-off the least advantaged group can become depends on how much of the social product needs to be set aside for investment in society's capital stock. Last, Rawls is concerned with what can be conceived of as an intergenerational distributive problem: How are the burdens and benefits of "capital accumulation and of raising the standard of civilization and culture" to be shared between generations?[17] This raises a unique challenge for Rawls since saving for future generations seems to violate the spirit of the difference principle.[18] As Samuel Freeman notes, "Rawls thinks that, just as it is unfair for the least advantaged to sacrifice their well-being for the sake of a majority, so too it is unfair for earlier generations to forgo their good for the sake of later generations."[19] It seems clear that any intergenerational savings would be contrary to the interests of earlier generations—specifically, the least advantaged members of early generations. Yet Rawls does not want to maintain that early generations have no duty of justice to save for future generations. The results of one generation consuming the entire social product—even if it greatly benefits the least advantaged group—would be disastrous. Consequently, early generations' sentiments of unfairness are, for Rawls, "entirely natural" yet ultimately "misplaced."[20] In devising the just

here. The former obtains when the index of social primary goods for the least advantaged group is maximized, the latter when inequalities are mutually beneficial.

16 Rawls, *Political Liberalism*, 274.

17 Rawls, *A Theory of Justice*, 1st ed., 286, and *A Theory of Justice*, rev. ed., 252.

18 Steven Wall argues that prioritarianism, which he takes the difference principle to be based upon, would allow for the intergenerational savings called for by the Rawls's savings principle, thereby providing a unified philosophical basis for both principles. He writes that "while prioritarianism gives priority to the interests of those who are badly off, it does not rule out the possibility that large benefits to the better off can be justified even if they would impose some sacrifice [to the worse off]" ("Just Savings and the Difference Principle," 88). While Derek Parfit's important discussion of prioritarianism suggests a link between the difference principle and prioritarianism, there is only a surface level similarity (Parfit, "Equality or Priority?"). As we see below, the philosophical basis for the difference principle is reciprocity, not priority.

19 Freeman, *Rawls,* 136.

20 Rawls, *A Theory of Justice*, 1st ed., 291, and *A Theory of Justice*, rev. ed., 254.

savings principle then, Rawls is trying to strike a happy medium by requiring early generations to save while also alleviating their burden to do so.

1.3. The Contents of the Just Saving Principle

Despite the lack of a determinate formulation on Rawls's behalf, what is clear is that the just savings principle would set the rate of saving based upon the developmental level a society has reached. In other words, the just savings principle would provide a societal savings schedule that would not be overly burdensome on any one generation. Rawls writes: "When people are poor and saving is difficult, a lower rate of saving should be required; whereas in a wealthier society greater savings may reasonably be expected since the real burden of saving is less."[21] Though Rawls is not explicit on the terminology, I follow Frédéric Gaspart and Axel Gosseries in understanding the just savings principle as applying in two different stages of societal development: an accumulation phase followed by a steady-state phase.[22] During the accumulation phase, the rate of savings should result in (real) increases in society's capital stock. The exact savings rate will depend on the developmental stage a society is in. A more advanced, wealthier society in the accumulation phase will have a higher rate of savings than a poorer one. Eventually, a society enters the steady-state phase; this occurs "once just institutions are firmly established."[23] It is at this point that "the net accumulation required falls to zero" and "society meets its duty of justice by maintaining just institutions and preserving their material base."[24] According to Rawls, once the steady-state stage is reached, considerations of justice between generations will allow for (real) net increases in society's capital stock to come to a halt, thereby making the need for saving minimal at most. This entails that once the steady state is reached, later generations are *not* entitled—as a matter of justice—to a higher material standard of life than preceding generations. Rawls reiterates this position in later work when he writes that we "should not rule out Mill's idea of a society in a just stationary state where (real) capital accumulation may cease."[25]

21 Rawls, *A Theory of Justice*, 1st ed., 287, and *A Theory of Justice*, rev. ed., 255.

22 Gaspart and Gosseires, "Are Generational Savings Unjust?"

23 Rawls, *A Theory of Justice*, 1st ed., 289, and *A Theory of Justice*, rev. ed., 255.

24 Rawls, *A Theory of Justice*, 1st ed., 289. There is a slight difference in the passage as it is found in the revised edition of *Theory*: "Once just institutions are firmly established *and all the basic liberties effectively realized*, the net accumulation *asked for* falls to zero" (*A Theory of Justice*, rev. ed., 255, emphasis added).

25 Rawls, "Justice as Fairness," 64. Rawls similarly writes in *The Law of Peoples*: "I follow Mill's view that the purpose of saving is to make possible a just basic structure of society; once

Going forward, I will refer to Rawls's savings principle as the *two-stage principle.*[26] This is done to avoid the question-begging phrasing Rawls employed. Labeling one's preferred saving principle "just" suggests there are no rival savings principles worthy of being deemed "just"—a point that will become more salient further on. With that said, one important feature of the two-stage principle is that Rawls devises it as a constraint on the difference principle.[27] Giving the two-stage principle lexical priority over the difference principle achieves this result. In Rawls's theory, the first principle of justice and the principle of fair opportunity have lexical priority over the two-stage principle, but the two-stage principle has lexical priority over the difference principle. What this means is that increasing the material standard of living for the least advantaged members of a living generation cannot come at the expense of securing or preserving just institutions for future generations. If a society collectively decides to promote production and consumption levels to their highest possible levels while complying with the difference principle, we could assume that this course of policy would be further constrained by the two-stage principle.

What is notable about the two-stage principle is that it provides an account of justice between generations that can be characterized as sufficientarian.[28] After all, what Rawls insists on is that justice between generations consists of reaching a certain basic level in terms of societal development and material well-being and then maintaining it. In *Theory,* Rawls states, quite candidly, that "it is a mistake to believe that a just and good society must wait upon a high material standard of life."[29] This judgment reflects the clarification that the difference principle does not require maximizing income and wealth to the highest permissible levels. According to Rawls, then, once the steady-state phase is reached, future generations are not entitled (as a matter of justice) to a higher material standard of life than preceding generations. What matters from the point of view of justice is that a sufficient material base and, in turn, material standard of living is maintained to preserve a just society. As I argue in section 4, the sufficientarian aspects of Rawls's account of justice between generations need to be given up so as to provide a more complete and satisfying account within his justificatory framework.

---

that is safely secured, real saving (net increase in real capital) may no longer be necessary" (107n33).

26 Attas, "A Transgenerational Difference Principle."

27 Rawls, *A Theory of Justice,* 1st ed., 292, and *A Theory of Justice,* rev. ed., 258.

28 The link between the two-stage principle and sufficientarianism is discussed in Meyer, "Intergenerational Justice."

29 Rawls, *A Theory of Justice,* 1st ed., 290, and *A Theory of Justice,* rev. ed., 257.

2. THE TWO-STAGE PRINCIPLE IN THE ORIGINAL POSITION

In the original edition of *Theory,* the parties in the original position have no reason to select the two-stage principle, much less any savings principle—a point Rawls explicitly acknowledges. This is due to the veil of ignorance and the motivational makeup of the parties:

> The parties, who are assumed to be contemporaries, do not know the present state of society. They have no information about the stock of natural resources or productive assets or the level of technology beyond what can be inferred from the assumption that the circumstances of justice obtain. The relative good or ill fortune of their generation is unknown.[30]

Consequently, "assuming generations are mutually disinterested, nothing constrains them from refusing to make any savings at all."[31] This should be evident since a savings principle would require every living person (both from the least advantaged group and most advantaged group) to make sacrifices for people in the future who will presumably be better off due to the cumulative effect of saving.

This counterintuitive result highlights what many see as a serious limitation of the social contract tradition and its reliance on cooperation among mutually disinterested individuals as a basis for social justice.[32] The lack of direct interaction among members of different generations suggests that the problem of savings is not within the "circumstances of justice," i.e., "what may be described as the normal conditions under which human cooperation is both possible and necessary."[33] Rawls is not shy about exposing this weakness:

> We should now observe that there is a peculiar feature of the reciprocity principle in the case of just savings. Normally this principle applies when there is an exchange of advantages and each party gives something as a fair return to the other. But in the course of history no generation gives to the preceding generations, the benefits of whose saving it has received. In following the savings principle, each makes a contribution to later generations and receives from its predecessors. The first generations may benefit hardly at all, whereas the last generations, those living when no further saving is enjoined, gain the most and give the least.[34]

30 Rawls, *Political Liberalism*, 273.

31 Rawls, *Political Liberalism*, 273n12.

32 See Barry, "Circumstances of Justice and Future Generations" and *Theories of Justice*; Hubin, "Non-Tuism."

33 Rawls, *A Theory of Justice*, 1st ed., 126, and *A Theory of Justice*, rev. ed., 109.

34 Rawls, *A Theory of Justice*, 1st ed., 290.

This explains why the difference principle alone cannot handle the problem of savings. As we saw above, the difference principle requires expansions in inequality to be mutually advantageous to be permissible. Yet it is difficult to imagine how intergenerational inequality could ever be mutually advantageous—the benefits of saving only flow in one direction.[35]

2.1. Rawls's Resolution to the Problem of Justice between Generations

How are intergenerational savings to be decided upon by the parties then? Unless the setup of the original position is modified in some way, there appears to be no way to resolve the problem of savings.[36] The initial solution Rawls proposed was to change his account of the motivational makeup of the parties in the original position. Instead of representing individuals, Rawls proposed that the parties instead represent "family lines" with "ties of sentiment between successive generations."[37] If the parties are understood this way, Rawls posits that they would care about their more immediate descendants and would therefore be motivated to save.

Rawls came to find this initial solution "defective" in light of criticisms that I will not review here.[38] Among the most serious criticisms is how unacceptably ad hoc changing the motivational assumptions of the parties is. As Jane English notes, Rawls's solution to the problem of savings is "in effect, being built into the premises of the theory in the form of a motivational assumption rather than being justified by the theory."[39] The result is that in subsequent work, Rawls retains the original motivational assumptions and proposes the

35 It is worth noting that for Rawls the concept of reciprocity in not simply mutual advantage. Reciprocity is a "moral idea situated between impartiality, which is altruistic, on the one side and mutual advantage on the other" (*Justice as Fairness*, 77). Though Rawls's understanding of reciprocity involves mutual advantage, it goes further in requiring the mutually advantageous arrangement to be fair and qualified with respect to an appropriate benchmark of equality (*Political Liberalism*, 16–17).

36 Recent attempts to resolve this problem that differ from the account I will ultimately propose can be found in Wall, "Just Savings and the Difference Principle"; Gaspart and Gosseries, "Are Generational Savings Unjust?"; Attas, "A Transgenerational Difference Principle"; Heyd, "A Value or Obligation?" Attas also provides a helpful overview of the literature surrounding Rawls's treatment of the subject of justice between generations ("A Transgenerational Difference Principle").

37 Rawls, *A Theory of Justice*, 1st ed., 292.

38 Rawls, *Political Liberalism*, 20n22. The earliest and most penetrating criticisms from philosophers can be found in Hubin, "Justice and Future Generations"; Barry, "Justice Between Generations"; English, "Justice Between Generations." For criticisms from economists, see Arrow, "Some Ordinalist-Utilitarian Notes on Rawls's Theory of Justice"; Harsanyi, "Can the Maximin Principle Serve as the Basis for Morality?"

39 English, "Justice between Generations," 93.

following explanation for the parties' selection of the two-stage principle in the original position:

> Parties are to agree to a savings principle subject to the condition that they must want all previous generations to have followed it. They are to ask themselves how much (what fraction of the social product) they are prepared to save at each level of wealth as society advances, should all generations have followed the same schedule.[40]

Rawls further adds that:

> The correct principle, then, is one the members of any generation (and so all generations) would adopt as the principle they would want preceding generations to have followed, no matter how far back in time. Since no generation knows its place among the generations, this implies that all generations, including the present one, are to follow it.[41]

This explanation for the selection of the two-stage principle is a noticeable improvement over the initial one. Yet this explanation faces one notable difficulty. Recall that the parties do not know the "relative good or ill fortune" of their generation. By this, Rawls presumably means that the parties do not know their historical status: Are they members of a relatively worse-off early generation or a more affluent later generation? This leads to a worry that Rawls does not consider in his brief treatment of justice between generations.

2.2. The Problem of the First Generation

The main issue with Rawls's later explanation is related to a problem he was explicitly concerned with in *Theory*: the first generation to save will not benefit from doing so.[42] Consider that, due to the veil of ignorance, the parties in the original position do not know what generation they belong to, nor do they know the level of economic development their society has reached. This would entail that they do not know whether society is in the accumulation phase or

40 Rawls, *Justice as Fairness*, 160.

41 Rawls, *Justice as Fairness*, 160. It is worth noting that Rawls credits Thomas Nagel and Derek Parfit for suggesting this better approach but also acknowledges that Jane English developed the same approach independently (*Justice as Fairness,* 160n39). Though unacknowledged, this account is likely influenced by the Golden Rule of Accumulation first introduced by Edmund Phelps. See Phelps, "The Golden Rule of Accumulation."

42 Stephen Gardiner also discusses a different variation of the problem of the first generation. The main difference with Gardiner's version of the objection, and his discussion of the fallbacks of Rawls's approach to justice between generations, is that it takes place in the context of the problem resource conservation rather than savings and investment for future generations. See Gardiner, "A Contract on Future Generations?" 110–14.

the steady-state phase. The correct principle (or savings schedule) is supposed to be the one that any generation would want preceding generations to have followed, but this excludes the possibility that the parties are members of an early generation. We do not need to assume this would be the first generation in all the history of mankind, but rather, the first generation within the circumstances of justice to start a fair system of social cooperation and begin the accumulation phase by forgoing some of their own consumption for those in the future.[43]

Now it seems clear that if there were *some* guarantees that the parties were not the first generation, the reasoning Rawls provides would be straightforward. Knowing that much of the uncompensated burden of the accumulation phase will not fall on their generation, of course the parties would have wanted preceding generations to have followed a savings schedule. But there is no such guarantee if we are to strictly abide by the requirements imposed by the veil of ignorance. We may just assume, as Rawls implicitly seems to, that the parties will not be members of a relatively poorer first generation. But like stipulating other-regarding motivational assumptions (as Rawls did initially), this is also unacceptably ad hoc.[44]

3. Why would the parties select the two-stage principle?

Rawls's stipulation that the correct savings principle is the one that the parties would have wanted previous generations to follow sets up an additional choice problem within the original position. When it comes to intergenerational savings, we may ask: If we retain the original motivational assumptions, would the parties really select the two-stage principle (or any societal savings schedule) if there were a possibility of being the first generation? We may further ask: What would mutually disinterested rational agents who lack information about their

43 This stipulation is meant to answer Daniel Attas's complaint that the problem of the initial generation is contrived. His chief objection is that the "problem we are facing is the losses that we will endure in moving from a no-saving unjust situation to a presumably just situation that involves some saving" ("A Transgenerational Difference Principle," 205). This would imply that the problem of the first generation is one of transitional justice "covered by nonideal theory and not by the principles of justice for a well-ordered society" (Rawls, *Political Liberalism*, 18). Yet it is not clear why we should assume that the first generation to begin the accumulation phase is necessarily one that is in a transitional stage. Recall that just institutions are not firmly established until the steady-state phase; this would have the implication that the entire accumulation phase is one of transitional justice in which the difference principle does not apply. The problem of savings is very much a problem for a just society, not a transitionally just society.

44 Note that Rawls's initial explanation does not fall prey to this problem since the first generation would still be motivated by ties of sentiment to the second generation.

historical status agree to when it comes to intergenerational savings? Despite the difficulty these questions pose, we do not need to reject Rawls's second strategy for explaining how the parties in the original position would be motivated to care about intergenerational savings. But if we wish to retain it, we need to explore the reasoning process of the parties in more detail—something that Rawls never does.

If the veil of ignorance were slightly modified so that the parties knew which generation they belonged to, and this generation turned out to be the first one, it is clear the parties would not opt for the two-stage principle as it would be contrary to their interests.[45] With the veil of ignorance back in place, an obvious place to start is by considering how maximin reasoning would guide the parties in their deliberations on savings. However, though initially it was thought that there was a relation between maximin reasoning and the two principles of justice, Rawls later clarifies that the maximin rule is mainly related to the first principle of justice.[46] Rawls does acknowledge that this is "a mistake unhappily encouraged by the faults of exposition in *Theory.*"[47] However, the difference principle (which includes the two-stage principle) is not supported on maximin reasoning but rather on grounds of publicity, reciprocity, and stability.[48] It is also a mistake to think that Rawls models the parties as being highly risk averse and, therefore, psychologically disposed to decide on maximin.[49] Hence, there should be no inconsistency in denying the use of maximin in selecting the two-stage principle.

