JOURNAL *of* ETHICS
& SOCIAL PHILOSOPHY

VOLUME XXII · NUMBER 2
July 2022

ARTICLES

SPECIAL CONTRIBUTION

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 Gould School of Law and the Dornsife College of Letters, Arts, and Sciences at the University of Southern California.

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 associate 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. 22, NO. 2 · JULY 2022
https://doi.org/10.26556/jesp.v22i2.1628

DEEPFAKES, DEEP HARMS

Regina Rini and Leah Cohen

DEEPFAKES are digitally altered audio or video recordings in which one person's face and/or voice are mapped onto the body of another person, creating misleading evidence of events that never took place. Deepfake videos can be created with open-source software based on machine-learning algorithms. Currently this technology is a niche interest, mostly isolated to internet pornography communities, but its use by other actors—some with still more malicious intent—is a very near technological possibility. The aim of this paper is to get out in front of that future and recognize some of the emergent harms the technology may bring about.

We think that the arrival of cheap and easy-to-use deepfake technology will have a number of significantly harmful effects, at both personal and social levels. For simplicity, this paper will focus on the *personal* harms of deepfakes.[1] Put simply: What happens *to* a person who has been deepfaked? What sort of harm is done by being falsely represented in a deceptive recording?

The structure of the paper is as follows. First, we provide a bit more background on the history and technology behind deepfakes. Then we marshal a parade of horribles: several distinct ways in which deepfakes may harm their targets. These include a new form of objectifying harm that we call *virtual domination*, in which a person's autonomy is invaded by their being represented as engaging in unconsented (and fabricated) sexual encounters; *illocutionary harm*, where a person is forced to engage in involuntary speech acts in order to dispute the content of deepfakes; and, most speculatively, existential trauma caused by *panoptic gaslighting*, when a person's memory and identity are undermined by a myriad of systemically targeted fabrications.

1. DISCOUNT DIGITAL DECEPTION

The deepfake technology available to consumers superficially resembles Hol-

1 For discussion of the social and political consequences, see Rini, "Deepfakes and the Epistemic Backstop."

lywood special effects, the sort where dead actors are revived anew onscreen. But these professional techniques are extremely expensive and time consuming. What makes deepfakes remarkable is that they provide an approximation of the same effect, much more cheaply and quickly. The knowledge, resources, and time needed to create deepfakes are substantially lower than for other kinds of video manipulation, and can be done as anonymously as the internet will allow.[2]

Another difference, of course, is that Hollywood stars usually *consent* to being digitally doppelgängered. That is rarely true of the targets of deepfakes. Pornography, always a central causal factor on the internet, drove the early use of deepfakes, including the origin of the term itself. In autumn 2017, an anonymous user going by the handle "deepfakes" posted multiple pornographic videos to the website Reddit. These videos featured the faces of famous actresses mapped onto the bodies of pornographic performers engaging in explicit sexual acts. By February 2018, Reddit and other sites—even Pornhub—had banned the videos.[3] But policing the internet is an unending task and the videos abound. A recent study by the digital security firm DeepTrace found that 96 percent of online deepfakes are pornographic.[4]

By late 2018 public attention began to focus on the potential political risks of deepfakes. In May of that year, the Flemish Socialist Party released a deepfake of Donald Trump appearing to urge Belgium to withdraw from the Paris Climate Accords. The video is noticeably unreal. The mouth of the speaker is out of sync with the rest of Trump's facial expressions. If that were not enough, the final line of the video clearly states, "We all know climate change is fake, just like this video." Yet still, according to *Politico*, "some commenters on the party's Facebook page had apparently not realized the video was a fake."[5]

By mid-2019, government and corporate policymakers had begun debating solutions. In June, the US House Intelligence Committee held hearings on the risks of deepfakes, while the state of Virginia banned the use of deepfakes in "revenge porn."[6] In January 2020, Facebook announced that it would ban deepfakes

2 Journalist Samantha Cole has done extensive work investigating deepfake technology and the internet communities that favor it. See Cole, "AI-Assisted Fake Porn Is Here and We're All Fucked."