If only the first principle of justice is tied to maximin reasoning, then why invoke considerations of rational choice in the selection of the two-stage principle? Could the two-stage principle be justified on grounds of publicity, reciprocity, and stability in a similar fashion to the difference principle? Reciprocity quite arguably plays the biggest role in supporting the difference principle, yet as we saw above, the reason why the savings problem is a problem in the first place is due to the lack of reciprocity that is characteristic of intergenerational

45 Note this point is being made *within* the original position where the parties are construed as rational and mutually disinterested. Members of an early generation may be happy to save for other reasons and may even have natural duties (i.e., pre-contractual and non-justice-based) to do so, as Rawls seems to suggest. See Heyd, "A Value or Obligation?"

46 To be more precise, maximin does still play a role in thinking about the second principle of justice since it rules out the principle of utility. But maximin does not play a role in justifying the difference principle over the principle of utility with a social minimum—a criticism first pointed out by R. M. Hare ("Rawls' Theory of Justice—II.)"

47 Rawls, *Justice as Fairness*, 43n3.

48 Rawls, *Justice as Fairness*, secs. 34–37.

49 Rawls, *Justice as Fairness*, sec. 31.

relations. Despite this, I will come back to considerations of reciprocity, as well as publicity and stability, in the penultimate section of this article. For now, it is worth recalling that the two-stage principle does not appeal to considerations of reciprocity as typically understood. As Rawls initially puts it: "We can do something for posterity but it can do nothing for us."[50]

3.1. The Maximin Criterion

Considerations of rational choice can still explain why the parties would select the two-stage principle even if there is a possibility of being the first generation. Though there is no inconsistency in denying the use of maximin, invoking considerations of rational choice requires us to consider the possibility of maximin reasoning reentering the original position. However, it should be emphasized that maximin provides a counterintuitive explanation by suggesting that no savings should be undertaken.[51] The worst-case scenario for the parties is that they are the first generation, and by refusing to save, they ensure that the worst possible outcome (being an early generation) is maximally improved.

To determine whether maximin reasoning applies to the selection of the two-stage principle, we can turn to Rawls's maximin criterion. The maximin criterion can elucidate the choice problem at hand and help us determine what decision rule it would be rational for the parties to adopt. Rawls posits three conditions that jointly ensure the use of maximin is rational in the original position:

1. There is no way to estimate probabilities.
2. There is little to be gained above the level that maximin guarantees.
3. There is the possibility of an outcome that one can hardly accept.

I will not repeat Rawls's argument for how these three conditions obtain in the main choice problem within the original position and how they are tied to the first principle of justice.[52] What is important to note is that Rawls suggests that the third condition alone may be sufficient, and what is crucial is that conditions 2 and 3 obtain to a high degree.[53] As I show below, in selecting a savings principle, conditions 2 and 3 are not met to any significant degree. However,

50 Rawls, *A Theory of Justice*, 1st ed., 291.

51 Arrow, "Some Ordinalist-Utilitarian Notes on Rawls's Theory of Justice" and "Rawls's Principle of Just Savings."

52 Rawls, *A Theory of Justice*, 1st ed., 154–56, *A Theory of Justice*, rev. ed., 134–35, and *Justice as Fairness*, 98–99. Hubin raises an important challenge to condition 2 when one grants that income and wealth are subject to diminishing marginal utility within Rawls's framework. See Hubin, "Minimizing Maximin."

53 Rawls, *Justice as Fairness*, 99.

first I say something in favor of condition 1, which is important for explaining how the parties would reason.

3.2. Ruling Out Expected Utility Maximization and Maximin

In their deliberations, the probability that would be most relevant to the parties' reasoning would be the probability of being any generation, particularly the probability of being the first generation. Recall that due to the veil of ignorance, the original position is supposed to be a situation marked up by uncertainty rather than risk.[54] On Rawls's interpretation of the original position, this means that probabilities cannot reliably be estimated—a major source of disagreement with John Harsanyi.[55] Harsanyi maintains that rationality requires the parties to assign equal probability to ending up as any member of society. This allows the parties to use expected utility maximization, which in turn leads them to select (contra Rawls) the principle of average utility.[56] I will not revisit this controversy here and will treat the choice problem of selecting the two-stage principle as one in which the parties do not have access to any relevant probabilities.[57] The main consideration in support of this stipulation is that, unlike the main choice problem in the original position, the selection of the two-stage principle is one in which the parties cannot invoke Harsanyi's equiprobability assumption due to their not knowing how many generations there are before them or after them. The number of generations there have been or will be is indefinite (though certainly not infinite). Further on, I return and expand on this point in addressing an objection to my central argument.

Establishing that the parties do not have any way of estimating probabilities means that expected utility maximization is off the table as a decision rule. However, maximin is also ruled out because conditions 2 and 3 of Rawls's maximin criterion are not met. Note first that the parties are modeled not only as rational but also as acquisitive. This means that they prefer higher levels of income and wealth to less. If savings are undertaken, the best-case scenario for the parties is that they end up in the steady-state phase. The worst-case scenario is that the parties are the first generation, and saving prevents them from obtaining a higher material standard of living than they could have otherwise

54 The distinction is commonly attributed to Frank Knight. Situations marked by risk involve well-defined probabilities on possible outcomes. Situations marked by uncertainty lack any quantifiable information about possible outcomes. See Knight, *Risk, Uncertainty, and Profit*.

55 Harsanyi, "Can the Maximin Principle Serve as the Basis for Morality?"

56 Also see Harsanyi, "Cardinal Utility in Welfare Economics and in the Theory of Risk Taking," and "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons."

57 For recent commentary on the Rawls-Harsanyi debate, see Moehler, "The Rawls-Harsanyi Dispute."

obtained. This is especially concerning if one turns out to be a member of the least advantaged group. The parties would reason that the further in time their generation lives, the better it is for them in terms of income and wealth if savings are undertaken. Further, they will assume that if no savings are undertaken, the material standard of life of each generation will roughly be the same across time. Though there is intergenerational equality, the material standard of life is much lower than it could have otherwise been.

Condition 2 for Rawls's maximin criterion is met when it is not worthwhile to take a risk for the sake of further advantage above the level maximin guarantees *if* this advantage is not significant. Yet it seems clear it is worthwhile for the parties to take a chance on the two-stage principle; they presumably have *a lot* to gain in terms of income and wealth if it turns out they are not an early generation (this is due to the cumulative effects of saving on economic growth). Of course, a potential gain significantly above the level maximin guarantees can be overridden by the possibility of a more significant loss. This is why Rawls stresses condition 3 when potential outcomes are "intolerable" and involve "grave risk" and "outcomes that one can hardly accept."[58] If the parties are an early generation, saving will undoubtedly be to their disadvantage. Yet the worst outcome of being on the losing end of the gamble hardly seems unacceptable. The worst outcome in the savings choice situation would not be akin to the worst possible outcome that the parties would face if they took their chances when selecting the principle of utility as their principle of social justice. Recall that with the two-stage principle, the savings rate for early generations would presumably be low enough to not be overly burdensome. It is, therefore, safe to conclude that conditions 2 and 3 of Rawls's maximin criterion are not met.[59]

3.3. *The Maximax Criterion*

If both maximin and expected utility maximization are ruled out as decision rules for the choice situation we are considering, an alternative decision rule needs to be identified. My suggestion is an overlooked decision rule for conditions of uncertainty: *maximax* (maximize the best possible outcome). Like the maximin rule, my suggestion is not that maximax be seen as a decision rule for rational choice in all cases of risk and uncertainty.[60] Rather, my suggestion is that the maximax rule is reasonable to apply when certain conditions are met.

58 Rawls, *Justice as Fairness*, 99, *A Theory of Justice*, 1st ed., 154, and *A Theory of Justice*, rev. ed., 134.

59 There may, of course, be other sets of conditions for when it is rational to adopt maximin reasoning. But they need not concern us here. Rawls's maximin criterion is by far the most well-known and most relevant for the inquiry at hand.

60 Rawls, *Justice as Fairness*, 97n19.

The above discussion of the maximin criterion and its relation to the choice situation at hand can be used to provide us with three conditions that are jointly sufficient for when it would be reasonable to apply such a rule:

1. There is no way to estimate probabilities.
2. There is a significant amount to be gained above a guaranteeable level.
3. There is no possibility of an outcome that one can hardly accept.[61]

The selection of the two-stage principle in the original position meets these three conditions: (1) the number of generations is indefinite, so there is no way to assign probabilities; (2) the cumulative effects of even one generation saving for the next are significant; and (3) the two-stage principle is designed to be as undemanding as possible. Therefore, it is rational for the parties to be guided by maximax reasoning in their deliberation.

When assessed next to the possibility of no savings being undertaken, maximax reasoning moves the parties to select the two-stage principle. The choice situation can be represented with the following payoff table (table 1). The numbers represent the general levels of income and wealth a generation (G) can expect based on the selected savings schedule.[62] We can stipulate that the outcome assigned a payoff of 5 represents the sufficiency level Rawls envisioned.

Table 1. No Savings vs. Two-Stage Savings

	G_1	G_2	G_3	G_n
No Savings	2	2	2	2
Two-Stage Savings	1	3	5	5

Note: G = generation.

Recall that the parties are acquisitive, so they prefer more social primary goods to less. Hence, outcomes with a higher level of income and wealth will be preferred to those with less. For simplicity, we can stipulate that it takes three generations to reach the steady-state phase. Any generation after the third (G_n) will be at the same level as the third generation (G_3). The table also shows why

61 These three conditions could perhaps also justify the use of Hubin's *quasi-dominance* decision rule for uncertainty, but I do not explore this possibility here. See Hubin, "Minimizing Maximin."

62 I focus on "general levels of income and wealth" instead of "levels of income and wealth for the representative least advantaged person." "General levels" is Rawls's terminology when discussing just savings and economic growth. It is unclear whether Rawls takes "general levels" to refer to a measure such as gross domestic product (GDP) per capita. But there would be no inconsistency in focusing on GDP per capita (or related measures) here since the parties are not adopting the perspective of the least advantaged in selecting a savings principle.

maximin reasoning leads to no savings, but more importantly, it shows why maximax reasoning leads to the selection of the two-stage principle.

Since Rawls did not go into very much depth when discussing the reasoning of the parties when selecting the two-stage principle, my goal has been to expand upon this neglected aspect of his theory. Now that this has been done, we can move on to the main conclusion of this article: why the parties in the original position would select a different savings principle if given the choice.

4. Expanding the Available Savings Principles

To recap: If the parties' decision is between the two-stage principle and no savings at all, the parties would opt for the two-stage principle. This should be clear since the parties would adopt maximax reasoning. If they are a later generation, the parties will enjoy a significantly higher material standard of living than if there had been no savings. Further, they will live in a society where just institutions are firmly established. If no savings principle is selected, the parties will undoubtedly have a much lower material standard of living if they turn out to be part of any generation that is not the first one. Hence, the parties would still select the two-stage principle over no savings at all since they would want to improve upon the best possible outcome of being a later generation (G_n).

But what if other options besides no savings and two-stage savings are on the menu? Rawls never discusses this possibility, and this is a commonly overlooked deficiency in his discussion of justice between generations. To be fair, Rawls does mention how the principle of utility would lead to an excessive rate of accumulation that would sacrifice early generations.[63] Though the principle of utility is ruled out in the original position, further on (section 4.3), I identify two savings principles that require high levels of savings and which pose a challenge to the maximax argument I am advancing. Before turning to those two principles, I identify and set forth the savings principle that I argue parties in the original position would select.

4.1. The Positive Savings Principle

The savings principle that I argue the parties would select if given the choice is what I will call the *positive savings principle*. As the name suggests, it requires the savings rate to be positive no matter what stage of societal development a generation is in. Like the two-stage principle in the accumulation phase, the positive savings principle relies upon positive savings rates from one generation

63 Rawls, *A Theory of Justice*, 1st ed., 286-7, and *A Theory of Justice*, rev. ed., 253. Whether utilitarianism requires such a policy is, of course, debatable.

to the next. It could also serve as a constraint on the difference principle. But unlike the two-stage principle in the steady-state phase, the savings rate needs to be high enough to increase (real) net capital accumulation from one generation to the next. Further, unlike the two-stage principle, the positive savings principle would not distinguish between an accumulation phase and a steady-state phase. However, we can still use the distinction to understand how the two-stage principle and the positive savings principle are similar and where they diverge.

We can stipulate that the positive savings principle would essentially require the same rates of savings for early generations as the two-stage principle. In this regard, they do not conflict. Early generations are still required to save for future generations at the expense of their material interests, but the rate will be low enough that it does not require significant sacrifices on their behalf. To save words, we can say that throughout the accumulation phase, the two-stage and the positive savings principles will result in the same savings schedule.

It is only when society reaches the "steady-state phase" that the two principles diverge. Bear in mind that the positive savings principle does not imply this distinction. It may turn out that the accumulation phase is, technically speaking, never-ending. Still, for purposes of this discussion, we can use the term "steady-state phase" to denote the level of societal development Rawls envisions as sufficient for a just society. When the steady-state phase is reached, the positive savings principle will still require additional savings so that (real) net accumulation increases from one generation to the next. The question that naturally arises is: How high should the rate of savings be at this stage? It will, of course, be high enough to preserve the material base of a just society. On this point, the two principles coincide again. But as we already know, maintaining a just society could allow for a net accumulation of zero. So, in addition, the positive savings principle should be understood as requiring that additional savings be undertaken so that the general level of income and wealth rises from one generation to the next (just as the two-stage principle does in the accumulation phase). In other words, what distinguishes the positive savings principle is that it requires continuous economic growth across generations.[64]

64 Wall argues that a similar principle would be selected in the original position on prioritarian grounds ("Just Savings and the Difference Principle"). My position and Wall's stand in stark contrast to the one developed by Gaspart and Gosseries, who defend the two-stage principle ("Are Generational Savings Unjust?"). Their reading of Rawls leads them to the conclusion that once the steady-state phase is reached, both saving and dissaving for future generations is (with some caveats) unjust. Attas defends the two-stage principle but on different grounds; he concludes that saving is permissible beyond the state–state phase subject to the condition that it benefits the least advantaged group ("A Transgenerational Difference Principle").

At this stage, it is worth noting that the saving and investment rate is not the only source of economic growth. On the Solow growth model, economic growth is explained by two additional factors: technological change and population growth.[65] The former is also arguably the most important determinant of economic growth.[66] Presumably, a just society's economy would grow from these two sources as well. Past a certain point of development, then, the need to grow an economy through savings and investment in capital may be diminished. In fact, because capital is subject to diminishing returns (the extra output from an additional unit of capital falls as the capital stock increases), we are faced with the worry that savings could become very burdensome for very later generations if the goal is to do more than preserve the material base. This is a worry that cannot be entirely dealt with in a satisfactory way due to the inexactness of the subject at hand. Since it would be extremely difficult to specify the savings rates at any stage of development, it is extremely difficult to specify how much the general level is to be raised from one generation to the next. This is especially complicated when considering the other determinants of economic growth. The positive savings principle does not rule out the possibility that a highly advanced society would adopt a savings rate so minimal that the next generation only enjoys a marginal increase in their material standard of living.

If the answer above is unsatisfactory, one consideration that is worth mentioning has to do with the circumstances of justice—specifically, the condition of moderate scarcity.[67] Due to continuous economic growth, a society may, after all, reach such a high stage of development that no further growth is needed. The society in question overcomes the condition of scarcity, thereby putting an end to the problem of distributive justice that the difference principle is designed to address in the first place.[68] However, such a possibility only adds independent support for the positive savings principle, and it is unclear whether it can be invoked in the original position. Technicalities aside, the important feature of the positive savings principle to bear in mind (and my goal in proposing such a principle) is that it offers a much-needed alternative to the sufficientarian aspects of the two-stage principle. Including a positive savings principle into the choice set casts doubt on whether Rawls is justified in embracing Mill's ideal of a just society in a stationary state.

65 Solow, "A Contribution to the Theory of Economic Growth."

66 Romer, "Endogenous Technical Change."

67 Rawls, *Justice as Fairness*, sec. 24.

68 Wall, "Just Savings and the Difference Principle," 94.

4.2. Positive Savings in the Original Position

Having explained some of the details of the positive savings principle, we now return to the original position. When given the choice between the two-stage principle and the positive savings principle, it is evident that the latter would be chosen. Table 2 represents the updated choice situation.

Table 2. Two-Stage Savings vs. Positive Savings

	G_1	G_2	G_3	G_n
No Savings	2	2	2	2
Two-Stage Savings	1	3	5	5
Positive Savings	1	3	4	$Y > 5$

Note: G = generation; Y = income and wealth.