3 Kelion, "Reddit Bans Deepfake Porn Videos."

4 Simonite, "Most Deepfakes Are Porn, and They're Multiplying Fast."

5 Von der Burchard, "Belgian Socialist Party Circulates 'Deep Fake' Donald Trump Video."

6 Cox, "Deepfake Revenge Porn Distribution Now a Crime in Virginia." See https://www.youtube.com/watch?v=tdLs9MlIWOk for the House Intelligence meeting. Rep. Yvette Jones introduced a bill targeting deepfakes at the US federal level, though (as of writing of this paper in late 2020) it has not been acted upon. See https://www.congress.gov/bill/116th-congress/house-bill/3230.

from its platform, though it granted a somewhat amorphous exception for "parody and satire."[7] Experts continue to debate whether technical or legal solutions are feasible.[8]

We will assume that a solution is not immediately forthcoming. Our goal instead is to illustrate why a solution is urgently *needed*. What are the harms that deepfakes might cause if left unchecked? Some are already recognized in journalism and legal scholarship. Law professors Bobby Chesney and Danielle Citron, for example, enumerate risks of election interference, corporate malfeasance, psychological espionage, and personal blackmail.[9] We agree that these are serious concerns, but we think that the philosopher's lens may help us see more subtle dangers. We turn now to these.

2. FRANKENPORN AND VIRTUAL DOMINATION

Deepfakes offer their creators a disturbing form of power over other people, one that seems inevitably to lend itself to pornographic misuse. The first face-swapped videos on Reddit featured actresses such as *Wonder Woman* star Gal Gadot engaged in simulated incest. The actresses were not consulted and did not consent to having their images used in such a manner, nor did the pornographic performers. This fact is what eventually led Reddit to shut down its deepfake forum, under its policy against "involuntary pornography." Yet the videos still circulate voluminously in dingier corners of the internet.

Deepfake technology crashes into long-running debates about pornography and the objectification of women. In the 1970s and '80s, feminist critics like Andrea Dworkin and Catharine MacKinnon argued that porn functions to affirm perceptions of women as playthings for male viewers, mere objects for gratification rather than full persons with autonomous wills. Other theorists have pointed to the ways in which pornography silences the viewpoints of women. These positions are controversial though, even within feminist communities. Some feminists argue that pornography, when executed carefully and respectfully, can be compatible with or even empowering of women's liberation. Such "porn-positive" feminists emphasize the agency of individual performers—not

7 Shead, "Facebook to Ban 'Deepfakes.'"

8 For discussion of proposed solutions (and their shortcomings), see Farid, "Digital Forensics in a Post-Truth Age"; Chesney and Citron, "Deep Fakes"; Harris, "Deepfakes"; Li and Lyu, "Exposing DeepFake Videos by Detecting Face Warping Artifacts"; Rini, "Deepfakes and the Epistemic Backstop."

9 Chesney and Citron, "Deep Fakes."

to mention female directors and distributors—in designing pornography that expresses women's sexuality without shame.[10]

This is an unresolved debate, one with a good deal of subtlety to it. Even if Dworkin and others are correct that the *general* social function of pornography is to objectify women, it still might be true that the *local* function of some feminist-created pornography is empowerment and destigmatization. Women in general do not consent to how they are represented in pornography, even if some women consent to their personal presentation in specific pornographic works. Weighing these two points is extremely difficult. But deepfakes obliterate any subtlety or nuance because *no one* consents to deepfake porn.

Journalist Samantha Cole interviewed women who have worked as pornographic performers to get their views about the emergence of deepfake porn. Retired performer Alia James told Cole: "It's really disturbing.... It kind of shows how some men basically only see women as objects that they can manipulate and be forced to do anything they want.... It just shows a complete lack of respect for the porn performers in the movie, and also the female actresses."[11]

There is something painfully literal in the sort of objectification at work in deepfake porn. Reddit forum users requested the creation of custom videos, with particular actresses swapped into particular sex acts, as casually as specifying the paint job at a car dealership or ordering toppings on a pizza. And as the technology improves, the ability to treat women's images as playthings will only grow. One emerging technique, developed by computer scientists without bad intentions, uses artificial intelligence to simulate the movements of a real person's entire body by mapping it onto an actor's poses.[12] Once a similar technique is available to deepfakers, they will no longer be limited to superimposing famous faces onto existing porn clips. Instead they will generate novel simulacra of their targeted celebrities—poseable, pliable representations ordered to do whatever the user desires.