Both principles have similar implications if the parties turn out to be members of the first generation to start the accumulation phase (G_1). On this consideration, neither principle has the upper hand. The same goes if the parties are members of a generation in the late accumulation phase (G_2). It is when the parties consider they are a generation in the "steady-state" phase (G_3) that the principles diverge. Under the positive savings principle, G_3 still needs to save for the next generation. This means that the general level for G_3 under the positive savings principle must be less than the general level under the two-stage principle. If the parties knew there would only be three generations, then maximax would lead to the two-stage principle. But assuming there are only three generations would once again be an ad hoc modification on Rawls's behalf. It is only when the parties consider they are a generation after the steady-state phase is reached (G_n) that the balance of reason tips in favor of the positive savings principle. This is because they are using maximax reasoning: the best scenario is that they are members of a later generation (G_n). By selecting the positive savings principle, they make the best possible outcome even better.[69]

Additionally, since in selecting a principle of intergenerational savings, we need to allow the parties to take an unquantifiable risk if we are to avoid the conclusion that a no savings principle is selected, the positive savings principle provides a higher possible reward (income and wealth) for the unquantifiable risk at stake (being the first generation). The later a generation is, the higher the parties can expect the general level of income and wealth to be. Since the parties can turn out to be members of any generation, this makes it even more plausible to suggest that they are willing to take their chances on intergenerational savings. In other words, when contrasted with the two-stage principle, the positive

69 Notice also that neither weak nor strong dominance reasoning is applicable here.

savings principle provides a bigger reward for the small unquantifiable risk at hand. This conclusion is consistent with the reasoning Rawls provides for the selection of the two-stage principle. It just happens that Rawls never provides alternatives to the two-stage principle, so no comparisons with other savings principles could be made.

4.3. Extreme and Aggressive Savings

An objection with the maximax solution I am proposing is that it would lead to counterintuitive savings principles if they were included in the menu of options. First, consider an *extreme savings principle*. The extreme savings principle would require significant sacrifices on behalf of early generations for the sake of later generations.[70] Could such a principle be compatible with Rawls's reasoning that the correct principle of intergenerational savings is the one that parties would have wanted previous generations to follow? Unless we were to substantially modify Rawls's theory of justice by giving the extreme savings principle lexical priority over the first principle of justice, the answer is clearly no. Even setting aside this worry and imagining an excessive saving rate compatible with occupational liberty, the maximax criterion would no longer be satisfied if this choice were to be offered. Though there is a lot to be gained, extreme savings would be overly burdensome and would involve an unacceptable outcome due to the high rate of savings it imposes. Though an extreme savings principle should be included in the menu of options, it would be rejected by the parties in the original position.

A more serious challenge to my central argument comes in the possibility of an *aggressive savings principle*.[71] With the exception of one "privileged" last generation, the aggressive savings principle leaves all generations at the level of the first generation that undertakes savings. As stipulated before, this level of saving is not overly burdensome, so one cannot reject aggressive savings on the same grounds as one rejects extreme savings. Table 3 represents the (once again) updated choice situation. Imagine Y^* is an incredibly high level of income and wealth only made possible by aggressive saving. Further, let Y^* denote a general level of income and wealth higher than any level made possible by the positive savings principle.

70 We can imagine how someone like Joseph Stalin would endorse such a rate of capital accumulation. Recall Stalin's infamous five-year plans to industrialize Russia at an unprecedented rate. This required major sacrifices from an entire generation.

71 I am thankful to an anonymous referee for suggesting the aggressive savings principle as an important challenge to my central argument.

Table 3. Positive Savings vs. Aggressive Savings

	G_1	G_2	G_3	G_n	G_{last}
No Savings	2	2	2	2	2
Two-Stage Savings	1	3	5	5	5
Positive Savings	1	3	4	$Y > 5$	Y
Aggressive Savings	1	1	1	1	$Y^* > Y$

Note: G = generation; Y = income and wealth.

On the aggressive savings principle, all generations throughout the history of a just society save for the last "privileged" generation—yet no generation is overly burdened in doing so. If the parties are guided by maximax reasoning, it would seem like they would choose the aggressive savings principle. The best-case scenario is that they are G_{last}, and aggressive savings makes this best possible outcome even better.

The counterintuitive result sketched above suggests that maximax is not a reasonable decision rule in the unique context of selecting a savings principle in the original position. But is it possible for the parties to consider the perspective of the last generation as the last column of table 3 implies? I argue that this kind of scenario cannot be represented in the payoff table, given the setup of the choice situation. The most right-hand column in table 3 should be eliminated as it does not represent a possible state of the world that the parties can envision. Recall that the choice situation is one of uncertainty—there is no way to assign probabilities to being any generation. As discussed earlier, this is because the parties do not know how many generations there will be. Yet, one may object that the setup of the choice situation is smuggling in probabilities by allowing the parties to consider being the first generation but not the last. There appears to be an asymmetry: despite the number of generations being indefinite, the parties can consider being G_1 (G_2 or G_3) but cannot consider being G_{last}. Is this asymmetry justified? I maintain that this asymmetry is justified, and below I explain why.

The most straightforward way to justify to the asymmetry in question is to appeal to a central organizing idea in Rawls's theory of justice. Recall that Rawls conceives of society as a system of fair cooperation over time from one generation to the next. Being a participant in a scheme of social cooperation across time is incompatible with adopting the perspective of a last generation. After all, Rawls's setup of the original position would (presumably) prohibit the parties from even entertaining the possibility of ending their society after one generation (this could be done to maximize one generation's consumption). Adopting the perspective of a last generation is incompatible with Rawls's general framework.[72]

72 I am grateful to a second anonymous referee for calling my attention to this point.

To build on this response, consider that "one generation to the next" also implies the kind of indefiniteness that prohibits the parties from adopting the perspective of a last generation. Outside the original position, the parties could come to learn they are the first generation to begin the accumulation phase—this information is available. However, in all but the most exotic scenarios, the same is not true if the parties are the last generation. Consider: we currently do not know how many generations of humans (or finite creatures that meet conditions for personhood) there will be in the future. Consequently, we have no way of knowing how many successive generations there will be once the accumulation stage of a just society begins. But we *can* know when the sequence of generations beginning the accumulation begins, i.e., we can identify the first generation to begin a fair system of social cooperation across time. Matters would be different if it were common knowledge that a massive asteroid was approaching Earth or that humans would become infertile within a fixed number of generations. In such situations, it would be possible to envision oneself as a member of the last generation. But such situations are beyond the parameters of Rawls's theory of justice. The possibility of a known last generation calls for radical revision to Rawls's theory of justice—or perhaps an entirely new theory altogether.

In brief, my response to the challenge of aggressive savings is as follows: though we can envision the start of a just system of social cooperation, we cannot envision its end. The same should be true of the parties in the original position: the parties can envision themselves being the first generation but not the last. Allowing the parties to adopt the perspective of the last generation would "stretch fantasy too far"—a consideration Rawls originally uses to reject an interpretation of the original principle in which everyone who ever lives is represented.[73] The challenge posed by the aggressive savings principle is neutralized once the parties realize they cannot envision being the last generation. But if this response is unsatisfactory, I offer additional considerations for the positive savings principle over the aggressive savings principle in section 5.

4.4. Is the Positive Savings Principle Compatible with the Difference Principle?

It may be objected that the positive savings principle is incompatible with the difference principle. Recall that Rawls states that a "feature of the difference principle is that it does not require continual economic growth over generations to maximize upward indefinitely the expectations of the least advantaged."[74] Though the positive savings principle does require continual

73 Rawls, *A Theory of Justice*, 1st ed., 139.

74 Rawls, *Justice as Fairness*, 63.

and gradual economic growth, it does not require maximal economic growth since the savings rates are presumably set low enough to not be burdensome on any generation.

A more serious complication arises because economic growth does not necessarily improve the position of the least advantaged group. Because the positive savings principle would be part of the difference principle, in raising the material standard of living from one generation to the next, the expectation is that it would benefit the least advantaged. Yet an increase in the material standard of living may be entirely due to the benefits economic growth has on the most advantaged group. If the material standard of living is understood as an average, then a shift from distribution D_3 (5, 12) to D_4 (5, 13) is an increase in the material standard of living. Note, however, that the two-stage principle faces the same problem during the accumulation phase. This issue is presumably dealt with by the background institutions for distributive justice.[75] The difference principle may be roughly satisfied by adjusting the social minimum and the constant marginal rate of taxation, as Rawls suggests in *Justice as Fairness*.[76] Ensuring that economic growth beyond the steady-state phase benefits the least advantaged group can presumably be achieved by similar policy mechanisms. If no policy mechanism is available, we once again arrive at the conclusion that the difference principle implies that no savings should be undertaken for future generations.

5. FURTHER CONSIDERATIONS IN FAVOR OF THE POSITIVE SAVINGS PRINCIPLE

The main goal of this article has been to demonstrate that the positive savings principle (or a family of savings schedules that leads to gradual and continual economic growth) is the savings principle that the parties in the original position would select on grounds of rational choice. As rational and mutually disinterested agents, the parties would want previous generations to follow the positive savings principle over the two-stage principle. This conclusion holds even if there is a possibility of being a member of the first generation. In section 2, I sidelined the possibility of appealing to considerations of publicity, reciprocity, and stability (on which the difference principle rests) to support the conclusion that Rawls's theory of justice requires continual economic growth. I turn to these considerations below and sketch how they may be used in relation to the problem of just savings.

75 Rawls, *A Theory of Justice*, 1st ed., sec. 43.

76 Rawls, *Justice as Fairness*, 161.

5.1. Indirect Reciprocity

The lack of reciprocity in intergenerational relations is the reason that Rawls initially thought the social contract tradition could not adequately deal with the problem of savings. On this point, Rawls may have been too hasty and not considered the possibility of appealing to *indirect* reciprocity. In contrast to direct reciprocity, the idea is that "cooperation can also be sustained by systems of indirect reciprocity, where there is no requirement that the person *to whom* one supplies a benefit be the person *from whom* one receives a benefit."[77] David Gauthier appeals to such a consideration in addressing a similar problem to his contractarian theory of morality:

> The generations of humankind do not march on and off the stage of life in a body, with but one generation on stage at any time. Each person interacts with others both older and younger than himself, and enters thereby into a continuous thread of interaction extending from the most remote human past to the farthest future of our kind. Mutually beneficial cooperation directly involves persons of different but overlapping generations, but this creates indirect co-operative links extending throughout history.[78]

At this stage, I will stay neutral regarding the viability of accounts of intergenerational justice that rely on indirect reciprocity.[79] Assuming that indirect reciprocity counts as reciprocity in the sense relevant to the parties' deliberation, we could appeal to the notion in determining which savings principle would more adequately reflect considerations of reciprocity. The question that arises is: Which savings principle best appeals to the notion of indirect reciprocity—the two-stage principle, the positive savings principle, or the aggressive savings principle?

There should be little doubt that, on grounds of reciprocity, the positive savings principle also triumphs over both the two-stage and aggressive savings principles. Since the positive savings principle requires every generation to save and invest for the future, no matter the stage of societal development, every generation (apart from the first to start saving) receives a benefit from

77 Heath, "The Structure of Intergenerational Cooperation," 33.

78 Gauthier, *Morals by Agreement,* 299. For an extended critique of Gauthier's approach, see Sauvé, "Gauthier, Property Rights, and Future Generations."

79 As expected, there are difficulties with appealing to generational overlap and indirect reciprocity. Most notably, there is the problem of policies whose negative costs will affect temporally distant generations instead of adjacent ones ("time bombs" for short). See Gardiner, "A Contract on Future Generations?" 103–6.

the antecedent generation and provides a benefit to a subsequent generation. Hence, every generation except for the first contributes toward and benefits from gradually raising the material standard of living. Contrast this with the two-stage and aggressive savings principles. Both principles, in essence, allow for intergenerational free riding.[80] Under two-stage saving, those lucky enough to find themselves in the steady-state phase have received considerable benefits at the expense of antecedent generations. Yet they are not expected to contribute to the same extent since their saving burden is minimal. Similar considerations apply to aggressive savings and its emphasis on a privileged generation reaping all the benefits of capital accumulation.[81] If the notion of a fair system of indirect reciprocity is appealing, then it seems that the positive savings principle better embodies this ideal when contrasted with the two-stage and aggressive savings principles.

5.2. Publicity and Stability

Rawls writes that considerations of publicity "require the parties to evaluate principles of justice in the light of consequences—political, social, and psychological—of the public recognition by citizens generally that these principles are affirmed by them and effectively regulate the basic structure."[82] Relatedly, considerations of stability require that "a political conception of justice must generate its own support and the institutions to which it leads must be self-enforcing."[83] These considerations, especially stability, do appear to justify concern for future generations by the parties in the original position. Yet, at first glance, they do not come on the side of any of the previously discussed saving principles.

Something can be said in favor of the positive savings principle over the two-stage and aggressive savings principle on grounds of publicity and stability

80 For a discussion of intergenerational free riding and its relevance to models of intergenerational reciprocity, see Gosseries, "Three Models of Intergenerational Justice."

81 One may object that the fact that a scheme of cooperation does not require equal sacrifice does not mean those who do not make *any* sacrifice are free riders. If hypothetical rational agents would agree to such an arrangement under fair conditions, it is a just arrangement of benefits and burdens, and there is no legitimate complaint of free riding. Yet, as suggested earlier, Rawls ultimately abandons the idea that his theory of justice is simply an extension of the theory of rational choice (*Justice as Fairness,* 82n2). My comments about reciprocity and free riding appeal to the notion of *reasonableness*, which is distinct from rationality, and which plays a more explicit role in Rawls's later work. Per Rawls, reasonableness is an "intuitive moral idea" that is "applied to persons, their decisions and actions, as well as to principles and standards, to comprehensive doctrines and to much else" (*Justice as Fairness,* 82).

82 Rawls, *Justice as Fairness*, 121.

83 Rawls, *Justice as Fairness*, 125.

if we consider additional empirical factors. Economists have long touted the positive consequences continuous economic growth has on human welfare.[84] Benjamin Friedman advances a related position that is relevant here. Friedman has made an extensive case for the link between economic growth and the flourishing of liberal values and democratic institutions throughout the last two centuries.[85] Friedman further argues that economic stagnation is linked to periods of declining civility, openness, and trust in democratic institutions. Friedman's conjecture is arguably controversial, and so it is questionable whether it is one of the "general facts about human society" the parties have access to behind the veil of ignorance.[86] Regardless, the plausibility of the link is highly relevant to considerations of publicity and stability. If Friedman is right, considerations of stability and publicity would come in favor of the positive savings principle and, in turn, continual economic growth.

6. Conclusion

It seems clear that Rawls would not endorse the positive savings principle since he is quite hostile to the view that social justice requires continual economic growth—a view that Rawls's aversion to can likely be explained by his belief that it bears a close relation to utilitarianism. There is no hiding this hostility: "To achieve a [just society] great wealth is not necessary. In fact, beyond some point it is more likely to be a positive hindrance, a meaningless distraction at best if not a temptation to indulgence and emptiness."[87] Despite this hostility, the aim of this article has not been to vindicate every aspect of Rawls's thinking. Rather, the aim has been to provide a more complete account of justice between generations from within Rawls's broader theory of justice. My main conclusion should not be of interest solely to those committed to Rawls's theory of justice but to anyone interested in answering the challenge of how the social contract tradition can provide a satisfactory account of questions pertaining to the intergenerational domain.

I conclude with some remarks about the viability and moral desirability of the positive savings principle and the notion that social justice requires continuous economic growth. Regarding viability, we must consider whether

84 Tyler Cowen offers the most recent defense along these lines. It should be noted that Cowen deviates from the standard defense by also appealing to the effects of economic growth on welfare viewed from a *significantly* longer time horizon than is typical for economists. See Cowen, *Stubborn Attachments.*

85 Friedman, *The Moral Consequences of Economic Growth.*

86 Rawls, *A Theory of Justice*, 1st ed., 136.

87 Rawls, *A Theory of Justice*, 1st ed., 290, and *A Theory of Justice*, rev. ed., 258–59.

continuous economic growth is, in fact, possible on a finite planet. This is not a question I can adequately take up here—the argument I have advanced only matters if certain empirical assumptions hold. Regarding moral desirability, one can argue that the positive savings principle captures a salient judgment regarding the future of humanity, i.e., that our children and our children's children live more prosperous lives than we do. There are also the various consequentialist considerations in favor of continual economic growth very briefly touched upon in the last section.[88] Aside from being justified by the justificatory framework of the original position then, it may also be said of the positive savings principle that it better matches our judgments in reflective equilibrium.[89]

University of Wisconsin–Madison
mpicchio@gmail.com

REFERENCES

Arrow, Kenneth J. "Rawls's Principle of Just Savings." *Swedish Journal of Economics* 75, no. 4 (December 1973): 323–35.

———. "Some Ordinalist-Utilitarian Notes on Rawls's Theory of Justice by John Rawls." *Journal of Philosophy* 70, no. 9 (May 1973): 245–63.

Attas, Daniel. "A Transgenerational Difference Principle." In Gosseries and Meyer, *Intergenerational Justice*, 189–218.