For now, deepfakes are limited to what Cole calls "frankenporn," with the digitally manipulated face of one woman stitched onto the body of another. Once again, this seems to be an unsubtle manifestation of the worst sort of objectification that feminist critics have always charged to pornography. In deep-

10 This is an extremely large debate. For important contributions, see Dworkin, "Against the Male Flood"; MacKinnon, *Feminism Unmodified*; Nussbaum, "Objectification"; Strossen, *Defending Pornography*; Langton and Hornsby, "Free Speech and Illocution"; Maitra, "Silencing Speech"; and Bauer, *How to Do Things with Pornography*. For a recent overview and reorientation, see Cawston, "The Feminist Case against Pornography."

11 Cole, "AI-Assisted Fake Porn Is Here and We're All Fucked."

12 Liu et al., "Neural Rendering and Reenactment of Human Actor Videos."

fake frankenporn, women really are reduced to body parts: a face from here, a torso from there, interchangeable and commodified. What is absent is any sort of independent mind or will.

In an important sense, the entity depicted in frankenporn *cannot* have a determinate will, since it is a composite of the parts of two different people, unified only in digital artifice. This entity is a mereological sum, constituted from the body parts of multiple humans. Its apparent intentions belong to neither of the women whose body parts appear. Instead, it seems to depict a "derived intentionality," like a fictional character, specified by the deepfake's creator.[13]

Yet it does seem to matter to deepfake-porn consumers that they are viewing the faces of *particular* women. Their requests target specific celebrities, or in some cases their own ex-girlfriends or acquaintances. This apparent need—to externalize a fantasy of some specific woman doing whatever the user demands—supports the familiar feminist claim that, for at least some men, sexual domination of women is as much about *power* as it is about physical gratification. There certainly appears to be an anti-feminist politics in the communities where deepfake porn is traded. As Cole puts it:

> In these online spaces, men's sense of entitlement over women's bodies tends to go entirely unchecked. Users feed off one another to create a sense that they are the kings of the universe, that they answer to no one. This logic is how you get incels and pickup artists, and it's how you get deepfakes: a group of men who see no harm in treating women as mere images, and view making and spreading algorithmically weaponized revenge porn as a hobby as innocent and timeless as trading baseball cards.[14]

Deepfaked frankenporn, then, is *virtual domination*, an extreme expression of sexual objectification aimed against specific women. As Dworkin puts it, "Objectification occurs when a human being, through social means, is made less than human, turned into a thing or commodity, bought and sold. When objectification occurs, a person is depersonalized, so that no individuality or integrity is available socially."[15] Frankenporn turns real people into digital toys. Even those unpersuaded by feminist objections to traditional pornography ought to recognize the moral wrong here.

13 "Derived intentionality" comes from Searle, *The Rediscovery of the Mind*.

14 Cole, "Deepfakes Were Created as a Way to Own Women's Bodies."

15 Dworkin, "Against the Male Flood," 15.

2. ILLOCUTIONARY HARM

We turn now to another potential harm of deepfakes, one more closely tied to their epistemic effects on public discourse. Deepfakes do not have to trick anyone in order to be harmful. Even if a deepfake is ultimately debunked, or never believed at all, it can still hurt the person it falsely depicts by changing the discursive context around them. This point is most clear for public figures, like politicians or celebrities. When deepfakes illegitimately force a public figure to react with undesired speech acts, they cause what we will refer to as *illocutionary harm.* In this section we will explicate what makes this a distinctive sort of harm, then catalog several forms it might take.[16]

There is a legend about the American politician Lyndon Baines Johnson (LBJ). Facing loss in a Texas congressional race, Johnson instructed an aide to spread rumors that his opponent engaged in sex with pigs. "We can't get away with calling him a pig-fucker," said the campaign manager. "No one's going to believe a thing like that." Johnson replied: "I know. But let's make the son-of-a-bitch deny it."[17]