Barry, Brian. "Circumstances of Justice and Future Generations." In *Obligations to Future Generations*, edited by Richard Sikora and Brian Barry, 204–48. Philadelphia: Temple University Press, 1978.

———. "Justice between Generations." In *Law, Morality and Society: Essays in Honor of H. L. A. Hart*, edited by P. M. S. Hacker and Joseph Raz, 268–84. Oxford: Clarendon Press, 1977.

———. *Theories of Justice: A Treatise on Social Justice*. Vol. 1. Berkeley: University of California Press, 1989.

Cowen, Tyler. *Stubborn Attachments*. San Francisco: Stripe Press, 2018.

English, Jane. "Justice between Generations." *Philosophical Studies* 31, no. 2 (February 1977): 91–104.

Freeman, Samuel. *Rawls*. London: Routledge, 2007.

88 Consequentialist considerations cut both ways, of course. If pursuing continuous economic growth is a hindrance to securing other social primary goods (as Rawls's comments suggest), then continuous economic growth would not be a requirement of social justice.

89 I would like to thank Jimmy Goodrich, Dan Hausman, Paul Kelleher, David O'Brien, Andrew Williams, and two anonymous referees for helpful comments and discussion.

Friedman, Benjamin M. *The Moral Consequences of Economic Growth*. New York: Knopf Doubleday Publishing Group, 2005.
Gardiner, Stephen M. "A Contract on Future Generations?" In Gosseries and Meyer, *Intergenerational Justice*, 77–118.
Gaspart, Frédéric, and Axel Gosseries. "Are Generational Savings Unjust?" *Politics, Philosophy and Economics* 6, no. 2 (June 2007): 193–217.
Gosseries, Axel. "Three Models of Intergenerational Reciprocity." In Gosseries and Meyer, *Intergenerational Justice*, 119–46.
Gosseries, Axel, and Lukas H. Meyer, eds. *Intergenerational Justice*. Oxford: Oxford University Press, 2009.
Hare, R. M. "Rawls' Theory of Justice—II." *Philosophical Quarterly* 23, no. 92 (July 1973): 241–52.
Harsanyi, John C. "Can the Maximin Principle Serve as a Basis for Morality? A Critique of John Rawls's Theory." *American Political Science Review* 69, no. 2 (June 1975): 594–606.
———. "Cardinal Utility in Welfare Economics and in the Theory of Risk-Taking." *Journal of Political Economy* 61, no. 5 (October 1953): 434–35.
———. "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility." *Journal of Political Economy* 63, no. 4 (August 1955): 309–21.
Heath, Joseph. "The Structure of Intergenerational Cooperation." *Philosophy and Public Affairs* 41, no. 1 (Winter 2013): 31–66.
Heyd, David. "A Value or Obligation? Rawls on Justice to Future Generations." In Gosseries and Meyer, *Intergenerational Justice*, 167–88.
Hubin, D. Clayton. "Justice and Future Generations." *Philosophy and Public Affairs* 6, no. 1 (Autumn 1976): 70–83.
———. "Minimizing Maximin." *Philosophical Studies* 37, no. 4 (May 1980): 363–72.
———. "Non-Tuism." *Canadian Journal of Philosophy* 21, no. 4 (1991): 441–68.
Knight, Frank H. *Risk, Uncertainty, and Profit*. Boston: Houghton Mifflin Company, 1921.
Mazor, Joseph. "Liberal Justice, Future People, and Natural Resource Conservation." *Philosophy and Public Affairs* 38, no. 4 (Fall 2010): 380–408.
Meyer, Lukas. "Intergenerational Justice." *Stanford Encyclopedia of Philosophy* (Summer 2021). https://plato.stanford.edu/archives/sum2021/entries/justice-intergenerational/.
Moehler, Michael. "The Rawls-Harsanyi Dispute: A Moral Point of View." *Pacific Philosophical Quarterly* 99, no. 1 (March 2018): 82–99.
Parfit, Derek. "Equality or Priority?" In *The Ideal of Equality*, edited by Matthew Clayton and Andrew Williams, 81–125. London: Palgrave Macmillan, 2002.
———. *Reasons and Persons*. Oxford: Oxford University Press, 1984.

Phelps, Edmund. "The Golden Rule of Accumulation: A Fable for Growthmen." *American Economic Review* 51, no. 4 (September 1961): 638–43.

Rawls, John. *Justice as Fairness: A Restatement*. Cambridge, MA: Harvard University Press, 2001.

———. *The Law of Peoples, with "The Idea of Public Reason Revisited."* Cambridge, MA: Harvard University Press, 1999.

———. *Political Liberalism*. Expanded ed. New York: Columbia University Press, 2005.

———. *A Theory of Justice*. 1st ed. Cambridge, MA: Belknap Press, 1971.

———. *A Theory of Justice*. Rev. ed. Cambridge, MA: Harvard University Press, 1999.

Romer, Paul M. "Endogenous Technical Change." *Journal of Political Economy* 98, no. 5 (October 1990): S71–102.

Sauvé, Kevin. "Gauthier, Property Rights, and Future Generations." *Canadian Journal of Philosophy* 25, no. 2 (June 1995): 163–76.

Scanlon, T. M. "Rawls on Justification." In *The Cambridge Companion to Rawls*, edited by Samuel Freeman, 139–67. Cambridge: Cambridge University Press, 2002.

Solow, Robert M. "A Contribution to the Theory of Economic Growth." *Quarterly Journal of Economics* 70, no. 1 (February 1956): 65–94.

Wall, Steven. "Just Savings and the Difference Principle." *Philosophical Studies* 116, no. 1 (October 2003): 79–102.

Journal of Ethics and Social Philosophy
VOL. 27, NO. 1 · APRIL 2024
https://doi.org/10.26556/jesp.v27i2.2434

RATIONALITY, SHMATIONALITY

EVEN NEWER SHMAGENCY WORRIES

Olof Leffler

THIS PAPER takes aim at constitutivist theories of the normativity of structural norms of rationality. Put generally yet briefly, constitutivists attempt to explain the force or applicability of various types of norms by appealing to how they constitute agency. The number of accounts of moral norms based on this strategy has recently skyrocketed.[1] However, constitutivism can in principle be used to formulate theories about other norms, too, and I shall focus on constitutivism about the applicability and force of structural principles of rationality.[2] This includes principles of means-ends coherence (sometimes called "instrumental rationality"), *enkrasia*, and the like.

Attempts to explain the applicability and force of principles of structural rationality using constitutivist means make much dialectical sense. The normative force of such principles has been much disputed. Some philosophers deny their force, and others even deny that they exist independently of normative reasons.[3] In virtue of such skeptical challenges, one may wonder if constitutivism may come to the rescue. Despite not endorsing constitutivism himself, John Broome writes:

1 See, for example, Katsafanas, *Agency and the Foundations of Ethics*; Korsgaard, *The Sources of Normativity* and *Self-Constitution*; Smith, "Agents and Patients," "The Magic of Constitutivism," and "Constitutivism"; Velleman, *How We Get Along*; and Walden, "Laws of Nature, Laws of Freedom, and the Social Construction of Normativity."

2 See, for example, Bratman, "Intention, Belief, and Instrumental Rationality," "Intention, Belief, Practical, Theoretical," "Intention, Practical Rationality, and Self-Governance," and *Planning, Time, and Self-Governance*; Brunero, *Instrumental Rationality*; Goldman, *Reasons from Within*; Korsgaard, *Self-Constitution*; Roughley, *Wanting and Intending*; Smith, "The Explanatory Role of Being Rational" and "A Puzzle about Internal Reasons"; and Southwood, "Vindicating the Normativity of Rationality" and "Constructivism and the Normativity of Practical Reason."

3 For the former, see Kolodny, "Why Be Rational?"; and Lord, *The Importance of Being Rational.* For the latter, see Henning, *From a Rational Point of View*; Kiesewetter, *The Normativity of Rationality*; and Raz, "The Myth of Instrumental Rationality."

> An account of the nature of rationality might imply that rationality is normative. For instance, it is plausible that rationality is constitutive of agency, so that if we were not rational we would not be agents. It may be that, being the acting creatures we are, we cannot help taking rationality as normative. If so, an argument might be built on that fact for the conclusion that rationality actually is normative.[4]

A common objection to constitutivism about moral norms, however, charges it with failing to explain why we cannot be so-called shmagents—namely, very much like agents but without commitments to the constitutive features of agency that would explain why we are subject to the norms that are constitutive of agency.[5] One may, then, very reasonably wonder whether shmagents also generate problems for constitutivism about structural rationality. Exactly that is what I shall argue, at quite some length, in this paper.

"Shmagency" worries gain particular pertinence because the shmagency objection has recently been the subject of much debate. In response to Enoch's original worry, many have argued that constitutivist norms are inescapable or valuable, thus immunizing them from the challenge and possibly even explaining their normativity.[6] But, simultaneously, novel versions of the challenge have been launched.[7] These developing and, in some ways, more sophisticated versions of the shmagency objection set the stage for this paper. Utilizing and extending them further, we can articulate shmagency worries that cause problems for constitutivism about structural rationality.

I start in section 1 by outlining constitutivism about structural rationality. In section 2, I outline the key attractions of that view. In section 3, I introduce the shmagency objection and develop two versions that generate problems for constitutivism about structural rationality. In the following sections, I apply them to several constitutivist views. Section 4 is dedicated to what I call first-person authority views, section 5 to single-mental-state views, and section 6 to systems-of-mental-states views. All these accounts of the normativity of structural norms of rationality suffer from the two shmagency objections articulated and defended in section 3. I wrap up in section 7.

4 Broome, *Rationality through Reasoning*, 204.

5 Enoch, "Agency, Shmagency" and "Shmagency Revisited."

6 For example, Ferrero, "Constitutivism and the Schmagency Challenge" and "Inescapability Revisited"; Katsafanas, *Agency and the Foundations of Ethics*; Korsgaard, *Self-Constitution*; Smith, "The Magic of Constitutivism"; and Velleman, *How We Get Along*.

7 Enoch, "Shmagency Revisited"; Leffler, "New Shmagency Worries"; and Tiffany, "Why Be an Agent?"

1. CONSTITUTIVISM ABOUT STRUCTURAL RATIONALITY

To discuss shmagency objections to constitutivism about structural rationality, it will help to first say something about which views count as constitutivism about structural rationality. Broadly speaking, I take constitutivism to involve a type of explanation of the force or applicability of various types of norms that appeals to how they are involved in constituting agency. But it will help to be more specific.

We will first need to narrow down our subject matter. As mentioned above, there are many types of constitutivism. Here, however, we are concerned with constitutivism about norms of structural rationality. These are norms of coherence that govern the structural rationality of combinations of mental states for agents. Typical examples include the following:

> *Instrumental Irrationality*: If A intends to ϕ, and A believes that ψ-ing is a necessary means to ϕ-ing, and A does not intend to ψ, then A is irrational.
>
> *Modus Ponens*: If A believes that p, and A believes that $p \rightarrow q$, and A does not believe that q, then A is irrational.[8]

I shall use these two norms to illustrate structural rationality. While they are formulated negatively in the sense that they specify when A is irrational, they can also easily be reformulated into positive requirements of rationality if one takes an agent to be, in relevant ways, in at least one respect instrumentally rational if they are not irrational in the way Instrumental Irrationality specifies, and in at least one respect epistemically rational if they are not irrational in the way Modus Ponens specifies.[9] As such, these norms are paradigmatic norms of structural rationality. While slightly different formulations of them may be given, they are the *type* of norms I am concerned with—yet there may, of course, also be other norms of the same type.

A second issue here is that the literature on structural rationality has been developing rapidly recently. This leaves it unclear which accounts of it one may want to count as constitutivist, and therefore also which versions are targeted by the shmagency objections. To clarify this, I shall introduce a schema for

8 These formulations of the principles are taken verbatim from Kiesewetter, *The Normativity of Rationality*, 15.

9 I do not, however, say that agents are rational *if and only if* they would not be irrational according to these norms, for there are presumably other norms of rationality, cases of arationality, or possibly even nonstructuralist aspects of a full theory of rationality—such as responsiveness to reasons—that they do not capture. For a view that incorporates both structural rationality and reasons responsiveness, see Worsnip, "What Is (In)coherence?" and *Fitting Things Together*.

constitutivism about structural rationality, and I shall use it below to show how various accounts of rationality are constitutivist.[10] The schema is:

> *Structural Rationality Constitutivism*: An account *T* of structural rationality *S* is constitutivist iff *T* entails that *S* is normative because *S* is, or is normative *in virtue of*, some property or properties of the constitutive feature or features *C* of an aspect of agency *A*, where *C* constitutes something as an *A*.

The variables mean the following:

> *T* = a theory that aims to explain the normativity of *S*.
> *S* = principles of structural rationality.[11]
> *C* = a constitutive aim, principle, or other relevant constitutive feature or features of agency.[12]
> *A* = anything conventionally associated with agency, such as action, agency itself, propositional attitudes, or selfhood.[13]

Some clarifications will also be helpful. First, by "*S* is normative," I mean to stipulate what "normative" is for present constitutivist purposes. That can be one or both of the following things: why structural principles hold for or apply to an agent or why they have normative force for her. "Holding for or applying to an agent" indicates that an agent is subject to the norm, and "having normative force for her" means that there is a way in which a norm authoritatively prescribes something for the agent.[14] It is sometimes unclear which of these

10 The schema and characterization are adapted from my *The Constitution of Constitutivism*, ch. 1.

11 I use the terms "principles," "norms," and "requirements" interchangeably here.

12 Here, a constitutive aim means that a goal constitutes some aspect of agency. For example, a belief might be constituted by aiming at truth. Constitutive principles are slightly different: perhaps the categorical imperative is constitutive of agency as per Korsgaard's *The Sources of Normativity* and *Self-Constitution*. But that does not mean that one aims at principles in the same way as truth might be the aim of belief: principles rather structure reasoning. For more on the distinction, see Katsafanas, *Agency and the Foundations of Ethics* and "Constitutivism about Practical Reasons."

13 For simplicity and readability, I sometimes lump these aspects of agency together under the umbrella term "agency."

14 The language of "normative force" and "authoritative prescriptivity" here could easily be contested, but I am using it stipulatively. What I am after is the extra property of norms in virtue of which they bind agents to following them independently of what the agents want themselves, but the literature is unclear on how to label that and indeed on how the property should be characterized. Even different constitutivist views imply different things. We can, however, bring out what I have in mind using a nonconstitutivist analogy. Foot famously takes the rules of etiquette and ethics to apply categorically to agents, so that

properties philosophers have in mind when they discuss whether rationality is normative, but context should make clear what I have in mind below.

Second, it matters that *S* is normative in virtue of some *property or properties* of the constitutive features *C*. *S* need not be normative *just* in virtue of the constitutive feature or features themselves: some writers on the normativity of rationality indicate this, whereas others do not.[15] But we can take some inspiration from the literature on constitutivism about moral norms to see that there often is a deeper underlying property that does explanatory work here.

This is so because many philosophers assume that the constitutive features only need to serve to *transmit* normativity from some other source.[16] Indeed, many think that some aspect of agency is independently valuable or inescapable and that *that* is what explains why their norms have force.[17] This is so even though value or inescapability need not be constitutive of agency. A full constitutivist explanation of a norm such as the categorical imperative (CI) being constitutive of agency might, instead, say that CI is constitutive of agency, agency is in some relevant sense inescapable, that kind of inescapability

---

they are always subject to them, but denies that the norms intrinsically have what I am here calling normative force or authoritative prescriptivity, so that agents need not follow them absent something external to etiquette or ethics itself (such as a reason) that prescribes that they do so ("Morality as a System of Hypothetical Imperatives"). What I am after with "authoritative prescriptivity" is the extra property of norms of structural rationality that would make them such that agents are bound to follow them independently of what they want themselves—and that Foot denied that etiquette and ethics have. The reader is however free to plug in their own terminology or characterization instead, perhaps calling it "normative oomph" or maybe "categoricity" (though not in Foot's sense).

Nevertheless, like me, many constitutivists are after this extra thing about some norms that is supposed to bind agents beyond their being subject to the norms, whether they are talking about rationality, morality, or something else. In this search, a strength of constitutivism is that it need *not* be committed to interpreting normative force as reason-givingness: normative force *qua* authoritative prescriptivity is more general than that. While some may think it consists of giving a reason, most appear to think that constitutive aims or principles are likely to possess some kind of normative force other than that, such being inescapable or valuable. In fact, the normative force of reasons is itself something that constitutivists might be inclined to explain using the constitutive features of agency (cf. Korsgaard, *Self-Constitution*). I return to this point in section 2 below.

15 For the former, see, for example, Brunero, *Instrumental Rationality*. For the latter, see, for example, Bratman, "Intention, Practical Rationality, and Self-Governance" and *Planning, Time, and Self-Governance*; and Roughley, *Wanting and Intending*.

16 Ferrero, "The Simple Constitutivist Move."

17 For example, Katsafanas, *Agency and the Foundations of Ethics*; Korsgaard, *Self-Constitution*; Velleman, *How We Get Along*; and Smith, "The Magic of Constitutivism"; cf. Ferrero, "Inescapability Revisited" and "The Simple Constitutivist Move."

explains normativity—and *therefore*, CI is normative.[18] An analogous line of argument can be developed using value.[19] Here, the idea is that some norms are constitutive of some valuable form of agency and are therefore valuable themselves. In either case, it is not being constitutive of agency by itself that explains normativity; rather, being so *transmits* normativity.

2. WHY CONSTITUTIVISM ABOUT RATIONALITY?

Constitutivism is now introduced. But why care? In the introduction, I indicated that it might serve to explain the normativity of norms of rationality. We may disentangle and expand on that point, for it is in fact based on several reasons to care about constitutivism. Of these, several will matter greatly in the critical discussion below.

A first, very general, reason to be interested in constitutivism about structural rationality is that constitutivism might be independently attractive. Perhaps one holds a general constitutivist position in the philosophy of action or thinks beliefs very plausibly are constituted by aiming at truth. If one simultaneously thinks that the normativity of structural rationality ought to be explained, one had better come up with an explanation that fits this picture.

A second and more specific point is that constitutivism might seem explanatorily *promising* with respect to some more important phenomenon, such as normative force. This seems to be what John Broome hints at in the quotation in the introduction. Perhaps one thinks constitutivism seems like a strong contender when it comes to the normative force of moral norms, and so, then, that it might also be a strong contender regarding norms of rationality.

Third, constitutivism might seem *especially* promising for explaining normative force because it need not do so using normative reasons. Much has been written about the relation between structural principles of rationality and normative reasons, but the normative force of rationality need not be understood in terms of reasons on constitutivist accounts: in fact, constitutivism is sometimes used to explain the force of reasons itself.[20] As indicated above,

18 This is the picture in Korsgaard, *Self-Constitution*. To be clear, the version of CI that Korsgaard thinks is most deeply constitutive of agency is the formula of universal law (roughly: "act only in a way such that your maxims could be made into universal law"). She also thinks that the hypothetical imperative, or HI (roughly: "take means to ends as a necessary feature of forming a will, on pain of irrationality"), is so constitutive. I return to both imperatives below.

19 The most paradigmatic example is Smith, "The Magic of Constitutivism."

20 The most paradigmatic example is Korsgaard, *Self-Constitution*.

constitutivists can perhaps take that force to depend on something entirely different from reasons, such as value or inescapability.

Fourth and finally, one might think norms of rationality are universally binding in the sense that they hold for all relevant entities, whichever they are: presumably all agents, at the very least.[21] If *S* norms are constitutive of agency, one might be tempted by a constitutivist explanation to guarantee universality, for if one takes *S* to be constitutive of some aspect of agency shared by all agents, constitutivism appears to guarantee it for them.

There are, then, at least four reasons to be interested in constitutivism about structural rationality. They generate a *prima facie* case for developing constitutivism about structural rationality, and several indicate that constitutivism is an attractive contender for generating the right kind of explanation of normative force. But whether constitutivism works is still an open question. In moral philosophy, it has suffered significant pushback, not least from the shmagency objection.[22] I now turn to it.

3. SHMAGENCY

The shmagency objection is probably the most prominent argument against constitutivism about morality.[23] I shall briefly introduce it, consider the two leading objections to it, and then show how responses to these lead to new and more sophisticated shmagency worries. I start with the reply from dialectical inescapability, which leads to shmagency as modal escapability. Then I turn to the reply from value, which leads to shmagency as underdetermination. It is these types of shmagency that, I argue, create major problems for constitutivism about structural rationality.

First things first. The shmagency objection is based on the idea that one can be very much like an agent without quite being one.[24] Hence, one might

21 For example, Brunero, *Instrumental Rationality*, ch. 7; and Way, "Reasons and Rationality."

22 See, for example, Enoch, "Agency, Shmagency" and "Shmagency Revisited"; Leffler, "New Shmagency Worries"; and Tiffany, "Why Be an Agent?"

23 There are also many other prominent objections, such as whether the constitutive aims individual constitutivists propose are plausible, and the problem of bad action, according to which it is unclear how we can act poorly if action is constituted by following some norm. The latter has often been presented as the main problem for constitutivism about rationality, such as by Kolodny ("Why Be Rational?") and Wedgwood (*The Value of Rationality*). For surveys of constitutivism in moral philosophy, including extensive further references to discussions of these problems, see Katsafanas, "Constitutivism about Practical Reasons"; Leffler, *The Constitution of Constitutivism*; Smith, "Constitutivism"; and Tubert, "Constitutive Arguments."

24 Enoch, "Agency, Shmagency."

be able to escape whichever norms are constitutive of agency. Consider chess. Chess has rules, and it also has aims—plausibly, to win, or at least to draw if one cannot win. But why play chess and be committed to its rules and aims? Perhaps one just does not care about winning and rather would prefer to go do something else. The same question can be asked about agency. Why be an agent and committed to its rules or aims? In other words, why be an agent rather than a shmagent?

There are two dominant responses to this worry. The first is to argue that agency is *dialectically inescapable*: it is such that attempting not to be an agent still involves agency and hence a self-contradiction. The second relies on positing some *value* that explains why agency is valuable and, hence, justifies agency rather than shmagency. I start by discussing the dialectical inescapability reply and proceed to defend the modal escapability worry to which it gives rise.

Many constitutivists argue that norms that are constitutive of agency are *dialectically inescapable*.[25] To clarify this point, we should start with a distinction between two perspectives from which we may wonder whether to be agents or shmagents: an *internal* and an *external* perspective. Asking the shmagency question internally is ordinarily considered unproblematic: doing so is for an agent to ask whether they have reason to be an agent, but then they do that while committed to the norms of agency.

The external question is different, but dialectical inescapability is thought to block the possibility that one could take up a standpoint external to agency and ask whether one should be an agent. This is because insofar as one is an agent, one cannot escape agency by deciding not to become one on pain of self-contradiction.[26] Hence, insofar as one is an agent, one cannot get out of one's agency without exercising one's agency—the act of escaping it is also subject to its norms. This is why agency differs from chess.

But there are new shmagency worries.[27] In response to the dialectical inescapability point, one might think that the real issue is not escapability for actual agents but rather *modal escapability*. Suppose someone is very much like an agent but not an agent by constitutivist standards from the start. They might be a *sophisticated shmagent*. Consider, then, an ambitious view such as Korsgaard's, according to which agency commits us to the categorical and hypothetical imperatives (CI and HI, respectively), and we explain the force of CI

25 Ferrero, "Constitutivism and the Schmagency Challenge"; cf. Leffler, "New Shmagency Worries."

26 As Ferrero puts it: agency is *the enterprise of the largest jurisdiction*, covering all actions, and *closed under reflection*, so that reflecting on or acting so as to escape agency still involves a commitment to its norms ("Constitutivism and the Schmagency Challenge").

27 Leffler, "New Shmagency Worries."

and HI by saying that they are constitutive of our agency together with some background premises such as the claim that acting, and hence our agency, is our inescapable plight.[28] This view appears unable to explain why CI still seems to bind shmagents whose psychologies do not constitutively feature it. In fact, for that very reason, it does not seem to apply to or have force for them. Such shmagents do, therefore, occupy a position external to agency—yet they might quite reasonably ask whether they should be agents or shmagents from that external perspective, for they may wonder whether they should take on a norm such as CI (or HI).

There are many ways to cash out this possibility, for there are many psychological profiles that lack constitutivist commitments. For example, instead of commitments to CI and HI, a Martian shmagent could have a Humean belief/desire psychology, where their movements ordinarily are explained by being caused by belief-desire pairs in the right way. Or a Saturnian shmagent could have a besire-based psychology, where their movements are explained by a mental state that both represents some fact and aims to make the agent realize that fact (probably together with extra means-beliefs).

Sophisticated shmagents indicate that constitutivism is extensionally inadequate for two reasons. First, even though sophisticated shmagents do not count as agents according to views like Korsgaard's, one can easily stipulate that "they are intelligent; are knowledgeable; perform what looks a lot like actions for what looks a lot like reasons; are capable of (what seems to be) deliberation and reflecting on what they do; and are able to prefer different behaviours."[29] If so, a norm such as CI should apply to and have force for them just as much as ordinary agents: they appear sophisticated enough to be part of our normative practices. But constitutivists who think agency has significant normative commitments, such as Korsgaard, cannot explain that.[30]

Second, the final reason mentioned in section 2 for going constitutivist applies here too. Many want to explain norms with universal normative force: they apply to and have force for all relevant entities. This is so whether we

28 Korsgaard, *Self-Constitution*, 1–2.

29 Leffler, "New Shmagency Worries," 132–33.

30 An important addendum is that many think that Kant's, or at least Korsgaard's, view is not just a view of moral principles but also a view of principles of rationality. It is unclear whether these should count as principles of structural rationality, but if they do, one might wonder whether a Kantian approach to rationality might work if one opts for a constitutivist explanation of *structural rationality*, whatever we make of morality. However, the very fact that Korsgaard's view is the standard example of a constitutivist view of morality that does not seem to deliver in the face of shmagency worries indicates that it will have problems on the rationality side too.

discuss morality or rationality. For example, Kantians presumably want to explain the force of CI for creatures whose behavior should be explained by the Humean theory of motivation or besires but lack intrinsic commitments to CI. The same seems true regarding norms of rationality. So it is not just that we should include some shmagents in our practices. Perhaps constitutivism explains the force of norms for too few entities.

Here, one might suspect that Korsgaard's so-called plight inescapability can help with shmagency as modal escapability. Plight inescapability says, roughly, that agents continuously face new situations where they have to act, which means that not taking action is also a way of taking action: agents are always bound to live up to the norms of agency, even if they try not to. But plight inescapability will not help. It is quite possible to be a shmagent who has to live up to the plight of shmagency, for *what* one has to live up to depends on how one is constituted, not on whether living up to it is inescapable.[31] If one has a belief/desire- or besire-based psychology, then that ends up being what one has to live up to. So constitutivist agency seems modally escapable whether or not it is plight inescapable.

Now to the second of the two novel shmagency worries. It, too, can be developed in response to a constitutivist response to the original objection. This time, the constitutivist response is that constitutivist-style agency is relevantly *valuable* (or otherwise normatively justifiable—feel free to trade in your value coins for some other normative currency here, but I shall use the language of value for simplicity).[32] Whether or not they intend to endorse it formulated in exactly this way themselves, versions of this argument are implicit in, *inter alia*, positions taken by Michael Smith and Michael Bratman.[33] Their thought is that normatively constituted agency is valuable, so it is *ipso facto* valuable to conform to the norms that constitute it. We can then reply to the shmagency objection by saying that being an agent rather than a shmagent is normatively valuable for to us, or that the value of agency matters more than that of shmagency.

In response to this worry, I formulated another new shmagency objection that I called shmagency as underdetermination.[34] That formulation, it has turned out, was unfortunately somewhat obscure, so I shall attempt to develop

31 I suspect that *no* type of inescapability is fully modally escapable because of the reason given in the main text. But discussing all possible types of inescapability that constitutivists have proposed would take us too far afield. See Ferrero, "Inescapability Revisited" and "The Simple Constitutivist Move" for important distinctions, however.

32 Previously, I toyed with calling this type of justification "normative inescapability" (Leffler, "New Shmagency Worries"). But there is something to be said for skipping that label: there are too many uses of "inescapability" anyway.

33 Smith, "The Magic of Constitutivism"; Bratman, *Planning, Time, and Self-Governance*.

34 Leffler, "New Shmagency Worries," 140–43.

it. Its core point is this: to reply to the shmagency objection, constitutivists who appeal to value are committed to saying that the value of agency supports or justifies being an agent *rather than* a shmagent, or at least is such that the value of agency matters much more than that of shmagency (e.g., we are to maximize the value of agency but not of shmagency; so, assuming their values count for as much *ceteris paribus*, the former now outweighs the latter). For if they would not, they would not have shown that the value in question justifies being an agent rather than a shmagent, so the value-based response to the question of whether to be an agent rather than a shmagent would not show that it is agency rather than shmagency that is justified. This means that constitutivists who opt for the value-based response to the shmagency objection need to show that the value of agency justifies agency rather than shmagency. But they do not. Sometimes, the value in question lends equal support to both, and sometimes, the value even supports shmagency rather than agency. So it is underdetermined whether the value supports agency rather than shmagency.

Let me articulate this point in greater depth. The shmagency-as-underdetermination worry already accepts the assumption that what is constitutive of agency has some value. But the problem is that unless it is shown that it is valuable to follow the norms that are constitutive of agency *rather than* those of shmagency, we do not have a response to the shmagency objection, for nothing would support our being agents and following its norms rather than the norms of some shmagent. The value of agency would then not be significant enough to do the theoretical work it is supposed to do to reply to the shmagency objection. This would suggest that constitutivism is false.

We may, again, use a version of Korsgaard's Kantian constitutivism to exemplify the point. As mentioned, she thinks that all (human) agents always are committed to CI. (For simplicity, ignore HI for now.) Assume also, now unlike Korsgaard, that you were to justify CI with some value of your choice. For simplicity again, perhaps we bring about happiness in the world if we are agents who have CI as a feature of our psychologies.

It is, however, easy to think of occasions on which being committed to CI will not help to bring about happiness in a way that an alternative does not do just as well or better. Assume that some version of the golden rule, saying that one ought to treat one's neighbor as one would want to be treated oneself, makes people equally happy as CI proper. Then it seems just as valuable to be a shmagent committed to the golden rule as an agent committed to CI. So the value of happiness does not support CI of agency rather than the golden rule of shmagency.

Alternatively, assume that a murderer comes knocking on the door every Tuesday to ask about a friend who is inside, we have normal desires and

commitments, and CI requires us not to lie to the murderer. Then it seems that a value-based justification of CI would entail that we are better off being shmagents who are committed to CI on every day of the week except Tuesdays. Then we can lie to the murderer and be happy that our friend does not get murdered. But people whose psychology disposes them to act on the golden rule or on a rule like CI-*except-on-Tuesdays* are not agents on Korsgaard's account: she needs all agents to always be committed to acting on CI. They are, rather, shmagents.

Generalizing, unless a constitutivist can show that the value they appeal to supports agency rather than some sort of shmagency in the vicinity of agency, they have not shown that the value supports being committed to the demands of agency rather than some form of shmagency. As such, they have not shown why value supports or justifies agency rather than shmagency.

Objection: following CI is likely to bring about *some* amount of happiness, even if other norms could also be valuable in virtue of the happiness they bring about. Does that not mean that it would be valuable to go with CI after all, though perhaps *pro tanto* rather than all things considered? Yes, it would be. But again, the underdetermination objection accepts that agency has some value. What constitutivists need is the comparative claim that the value justifies or supports agency *rather than* shmagency (or that the value of agency needs to be treated as more important). Otherwise, they have not shown that it is agency rather than shmagency that is justified or supported. And this is quite orthogonal to the *pro tanto*/all-things-considered distinction: we may reasonably wonder whether the value in question supports CI rather than some alternative both *pro tanto* and all things considered. However, it is hard to see why agency would be more valuable than shmagency or have value that we would have to treat as more important than that of shmagency. It is very plausible that shmagency is just as valuable or even more valuable than agency. We see that with the golden rule or CI-except-on-Tuesdays examples.[35]

Another objection: Could the constitutivist perhaps stipulate that the value of rationality is to be maximized, making it look straightforward that constitutive norms of rationality will count for a lot? Unfortunately, maximizing that value rather than some other seems quite implausible. There are always cases of so-called rational irrationality.[36] If a burglar threatens to kill your family and you have to have the combinations of mental states that an irrational shmagent

35 This point is analogous to a familiar objection to rule consequentialism: much like it is rule fetishistic to cling to a rule justified by some value in a moral context when there is some other rule that brings about just as much or more of the value, it seems constitution fetishistic to say that whatever value supports our being constituted as agents supports it rather than other, equally or more valuable, shmagency constitutions.

36 For seminal discussion, see Parfit, *Reasons and Persons*.

rather than a rational agent would have to save them, the value of shmagency surely trumps that of rationality. In this case, it is more valuable to be a shmagent rather than an agent: being an agent might even have negative value. Yet it is hard to specify when and where shmagency might be equally or more valuable than agency. So constitutivists have much explanatory work to do if they want to show how it is agency rather than shmagency that is justified or supported by some value.

In summary, all this means that the shmagency objection remains deeply concerning for constitutivists in spite of their standard replies. Even worse, the new shmagency worries risk being problematic for various types of constitutivism beyond Korsgaard's—including, as I shall argue, various types of constitutivism about rationality. The new worries show that leading constitutivist replies from inescapability or value do not help to defend it, even though arguments for constitutivism about structural rationality often rely on inescapability or value.

Hence, I shall proceed to launch the shmagency challenges of modal escapability and underdetermination against constitutivism about the normativity of structural principles of rationality and argue that they go unmet. So constitutivists about structural rationality suffer from versions of these new shmagency worries. As a result, they appear unable to explain why the norms of structural rationality apply to and have force for all relevant entities.

4. First-Person-Privilege Views

I start with Nicholas Southwood's view.[37] It fits the constitutivist schema well: Southwood argues that requirements of structural rationality *S* are normative in virtue of being constitutive *C* of having a first-personal standpoint *A*.

For Southwood, a standpoint is "constructed out of our particular beliefs, desires, hopes, fears, goals, values, and so on, and relative to which things can go well or badly. Our standpoints describe what matters to us; they are ones in which we are invested."[38] It is because the requirements of rationality are constitutive of our standpoints that they apply to us.

Southwood also thinks that the normative force they have is a special kind of first-personal normative force. What that might be is unclear, but we can run

37 Southwood, "Vindicating the Normativity of Rationality" and "Constructivism and the Normativity of Practical Reason." For other criticism, see Broome, "Replies to Southwood, Kearns and Star, and Cullity"; Coons and Faraci, "First-Personal Authority and the Normativity of Rationality"; and Levy, "Does the Normative Question about Rationality Rest on a Mistake?"

38 Southwood, "Vindicating the Normativity of Rationality," 26.

with the idea for now, for there are deeper worries ahead. Southwood can be read as taking it to be constitutive of standpoints to adhere to norms of structural rationality. But could we be such that we just have something very much like standpoints (of Southwood's type) without committing ourselves to the norms? Call them "shmandpoints."

I think we can have shmandpoints. It does not seem like we, descriptively, necessarily have Southwood's first-personal standpoints in the sense that we have things that matter to us or we are invested in. Perhaps we can be easily swayed by fashion, whether in the form of winds of political rhetoric or just changing social mores more broadly, therefore attaining or retaining new desires, emotions, goals, and values. Maybe we even do that without responding to reasons: some people happily try out what is new just because they can. Or perhaps we are just inclined to change our minds without being responsive to reasons: one day we feel like taking a swim, on another like taking a walk. If so, things would not seem to matter to us or like we were invested in them, for if something did matter to us or we were invested in it, we would not be willing to give it up for no reason. Yet we may have ephemeral and fickle desires, emotions, goals, or values that do not require reasons to change. They may come and go without us having much commitment to them.[39]

If it is possible to have a shmandpoint rather than a standpoint, standpoints seem modally escapable. And there are certainly possible creatures who do. Call them Mercurians, though possible humans also fit the profile. By stipulation, they are born disposed to be easily swayed by fashion or otherwise with an inclination to change their minds for no reason. Again, they only have shmandpoints, not standpoints. From there, however, they can ask the question of whether to have shmandpoints or standpoints, and hence ask the shmagency question from an external point of view. This means that Southwood fails to explain the applicability and, therefore, force of norms of rationality for the Mercurians.

Southwood's view, therefore, seems extensionally inadequate in virtue of both motivations for the modal escapability version of the shmagency challenge mentioned above. First, we can stipulate that the Mercurians (or fickle humans) have all the other properties that make them appropriate to include in our normative practices—intelligence; knowledgeability; the ability to perform something like actions for things that are much like reasons; capacities for

39 One could attempt to stipulate that just having certain desires, emotions, goals, or values means that they matter to one or that one is invested in them in some thinner sense, such that one could still give them up for no reason. But it seems Moore contradictory to say "*x* matters to me, but I would be willing to give up *x* for no reason" or "I am invested in *x*, but I would be willing to give up *x* for no reason." Here we are no longer in the territory of mattering or investment.

deliberation and reflection; and preferences for different things—so the norms should apply to them. Second, because of this stipulation, it is also plausible that if we want to explain norms of rationality with universal force, they should apply to the Mercurians.

Southwood might reply that my focus is off: maybe there are Mercurians, but even their shmandpoints would be governed by norms of structural rationality, so the objection is beside the point. But presumably, the Mercurians need not be committed to a norm like Instrumental Irrationality either. Perhaps they sometimes—often enough to survive—manage to take means to ends because their desires direct attention to the means to their satisfaction rather than out of a separate capacity or disposition to do so. So the norm need not be *constitutive* of a shmandpoint consisting of ephemeral and fickle desires, emotions, goals, or values.

Another line of argument would be to claim that it is valuable to have standpoints, *ipso facto* taking the second of the two constitutivist reply routes to shmagency outlined in section 3. But it seems unlikely that having practical standpoints that we are invested in necessarily is going to be very valuable. One can easily see how changing one's values and normative commitments—including commitments to rationality—along with fashion trends could be prudentially beneficial. Or just think of the burglar case in which you have to be structurally irrational to save your family and add that there may be very many burglars in the world. In turn, this means that it is very hard to explain to what extent, if any, the value of having a practical standpoint supports having a standpoint rather than a shmandpoint. This opens up space for a version of the underdetermination worry. Why have standpoints rather than shmandpoints? Good question.

Southwood has, however, developed his view. Instead of discussing his earlier view further, we can make a fresh start with it. In more recent work, he suggests that the norms of practical reason apply and have force because they govern answers to the question of "what to do," where that is a question of truths that determine what "the thing to do" is.[40] "The thing to do" just means a correct answer to the question we attempt to answer when reasoning practically. This is the question of "what to do"—and the question of "what to do," Southwood adds, is the question one attempts to answer when one uses one's faculty of practical reason. It is not answered just by appealing to what is required by being an agent.

Southwood's new view does, therefore, not seem constitutivist. But it can be reformulated. Perhaps the theory *T* could say that the faculty of practical

40 Southwood, "Constructivism about the Normativity of Practical Reason."

reason is (at least partially) constitutive of agency, and, hence, that answering the "what to do" question *C* is (at least partially) constitutive of agency *A*—and that *that* involves norms of structural rationality *S*. Alternatively, perhaps practical reason *C*, including *S*, is an *aspect* of agency *A* whether or not it is constitutive of agency *simpliciter*. With such maneuvers, we can generate forms of constitutivism based on Southwood's later view. In virtue of the attractions of constitutivism in section 2, these are interesting to discuss regardless of which view Southwood now endorses.

However, the reformulated views have no responses to shmagency objections either. Let us start with the modal escapability worry. Instead of asking "what to do" questions about "the thing to do" guided by practical reasoning, it is easy to imagine creatures who could aim for "the thing to shmo" rather than "the thing to do" and make use of some ability of "practical shmeasoning" rather than "practical reasoning" to get there. And practical shmeasoning need not be guided by norms of structural rationality such as Instrumental Irrationality; it could be guided by some other set of norms instead, such as *Instrumental-Irrationality-Except-at-2:00 AM-on-Tuesdays*. Call that a principle of "shmationality."

Doing is modally escapable when contrasted with shmoing. At the very least, there can shmagents who have a faculty of practical shmeasoning that allows them to *shmo* using principles of shmationality rather than *do* using principles of rationality. These shmagents need not be incompetent or unsophisticated; in fact, it will be very hard to differentiate committed followers of Instrumental Irrationality from followers of Instrumental-Irrationality-Except-at-2:00 AM-on-Tuesdays, so they may very well ask the external shmagency question from their perspective. But then the normativity of Instrumental Irrationality is not explained in their case, much like how the normativity of norms constitutive of standpoints is not explained in the case of agents who are committed to shmandpoints. Nevertheless, as before, we want to include them in our normative practices and explain the force of norms of rationality for them. We can stipulate that they are sophisticated enough for that.

Furthermore, shmagency as underdetermination reappears here too. Perhaps it is valuable to be an agent who settles on the thing to do using norms of structural rationality. But when and to what extent? It is unclear why any value would support Instrumental Irrationality over Instrumental-Irrationality-Except-at-2:00 AM-on-Tuesdays, given their similarity. And things get trickier still at some worlds. Somewhere in modal space, an evil demon punishes us for eternity if we go with the former rather than the latter. So the underdetermination version of the shmagency objection applies here too. It is unclear why it would be more valuable to be a doer rather than a "shmoer," and hence we lack reason to suppose it is valuable to be agents (doers) rather than shmagents (shmoers).

To sum up, regardless of version, Southwood's first-person-privilege view seems rather implausible when construed as a form of constitutivism. It allows us to be shmagents both modally and evaluatively. In fairness, Southwood does not appear to treat his later view as constitutivist, and to the extent that it is not a constitutivist view, it might be off the hook from the objections presented here. But that simultaneously means that it might not have the potential theoretical benefits of constitutivism.

5. SINGLE-MENTAL-STATE VIEWS

It is now time to consider single-mental-state view. On such accounts, tokens of some particular type of mental state are wholly or partially constituted by some norm of rationality—as well as such that they can explain the applicability and force of that norm. For example, it might be constitutive of an intention to ϕ to be disposed to take the necessary means ψ one believes there are to ϕ, as per Instrumental Irrationality.[41] I shall consider three types of single-mental-state view of this kind, regarding beliefs, intentions, and desires. I shall first outline each and then argue that they suffer from the shmagency objections.[42]

5.1. Beliefs

Perhaps the most famous nonmoral constitutivism is constitutivism about belief. The idea here is that beliefs are (at least in part) constituted by aiming at truth.[43] While this thesis can be descriptive, and hence concern whether we

41 This brief description raises further theoretical questions that also appear for the views I shall discuss below. (I want to thank an anonymous reviewer for pushing me to make them explicit.) For one example, is it possible to intend without taking known necessary means, and therefore even possible to intend instrumentally irrationally? This is a version of the problem of bad action for constitutivism but applied to intentions rather than actions: presumably, constitutivists need an answer (cf. note 23 above). Other questions are familiar from the literature on norms of structural rationality: for example, should we cash out Instrumental Irrationality in wide- or narrow-scoping terms, such that *A* either can stop intending to ϕ or start to intend to ψ; or does *A* have to start to intend to ψ on pain of irrationality? Fortunately, we can sidestep these concerns here: they are orthogonal to shmagency worries.

42 Some philosophers sometimes appear to embrace versions of both single-mental-state views and what I call system-of-mental-states views below, including Brunero (*Instrumental Rationality*), Bratman (*Planning, Time, and Self-Governance*), and Goldman (*Reasons from Within*). Sometimes they talk about the constitutive norms of token mental states, and sometimes about systems fitting together. Insofar as I discuss both types of views in different places, I will also cover their views throughout my discussion.

43 This is a familiar view and references snowball quickly. For now, however, see Bratman, "Intention, Belief, and Instrumental Rationality," "Intention, Belief, Practical, Theoretical," and *Planning, Time, and Self-Governance*; Brunero, *Instrumental Rationality*; Railton, "On

tend to actually believe the truth, the important part here is normative. The idea is that beliefs are constituted wholly or in part by some norm according to which we aim to get at the truth. Perhaps we aim to believe *p* based on the evidence for the truth of *p*, or give up *p* if the evidence contradicts *p*. Responding to the evidence in such ways could be at least part of what it *is* to believe that *p*.

On this account, we can hope to explain at least epistemic norms such as Modus Ponens. As Modus Ponens is truth preserving, if one aims at believing the truth, one should form a belief that *q* on pain of irrationality if one believes that *p* and that $p \rightarrow q$. Possibly, however, beliefs may also contribute to explaining structural norms of practical rationality. If one is a so-called cognitivist about norms of practical rationality, one takes action to involve belief in some relevant manner, and it is the norms of belief that explain norms of practical rationality.[44] For example, if intentions are a species of belief, a norm such as Instrumental Irrationality could be cashed out as saying that one is irrational if one holds inconsistent beliefs and does not make one's intentions consistent with beliefs about the necessary means to satisfying them.

Understood as such, the single-mental-state view of norms of rationality constituting beliefs fits the constitutivist schema. The theory *T* of some norms of structural rationality *S* takes them to be constitutive *C* of belief, and these are an aspect of our agency *A*. Probably, additional assumptions are needed to explain why the norms have force—perhaps believing is inescapable or valuable—but at least we can explain why they *apply* to people.

It is, however, perfectly possible to have "shmeliefs" rather than beliefs, in the sense that one has mental states about reality that need not be responsive to a truth norm. There are many possibilities familiar from the literature on belief and belief-adjacent mental states that do not seem esoteric enough to have to be the mental states of aliens: ordinary humans quite possibly instantiate them all. I shall focus on three examples.

First, it seems possible to entertain thoughts—in a broad, colloquial sense—without a truth aim. The summarizing label of "entertaining" is mine, but it usefully brings together many nonbelief attitudes one may contrast with belief.[45] For example, one may entertain a proposition by *pretending* that it is true when

---

the Hypothetical and Non-Hypothetical in Reasoning about Belief and Action"; Velleman, "On the Aim of Belief"; Velleman and Shah, "Doxastic Deliberation"; and Williams, "Deciding to Believe."

44 Harman, "Practical Reasoning" and *Change in View*; Setiya, "Cognitivism about Instrumental Reason"; Velleman, *Practical Reflection* and "What Good Is a Will?"; and Wallace, "Normativity, Commitment, and Instrumental Reason."

45 For inspiration and cases, see Velleman's "On the Aim of Belief" and the follow-up paper by Velleman and Shah, "Doxastic Deliberation."

playing with a child. Or one may entertain propositions *for argument's sake* without aiming at truth with the acceptance of that proposition. Or one may entertain a proposition as a *hypothesis*, holding it conditionally on further testing but nevertheless independently of its truth value when forming it.

Second, perhaps it is possible to "alieve" something without believing it.[46] Aliefs are mental states with a special kind of associative structure: they involve affective, representational, and behavioral content activated by the environment. If it is dark outside, one may alieve that spooky ghosts are on the move, that going outside is dangerous, and that one should not go out for a walk. But aliefs are not beliefs. They need not involve regarding something as true. This does, in fact, explain at least some cases where they stand in tension with beliefs, as they are thought paradigmatically to do. One may truly believe going outside is safe yet alieve that it is not.

Third, perhaps there are intuitions that are *seemings* about reality without being beliefs.[47] George Bealer, for example, characterizes intuitions phenomenologically as intellectual seemings that are neither beliefs nor mere hunches: they present certain facts as necessary. Importantly, they do so fairly nonplastically, that is, without changing often even in response to evidence. For example, I have a strong intellectual seeming that the gambler's fallacy is not fallacious, despite learning the opposite when I studied statistics many years ago.

Assume, then, that there are creatures who only have belief-like mental states that are not guided by a truth norm, whether these are entertainings, aliefs, or seemings that *p*. You decide whether they are humans or Neptunians; in either case, they are shmelievers. The shmelievers need not be committed to altering their belief-like states in accordance with the evidence, for none of these mental states are constituted by a truth norm. The shmelievers are then, in a sense, external to agency, but they may very well wonder whether to be believers or shmelievers from their perspectives.

However, the truth norm for belief should be applicable to and have normative force even for the shmelievers. It seems epistemically outrageous that someone could have mental states about reality that are not regulated by the evidence: imagine being in a political discussion with someone who claims only to entertain, intuit, or alieve in propositions such that they see no need to respond to evidence that contradicts their claims. Yet a single-mental-state constitutivist cannot straightforwardly explain why evidential norms would apply to this person. Perhaps the norms in fact do, but if so, that would be in

46 Gendler, "Alief and Belief."

47 Bealer, "Intuition and the Autonomy of Philosophy"; cf. Huemer, "Compassionate Phenomenal Conservatism."

virtue of something other than the mental states of the shmelievers: for example, in virtue of the value of good public deliberation. Hence, the truth norm of belief is modally escapable and the constitutivist view extensionally inadequate.

Instead, one plausibly needs to appeal to some value of belief to defend norms of rationality as constitutive of belief. It seems extremely plausible that it is valuable, in general, to have (true) beliefs. They are crucial not just for good public deliberation but to represent means in action; they are likely to be integral to our identities; and, in cases such as Pascal's, they might even bring great rewards.

But then the underdetermination worry looms. Such pragmatic considerations do not say much about whether or when it is valuable to believe or shmelieve. This means that the norms of rationality implicit in belief remain underdetermined. Gesturing at the instrumental benefits of believing truly would not get us a view that tells us whether to be believers that *p* or shmelievers that *p*, for just speaking of instrumental benefits does not guarantee that the norms constitutive of beliefs *rather than* of shmeliefs will bring the benefits in question.

Now, there is of course a literature on instrumentalist justifications of epistemic norms.[48] Perhaps true beliefs matter in general because they are likely to be conducive to us achieving our aims. Or perhaps epistemic reasons constitutively are reasons to believe that *p* because they improve the satisfiability of our aims. Could a constitutivist not opt for a value-based justification of beliefs (or other epistemic phenomena) such as that?

Unfortunately, the problem with such views is that truth need not be what uniquely satisfies our aims; it has yet to be shown why being believers rather than shmelievers is justified. Sometimes it seems better to alieve in Santa than to believe that Santa does not exist. What kind of believers (or shmelievers) it is valuable to be in the light of potentially diverging concerns is exactly the question we are trying to answer. As such, shmagency as underdetermination is a pertinent worry here too.

Another response to shmagency as underdetermination is to argue that the deliberative question of whether to believe that *p* is transparent once the question of whether *p* is answered. The latter settles the former, so there is no further question to be asked about their interrelation: it is always truth that settles what to believe in deliberation. This could even rule out Pascal's wager cases, where practical benefits seem to come into play, since we may not want to count them as genuine deliberation.[49]

48 For some prominent examples, see Cowie, "In Defence of Instrumentalism about Epistemic Normativity"; and Kornblith, *Knowledge and Its Place in Nature.*

49 Velleman and Shah, "Doxastic Deliberation," 530n15.

However, this argument just pushes the underdetermination worry one step further down the line. Sure, you can define deliberation as involving only reasoning about genuine beliefs. But then the pragmatic question will instead become: Is it more valuable to be a deliberator or a "shmeliberator"?—namely, someone who "shmeliberates" rather than deliberates with the aim of holding nonbelief mental states such as entertainings, aliefs, or intuitions that are useful rather than true. So the underdetermination worry reappears. Therefore, I conclude that single-belief constitutivism suffers from both shmagency objections.

5.2. *Intentions*

Another common way to explain rational norms makes use of intentions. There are many theories of intentions, however, and I cannot discuss them all here. I shall instead assume that on the relevant accounts, intentions are part of the explanation of the force and applicability of the norms that constitute them and are distinct from beliefs or desires (treated in sections 5.1 and 5.3, respectively). If we are inspired by Bratman, for example, we might take intentions to be mental states that functionally aim to execute and coordinate our plans.[50]

The core idea on single-intentions accounts is that norms of structural rationality are constitutive of intentions. For example, Instrumental Irrationality is a plausible contender for that. It might very well be constitutive of an intention to be disposed to avoid that type of irrationality: if A intends to ϕ, and A believes that ψ-ing is a necessary means to ϕ-ing, and A does not intend to ψ, then A is irrational.[51] This would also make single-intentions views fit the constitutivist schema. Our theory T of some norms of structural rationality S, such as Instrumental Irrationality, is to treat them as constitutive C of some aspect of our agency—namely, intentions A.

Intentions construed as such are, however, modally escapable. Consider again the Martians and Saturnians. They are similar enough to Korsgaard-style constitutivist agents committed to CI, but they are not disposed to follow it. Instead, the Martians have beliefs and desires, and the Saturnians have besires (and beliefs). They seem similar enough to agents to be such that we should

50 Bratman, *Intention, Plans, and Practical Reason* and *Planning, Time, and Self-Governance*. Note, however, that Bratman's full view is complex and might be best interpreted as a systems-of-mental-states view, as per section 3 below.

51 Depending on how one fills in the details here, perhaps A does not truly intend to ϕ unless they also intend to ψ. On a weaker account, perhaps A has to start to *try* to form the new intention but need not necessarily succeed. The former possibility raises a version of the problem of bad action about intentions: cf. notes 23 and 41 above for details and references. But again, shmagency is a concern separate from these details.

explain how norms hold for them, yet they need not count as agents in Korsgaard's view, for they lack commitments to CI.

Mutatis mutandis, this line of reasoning carries over to intentions. It might be constitutive of action to act based on an intention that is constituted by a norm of rationality. But then there are many possible creatures who do something *very much like* acting, and in whose cases constitutivists about norms of rationality constitutive of intention fail to explain norms of rationality. It is quite possible that they have intentions in either some sense other than ordinary humans—perhaps belief-desire pairs count as Martian intentions—or that they lack intentions at all—perhaps belief-desire pairs should not be interpreted as intentions. Regardless, the Martians and Saturnians escape Instrumental Irrationality, so they can ask an external shmagency question. Yet modal escapability is problematic for by now familiar reasons. We can stipulate that the Martians and Saturnians are similar enough to plausibly be included in our normative practices, and we want to explain how the norms hold universally, not just for some relevant entities.

One may, however, be tempted to think that it is not so bad to avoid explaining a norm such as Instrumental Irrationality for the Martians and Saturnians.[52] After all, Martians and Saturnians do not have intentions construed as distinct mental states, and the thought here is that Instrumental Irrationality is constitutive of such intentions. Why should they have it?

Martians and Saturnians should have it because taking means to ends no doubt matters to them too, much like it does to humans with intentions. That is how "actions" are brought about according to the Humean picture that holds for the Martians, as well as, plausibly, according to the besire model of motivation that holds for the Saturnians. But then, Martians and Saturnians also seem able to have combinations of mental states where they have ends given by desires or besires that they fail to combine with relevant and available means beliefs, and hence seem irrational. Why this is problematic should be explained, whether we are concerned with entities who count as agents by some constitutivist standard or not.

Instrumental Irrationality is one way to articulate a norm of means-ends coherence, but there could also be others. Hence, the constitutivist who wants to explain the normativity of that norm using intentions faces a choice that stems from the Martians and Saturnians. They can either say that the normativity of Instrumental Irrationality should be given the same explanation for Martians, Saturnians, and humans, or they could opt for some other explanation regarding the Martians and Saturnians. The former option seems unpalatable,

52 I want to thank an anonymous reviewer for helping me think through this objection.

as Martians and Saturnians *ex hypothesi* lack intentions, but why should we not go with the latter?

There is an analogous worry regarding normative reasons for shmagents. When I discussed it previously, I argued that the "reasons" of sophisticated shmagents are similar enough, pretheoretically, to human reasons to seem apt to be given the same explanation.[53] However, this is less clear in the case of principles of rationality, as *ex hypothesi*, the Martians and Saturnians lack the intentions that might be partially constituted by Instrumental Irrationality. They are in this respect dissimilar to us. In fact, this point even risks undermining the reasons for which modal inescapability seems problematic. First, instead of treating them like us in our normative practices, perhaps their difference from humans with intentions means they can be part of our normative practices *in a different way*. Second, instead of explaining means-ends coherence with universal applicability to them and us, the dissimilarity might indicate that they no longer count as relevant entities in whose cases we need to explain it.

However, whether or not these reasons are ultimately undermined, I think we have good reason to strive for an identical explanation of the normativity of means-ends coherence for sophisticated shmagents and humans. This reason comes from theoretical virtue. A theory that explains how instrumental rationality works for us in one way but in other ways for sophisticated shmagents—presumably, in different ways depending on the different ways in which shmagents are not agents—seems awkwardly disunified.

Such a view would fail to possess many familiar theoretical virtues. It is not parsimonious, as it admits of several explanatory mechanisms rather than just one. It lacks theoretical unity for the same reason. It is ad hoc, as it seems to invite novel explanations in response to novel counterexamples (there is one for Martians, one for Saturnians, etc.). It is not conservative, as it does not integrate our new explanations with previous theories by explaining means-ends coherence norms for Martians and Saturnians in the same way as it explains them for humans. And perhaps most importantly, it lacks many aspects of explanatory power: it is very sensitive to changes in background conditions about how the psychology of some particular creature works; it lacks a precise *explanandum*, as it tries to explain different kinds of means-ends coherence; and it lacks cognitive salience, as it requires reformulation to explain means-ends coherence norms for many different conceivable psychologies.[54] So we had better avoid a disunified view.

53 See Leffler, "New Shmagency Worries," 130n26, 134–35.

54 Here I rely on the compelling approach to explanatory power from Ylikoski and Kuorikoski, "Dissecting Explanatory Power."

The modal escapability version of the shmagency objection remains, then. Could we instead reply to the original shmagency worry by appealing to the value of intentions? Then we might perhaps also explain why having them rather than being Martians or Saturnians without them is justified by their value. And intentions certainly seem to have their uses. First, as Bratman famously has emphasized, they are *efficient*: they help to coordinate our actions with ourselves socially and over time. Second, as Bratman has also emphasized, they may help us be *self-governing*. The thought, roughly, is that we can govern our own actions if we have long-term plans, whereas agents without plans fail to do so well. Third, we may think that intentions allow us to be part of our ordinary practices of moral responsibility, for adhering to the norms of practical rationality is what makes us the sources of our actions and therefore able to be held responsible and take responsibility ourselves. Call the latter *moral participation*.[55]

However, underdetermination worries nevertheless remain. It is unclear *how* and *when* it is valuable to have intentions. Even though intentions may have several kinds of value, which may make us subject to their implicit norms of rationality, cases of rational irrationality still abound: if a burglar will kill your family if you do not act in a way that would count as irrational with respect to your intentions, that surely outweighs whatever value the intentions may have. Or if that case is not extreme enough, perhaps an evil demon will punish you forever unless you act in a way that contradicts whichever norm might be constitutive of intention.

Cases multiply easily, and as long as they are possible, what is valuable here does not seem to be intentions that feature Instrumental Irrationality but rather following some norm similar to it that allows for relevant exceptions. In other words, these cases indicate that agency does not appear justified or supported by value, in contrast with shmagency. This is shmagency as underdetermination again.

Perhaps it could be replied that the values of intention add up. Efficiency, self-governance, and moral participation *together* may make it very valuable to have intentions governed by the right norms. That seems right, but unhelpful. For as long as there are possible cases of rational irrationality available—and there always are, because we can always formulate more extreme versions (perhaps unless one makes oneself structurally irrational, life as we know it will cease tomorrow and everyone who has ever lived will be tormented forever)—we are led into a situation where the value of having intentions in fact does not support having intentions but rather something much like them. As such,

55 For efficiency, see Bratman, *Intention, Plans, and Practical Reason*. For self-governance, see Bratman, *Planning, Time, and Self-Governance*. For moral participation, see Roughley, *Wanting and Intending*.

it seems better to have "shmintentions" that are not governed by the norms constitutive of intentions. That, too, is shmagency as underdetermination again.

5.3. *Desires*

The final single mental state I shall consider is desire. On one account, taking means to ends is instrumentally rational because it is constitutive of desire to do so.[56] Here, the relevant aspect of agency *A* is desires, the constitutive feature *C* is a norm of instrumental rationality, and structural rationality *S* in the form of instrumental rationality might be explained by how it is constitutive of desires. This yields a theory *T* of instrumental rationality. Admittedly, the norm that is explained here would have to differ slightly from Instrumental Irrationality, as it is formulated in terms of intentions rather than desires. But that is a minor tweak.

A bigger issue is that desires are modally escapable. We can easily imagine creatures without them. The Saturnians, for example, have besires instead. This gives rise to the same worries about extensional inadequacy as it does about single-belief or single-intention views. There is a space external to agency where someone asks shmagency questions about desires, but the norms constitutive of desires do not hold for them, for they lack desires; yet we want to include besiring creatures in our normative practices. This is because we can stipulate that besiring creatures are relevantly similar to us by being intelligent, knowledgeable, and so on. And the universality intuitions give us some reason to want to explain the force of Instrumental Irrationality for them, too.

But are desires really escapable? The Humean theory of motivation is a venerable account of the explanation of action, and it says that an action is an action in virtue of being caused and rationalized in the right way by belief-desire pairs. If desires are constitutive of action, perhaps they are inescapable in some important sense for all relevant entities. Maybe desires can do a surprising amount of work if we accept the Humean theory of motivation.

However, not even the Humean rationale works. This is so for two reasons. First, the Saturnians might well be shmagents who "shmact" rather than act. But the point of the modal escapability worry is to show how entities who do not count as agents according to some constitutivist views still have psychologies that should be subject to norms. Second, even supposing that desires are necessary for action and that the Saturnians act, *what kind* of desires feature in action still seems variable. Maybe actual agents have desires that are partially constituted by principles of rationality, but other possible creatures may have desires

56 For examples of defenses, see Goldman, *Reasons from Within*; Smith, "A Puzzle about Internal Reasons"; and Railton, "On the Hypothetical and Non-Hypothetical in Reasoning about Belief and Action."

that are phenomenal, consisting of experienced urges that are not constituted by principles of rationality. Perhaps these are the desires of Jupiterians. But as before, we want to explain the normativity of rationality for the Saturnians and the Jupiterians—they can be stipulated to be sophisticated enough to be part of normative practices, and we want universality in our explanations—so the problem remains regardless.

Are, then, desires that are constituted by principles of rationality valuable and therefore such that we can avoid the shmagency objection? I am not sure why they would be, but even if we could come up with some reason for thinking so, the by now familiar underdetermination worry would remain. It would do so regardless of why we would consider them to be valuable, for to what extent it is valuable to have desires governed by a norm of instrumental rationality rather than, for example, besires or desires that are phenomenal urges is a question that turns on the rational irrationality counterexamples we can construct. And we can always construct more.

It is, instead, time to summarize. So far, in section 4, I have criticized first-person authority views that take principles of rationality to be constitutive of standpoints. In section 5, I criticized single-mental-state views. I focused on beliefs, intentions, and desires, and argued that shmagency objections were problematic for them all. But what if we were to think of mental states as hanging together in *systems* that also are partially constituted by principles of rationality?

6. SYSTEMS-OF-MENTAL-STATES VIEWS

Another possibility is this: maybe mental states hang together in *systems*, and it is constitutive of these systems to be subject to norms of rationality. The core idea here is that mental states can be properly or improperly organized in systems of interrelated states, where these systems also are partially constituted by rational requirements.[57] What differentiates these from single-mental-state views is that the requirements are constitutive of systems that feature tokens of many types of mental states, such as intentions, beliefs, or desires, rather than of single tokens of the states.

There are many systems views. According to Bratman, there can be different kinds of agency, but a kind of self-governing cross-temporal agency is constituted in part by certain principles of rationality.[58] He even thinks that it is

57 Southwood discusses similar views using the terminology of functioning ("Vindicating the Normativity of Rationality"), but as not all views here need be functionalist, I opt for the broader language of systems.

58 Bratman, "Intention, Belief, Practical, Theoretical," "Intention, Practical Rationality, and Self-Governance," and *Planning, Time, and Self-Governance.*

constitutive of individual intentions to organize our actions together with our other intentions—though, presumably, the connections between mental states here also involve other states, such as beliefs, made explicit in a norm such as Instrumental Irrationality.

Some different versions of the systems-of-mental-states view have also been defended by Smith. Sometimes, Smith has indicated that an agent's entire psychology must hang together.[59] On one interpretation of this idea, it is natural to think that principles of rationality are partially constitutive of an agent's psychology, at least when the agent is functioning perfectly. Smith has also sometimes hinted at a weaker view.[60] Here, he argues that Humean agents need to combine beliefs and desires using modally sensitive rational capacities to act. The idea is then that the belief-desire pairs that generate actions on the Humean theory of motivation in fact should be thought of as belief-desire-rationality triples. This would make action in part constituted by rationality and make a belief-desire psychology a kind of system that is regulated by requirements of rationality.

Other versions of this view have recently been developed in the context of the burgeoning post-Broome literature on structural rationality. Alex Worsnip, for example, explains the normativity of rationality in a slightly different way from most constitutivists but nevertheless wants to locate principles of rationality in people's psychologies to give an account of their ontology.[61] And John Brunero attempts to explain the force of at least some norms of rationality in this way. He calls this view "non-normative disjunctivism." This view "looks to the logical relations among the contents of your attitudes, and the constitutive aims of those attitudes, to explain why something is amiss in [the case of irrationality]."[62] Taking beliefs to aim at truth and intentions to aim at effective controlled action, Brunero argues:

> If your belief that you must intend to buy a ticket in order to take the train to Charleston achieves its constitutive aim (truth), then you'll take the train to Charleston only if you intend to buy a ticket. But since you don't intend to buy a ticket, you won't take the train to Charleston. If this is so, your intention won't succeed with respect to its constitutive aim. So, given your combination of attitudes, either you're wrong about the need to intend to buy the ticket, and so your belief fails to achieve its constitutive aim, or you won't take the train to Charleston, and so your intention will fail to achieve its constitutive aim. In other words, your

59 Smith, *The Moral Problem* and "The Magic of Constitutivism."

60 Smith, "The Explanatory Role of Being Rational."

61 Worsnip, "What Is (In)coherence?" and *Fitting Things Together.*

62 Brunero, *Instrumental Rationality*, 197.

> failing to intend to buy a ticket has ensured a constitutive aim failure: either your belief is false or your intention will not succeed.[63]

Here, beliefs and intentions in combination do in some cases ensure the failure of at least one of the attitudes. Hence the disjunctivism. But whether we are concerned with Bratman's, Smith's, Worsnip's, or Brunero's views, systems of mental states count as constitutivist. The systems are aspects of—or even all of—our agency *A*, and the norms of rationality *S* themselves are partially constitutive of them *C*. These norms are therefore supposed to apply to us, and, in combination with further premises, are often taken to have normative force.

As Southwood has noted, however, this type of view seems *prima facie* obviously susceptible to the shmagency objection.[64] Why be the type of agent Bratman, Smith, Worsnip, Brunero, or others posit rather than some shmagent? Good question. Modal escapability does in fact seem particularly pertinent here. Neither the shmandpoint Mercurians, the belief-desire-pair Martians, the besire Saturnians, the shmelief Neptunians, nor the desires-as-urges Jupiterians are committed to these psychological systems, so there are standpoints external to agency from which they may ask shmagency questions. And as usual, we can stipulate that they are sophisticated enough to be included in our normative practices and that we want to explain the universal force of rational requirements in a way that includes them. Hence, modal escapability is a problem here too.[65]

It seems much more feasible to defend systems views using other values. That is, in fact, what both Bratman and Smith have done. For Bratman, planning agency involves efficiency and self-governance, and we may also want to add Neil Roughley's moral participation point to the picture. And for Smith, agency is a goodness-fixing kind.[66] To be good *qua* agent is to be fully coherent, so being fully good *qua* agent and to be fully coherent amount to the same thing.

63 Brunero, *Instrumental Rationality*, 196–97.

64 Southwood, "Vindicating the Normativity of Rationality."

65 As in section 3.2 above, it might perhaps be thought that as writers such as Bratman and Brunero talk about instrumental rationality in the context of intentions rather than beliefs and desires, some of these worries need not be issues for them. They might only be concerned with creatures with systems of mental states that feature intentions in their sense of the word. However, this worry also received a reply above. We should explain the normativity of norms such as Instrumental Irrationality for creatures such as belief-desire-pair Martians and not just for those with a certain type of intentions as long as they take means to ends, or else our theory will be problematically disunified: it will lack parsimony and unity, be ad hoc and unconservative, and lack explanatory power.

66 Again: for efficiency, see Bratman, *Intention, Plans, and Practical Reason*. For self-governance, see Bratman, *Planning, Time, and Self-Governance*. For moral participation, see Roughley, *Wanting and Intending*. And for goodness-fixing agency, see Smith, "The Magic of Constitutivism."

May we then defend the normativity of rationality by appealing to the value of agency on systems-of-mental-states accounts? Unfortunately, the same issues as before remain—and for the same reasons. It is unclear how or when systems are valuable, so it is quite unclear to what extent these values support having systems rather than "shmystems" of mental states: shmystems that are very much like systems of mental states, but are governed by norms of rationality rather than other, similar, norms. Hence, underdetermination remains a worry.

Or does it? In section 5.2 above, I mentioned we could add up the reasons of efficiency, self-governance, and moral participation to strengthen the case for Bratman's famous account of intentions. While that argument did not work, we may perhaps want to add further values still to them, such as the value of being fully good *qua* agent. Is *that* enough to say that agency rather than shmagency is supported by the value of agency?

I doubt it. We can always formulate new cases of rational irrationality, whether they are cases of burglars, demons who will punish us for eternity, or cases where, unless one is rationally irrational, life as we know it will cease tomorrow and everyone who has ever lived will be tormented forever. As long as there are such possibilities, it seems better to be such that one is constituted by a norm that is similar to a norm of rationality but allows exceptions to accommodate the cases. Shmagency rather than agency seems justified by value again.

A more promising reply is to say that agential goodness is a *kind* of value that is different from others. The thought here is that the full goodness of an agent *qua* agent does not admit of exceptions in the way that ordinary values or reasons might do. Those values and reasons may be weighed against other considerations, but perhaps the value of agency is not a type of value that may be. Then agential goodness seems more universally applicable: if good agency is coherent agency, then what is best for an agent *qua* agent is to be fully coherent, no exceptions allowed. Might that justify a systems-of-mental-states view?

The problem here is that the view seems normatively inadequate. Again, we can always formulate extreme cases of rational irrationality. Is it truly better for an agent even *qua* agent to be fully coherent and rational than to avoid cases where all life ceases to be and everyone, including that agent, gets tormented forever? I doubt that. Even if consequences for other agents do not matter at all to the agent, they will be tormented forever themselves. That does not seem to be a good way for an agent to be an agent.

One final possibility. Maybe systems-of-mental-states views are different from single-mental-state views in another way. While failing to live up to some single-mental-state norm *ipso facto* could entail that one does not have the mental state constituted by being disposed to follow that norm, when we are concerned with systems of mental states, one can sometimes fail to live up to

the norms that constitute some mental states but still have *systems* of mental states that feature them. If so, maybe cases of rational irrationality can be mitigated. Maybe it is sometimes more valuable to be irrational with respect to some intention or belief token one has, but one can still maintain a system of mental states guided by norms of rationality on aggregate.

However, it is easy enough to reformulate the cases to generate underdetermination concerns again. Perhaps life as we know it will cease tomorrow and everyone who has ever lived will be tormented forever if one has a *system* of mental states governed by certain rational norms that are similar to but other than Instrumental Irrationality, such as Instrumental Irrationality-Except-at-2:00 AM-on-Tuesdays. So underdetermination remains worrisome.

7. CONCLUSION

Recap time. I introduced constitutivism about structural rationality in section 1. In section 2, I presented some of its attractions. In section 3, I introduced the shmagency objection and defended two recent versions: modal escapability and underdetermination.

Then, I applied these objections to constitutivism about the normativity of structural rationality. In section 4, I argued that the shmagency worries remain significant for first-person-privilege views. In section 5, I argued the same thing for single-mental-state views. In section 6, I showed that the worries apply to systems-of-mental-states views.

Is there a general way to avoid the worries? I am skeptical as long as all the original motivations for constitutivism about rationality are adhered to. However, the assumption that sophisticated shmagents ought to be incorporated into our normative practices and the intuition of universality, which entails that norms of rationality are supposed to be normative for a very extensive range of possible agents and shmagents, might maybe be done away with. If so, maybe norms of rationality can be binding for *some* agents (or sophisticated shmagents) even if they are not for all agents (or sophisticated shmagents). With a suitable take on which norms these are, perhaps the norms will be binding for at least the vast majority of human beings. Such a view may still be defensible. Whether it is, however, will have to be a topic for elsewhere.[67]

University of Vienna
olof.leffler@univie.ac.at

67 This paper develops and extends some arguments in my "New Shmagency Worries" and *The Constitution of Constitutivism* (esp. chs. 1 and 3 and app. B) by running them as problems for constitutivism about structural rationality rather than moral reasons. I am grateful to

REFERENCES

Bealer, George. "Intuition and the Autonomy of Philosophy." In *Rethinking Intuition: The Psychology of Intuition and Its Role in Philosophical Inquiry*, edited by Michael DePaul and William Ramsey, 201–39. Lanham, MA: Rowman and Littlefield, 1998.

Bratman, Michael. "Intention, Belief, and Instrumental Rationality." In *Planning, Time, and Self-Governance*, 52–75.

———. "Intention, Belief, Practical, Theoretical." In *Planning, Time, and Self-Governance*, 18–51.

———. *Intention, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press, 1987.

———. "Intention, Practical Rationality, and Self-Governance." In *Planning, Time, and Self-Governance*, 76–109.

———. *Planning, Time, and Self-Governance: Essays in Practical Rationality*. Oxford: Oxford University Press, 2018.

Broome, John. *Rationality through Reasoning*. Oxford: Oxford University Press, 2013.

———. "Replies to Southwood, Kearns and Star, and Cullity." *Ethics* 119, no. 1 (October 2008): 96–108.

Brunero, John. *Instrumental Rationality: The Normativity of Means-Ends Coherence*. Oxford: Oxford University Press, 2020.

Coons, Christian, and David Faraci. "First-Personal Authority and the Normativity of Rationality." *Philosophia* 38, no. 4 (December 2010): 733–40.

Cowie, Christopher. "In Defence of Instrumentalism about Epistemic Normativity." *Synthese* 191, no. 6 (April 2014): 4003–17.

Enoch, David. "Agency, Shmagency: Why Normativity Won't Come from What Is Constitutive of Action." *Philosophical Review* 115, no. 2 (April 2006): 169–98.

———. "Shmagency Revisited." In *New Waves in Metaethics*, edited by Michael S. Brady, 208–33. New York: Palgrave MacMillan, 2011.

Ferrero, Luca. "Constitutivism and the Shmagency Challenge." In *Oxford*

everyone who provided input on those pieces way back when. For feedback on this manuscript, I want to thank Janis Schaab and Michael Bratman for correspondence, an audience at EIDOS at Durham—especially Susan Notess and Sarah Uckelman—and two anonymous reviewers for the *Journal of Ethics and Social Philosophy*. Research in this paper has been conducted within the European Research Council (ERC) project "The Normative and Moral Foundations of Group Agency" at the University of Vienna, which in turn was funded by the ERC under the European Union's Horizon 2020 research and innovation program (grant agreement no. 740922). Much like I thanked the EU for the Erasmus+ scheme that funded part of the research behind those older pieces, I want to thank the EU for this too.

Studies in Metaethics, vol. 4, edited by Russ Shafer-Landau, 303–32. Oxford: Oxford University Press, 2009.
———. "Inescapability Revisited." *Manuscrito* 41, no. 4 (2018): 113–58
———. "The Simple Constitutivist Move." *Philosophical Explorations* 22, no. 2 (April 2019): 146–62.
Foot, Philippa. "Morality as a System of Hypothetical Imperatives." *Philosophical Review* 81, no. 3 (July 1972): 305–16.
Gendler, Tamar Szabó. "Alief and Belief." *Journal of Philosophy* 105, no. 10 (October 2008): 634–63.
Goldman, Alan H. *Reasons from Within: Desires and Values*. Oxford: Oxford University Press, 2011.
Harman, Gilbert. *Change in View: Principles of Reasoning*. Cambridge, MA: MIT Press, 1986.
———. "Practical Reasoning." *Review of Metaphysics* 29, no. 3 (March 1976): 431–63.
Henning, Tim. *From A Rational Point of View: How We Represent Subjective Perspectives in Practical Discourse*. Oxford: Oxford University Press, 2018.
Huemer, Michael. "Compassionate Phenomenal Conservatism." *Philosophy and Phenomenological Research* 74, no. 1 (January 2007): 30–55.
Katsafanas, Paul. *Agency and the Foundations of Ethics*. Oxford: Oxford University Press, 2013.
———. "Constitutivism about Practical Reasons." In *The Oxford Handbook of Reasons and Normativity*, edited by Daniel Star, 367–91. Oxford: Oxford University Press, 2018.
Kiesewetter, Benjamin. *The Normativity of Rationality*. Oxford: Oxford University Press, 2017.
Kolodny, Niko. "Why Be Rational?" *Mind* 114, no. 455 (July 2005): 509–63.
Kornblith, Hilary. *Knowledge and Its Place in Nature*. Oxford: Oxford University Press, 2002.
Korsgaard, Christine M. *Self-Constitution: Agency, Identity, and Integrity*. Oxford: Oxford University Press, 2009.
———. *The Sources of Normativity*. Cambridge: Cambridge University Press, 1996.
Leffler, Olof. "The Constitution of Constitutivism." PhD diss., University of Leeds, 2019.
———. "New Shmagency Worries." *Journal of Ethics and Social Philosophy* 15, no. 2 (June 2019): 121–45.
Levy, Yair. "Does the Normative Question about Rationality Rest on a Mistake?" *Synthese* 195, no. 5 (May 2018): 2021–38.
Lord, Errol. *The Importance of Being Rational*. Oxford: Oxford University Press,

2018.
Parfit, Derek. *Reasons and Persons*. Oxford: Oxford University Press, 1984.
Railton, Peter. "On the Hypothetical and Non-Hypothetical in Reasoning about Belief and Action." In *Ethics and Practical Reason*, edited by Garrett Cullity and Berys Gaut., 53–80. Oxford: Oxford University Press, 1997.
Raz, Joseph. "The Myth of Instrumental Rationality." *Journal of Ethics and Social Philosophy* 1, no. 1 (April 2005): 1–28.
Roughley, Neil. *Wanting and Intending: Elements of a Philosophy of Practical Mind*. Dordrecht: Springer, 2015.
Setiya, Kieran. "Cognitivism about Instrumental Reason." *Ethics* 117, no. 4 (July 2007): 649–73.
Smith, Michael. "Agents and Patients, or: What We Learn about Reasons for Action by Reflecting on Our Choices in Process-of-Thought Cases." *Proceedings of the Aristotelian Society* 112, no. 3 (October 2012): 309–31.
———. "Constitutivism." In *The Routledge Handbook of Metaethics*, edited by Tristram McPherson and David Plunkett, 371–84. New York: Routledge, 2017.
———. "The Explanatory Role of Being Rational." In *Reasons for Action*, edited by David Sobel and Steven Wall, 58–80. New York: Cambridge University Press, 2009.
———. "Four Objections to the Standard Story of Action (and Four Replies)." *Philosophical Issues* 22, no. 1 (October 2012): 387–401.
———. "The Magic of Constitutivism." *American Philosophical Quarterly* 52, no. 2 (April 2015): 187–200.
———. *The Moral Problem*. Oxford: Blackwell, 1994.
———. "A Puzzle about Internal Reasons." In *Luck, Value and Commitment: Themes from the Philosophy of Bernard Williams*, edited by Ulrike Heuer and Gerald Lang, 195–218. Oxford: Oxford University Press, 2012.
Southwood, Nicholas. "Constructivism and the Normativity of Practical Reason." In *The Many Moral Rationalisms*, edited by Karen Jones and François Schroeter, 91–109. Oxford: Oxford University Press, 2018.
———. "Vindicating the Normativity of Rationality." *Ethics* 119, no. 1 (October 2008): 9–30.
Tiffany, Evan. "Why Be an Agent?" *Australasian Journal of Philosophy* 90, no. 2 (2011): 223–33.
Tubert, Ariela. "Constitutive Arguments." *Philosophy Compass* 5, no. 8 (May 2010): 656–66.
Velleman, J. David. *How We Get Along*. Oxford: Oxford University Press, 2009.
———. "On the Aim of Belief." In *The Possibility of Practical Reason*, 244–81. Oxford: Oxford University Press, 2000.

———. *The Possibility of Practical Reason*. Oxford: Oxford University Press, 2000.

———. *Practical Reflection*. 2nd ed. Stanford, CA: CSLI Publications, 2007.

———. "What Good Is a Will?" In *Action in Context*, edited by Anton Leist, 193–215. Berlin: De Gruyter, 2007.

Velleman, J. David, and Nishi Shah. "Doxastic Deliberation." *Philosophical Review* 114, no. 4 (October 2005): 497–534.

Walden, Kenneth. "Laws of Nature, Laws of Freedom, and the Social Construction of Normativity." In *Oxford Studies in Metaethics*, vol. 7, edited by Russ Shafer-Landau, 37–79. Oxford: Oxford University Press, 2012.

Wallace, R. Jay. "Normativity, Commitment, and Instrumental Reason." *Philosophers' Imprint* 1, no. 3 (December 2001): 1–26.

Way, Jonathan. "Reasons and Rationality." In *The Oxford Handbook of Reasons and Normativity*, edited by Daniel Star, 485–503. Oxford: Oxford University Press, 2018.

Wedgwood, Ralph. *The Value of Rationality*. Oxford: Oxford University Press, 2017.

Williams, Bernard, "Deciding to Believe." In *Problems of the Self*, 136–51. Cambridge: Cambridge University Press, 1973.

Worsnip, Alex. *Fitting Things Together: Coherence and the Demands of Structural Rationality*. Oxford: Oxford University Press, 2021.

———. "What Is (In)coherence?" In *Oxford Studies in Metaethics*, vol. 13, edited by Russ Shafer-Landau, 184–206. Oxford: Oxford University Press, 2018.

Ylikoski, Petri, and Jaakko Kuorikoski. "Dissecting Explanatory Power." *Philosophical Studies* 148, no. 2 (March 2010): 201–19.

JOURNAL *of* ETHICS & SOCIAL PHILOSOPHY
http://www.jesp.org
ISSN 1559-3061

The *Journal of Ethics and Social Philosophy* (JESP) is a peer-reviewed online journal in moral, social, political, and legal philosophy. The journal is founded on the principle of publisher-funded open access. There are no publication fees for authors, and public access to articles is free of charge. Articles are typically published under the CREATIVE COMMONS ATTRIBUTION-NONCOMMERCIAL-NODERIVATIVES 4.0 license, though authors can request a different Creative Commons license if one is required for funding purposes.

Funding for the journal has been made possible through the generous commitment of the Division of Arts and Humanities at New York University Abu Dhabi.

جامعة نيويورك أبوظبي
NYU ABU DHABI