Imagine a 2020s version of the LBJ legend. Now the porcine indecency is no longer mere rumor; instead it has been deepfaked, with the opponent's head digitally inserted into a video of the alleged act. (Best to imagine this case only schematically.) The video quickly goes viral online. On cable news, experts debate the video's veracity while blurred-for-TV excerpts play in the background. Late-night comics quickly join in. Most people realize that it is probably fake, but they still laugh along. At first, the politician tries to simply ignore the video, but soon it is everywhere. It becomes hard to do any interview, as even respectable journalists start asking thinly coded questions. Opposing party operatives turn up at rallies dressed in pig costumes. Finally, the politician's aides say: this is only going to stop if you address it directly, once and for all. The press conference is called, the podium prepared. And so there, on live TV, is the son-of-a-bitch denying that he had sex with a pig.

This is bad. All else equal, people who aspire to public office should not have

16 As this paper went to final editing, we became aware of a very recent paper by Henry Schiller exploring the same term ("Illocutionary Harm"). Our use of the term is not the same as Schiller's, though there is some interesting overlap.

17 That is the frequently told legend, anyway. It is almost certainly not true. The most plausible source we have found is Joseph Califano, an LBJ aide in the 1960s. In Califano's version, LBJ was actually the protesting young staffer in this story. It was LBJ's mentor, Richard Kleberg, who played the "let him deny it" card against an opponent. Also, the barnyard consort was a sheep, not a pig. See Califano, *The Triumph and Tragedy of Lyndon Johnson*, 118. We have kept the legend in our main text since it is what frequently appears in political journalism.

to call press conferences to deny false allegations of unnatural congress with livestock. In fact, we think that a person placed in this position has been harmed, *even if their denial is effective*. That is, even in the unlikely event that everyone immediately accepts the denial and ceases to believe that the video is veridical, the denier has still been harmed simply by having to issue the denial.

The key idea of illocutionary harm is this: a person can be harmed by being illegitimately compelled to perform an undesired speech act. Setting deepfakes aside for the moment, think of simpler examples. Totalitarian regimes frequently force their citizens to engage in compelled speech. Under Mao, the Chinese Communist Party ordered comrades to write "self-criticism statements," confessions of their complicity in capitalist villainy. Many were also pressured to falsely testify against friends and family. Similar things happened in the Soviet Union and Nazi Germany. These are all examples of illocutionary harm. Importantly—and perhaps controversially—we hold that this is a distinctive *type* of harm, in that it does not wholly reduce to other types of harm or wronging. Against our view, one might insist that illocutionary harm reduces to some combination of psychological anguish or reputational effects. But we think this misses a key feature of compelled speech. When a person is illegitimately compelled to speak, they are abused *specifically in their capacity as a speaker*.

What does that mean? We have in mind here something akin to Miranda Fricker's account of testimonial injustice. According to Fricker, when a person's testimony is unfairly dismissed on the basis of their membership in a derogated social category, that person has been "wronged in one's capacity as a knower." In addition to whatever material harms might result from not being believed, the target is undermined "in a capacity essential to human value" and so "suffers a great injustice."[18]

Similarly, we think that illegitimately compelled speech involves a distinctive type of harm, a harm to one's capacity as a user of information. In this we follow Rachel McKinney's work on "extracted speech."[19] McKinney's primary examples concern coercion, such as when psychological pressure drives innocent people to confess to crimes. Such pressure "amounts to wrongly undermining, bypassing, or overriding an agent's ability to speak voluntarily," and wrongs victims "as communicative agents."[20]

McKinney distinguishes two ways that extracted speech can be wrongful. First, in a forward-looking way, it can *license* future wrongs against victims (as when an extracted confession licenses the unjust conviction of an innocent

18 Fricker, *Epistemic Injustice*, 44.

19 McKinney, "Extracted Speech."

20 McKinney, "Extracted Speech," 266, 259.

defendant). Second, the mere act of extracting involuntary speech can *itself* be wrongful (regardless of further consequences) when it comes about through subverting a person's communicative agency.

We think that deepfakes pose similar risks. To start off simply, take McKinney's first form of wrongfulness: licensing future wrongs against victims. Suppose a deepfake succeeds in tricking some part of an audience into believing that the target said words they never actually used. This will often license illegitimate treatment.

This sort of harm has already been caused by much simpler manipulations than deepfakes. In 2016, the then-governor of Jakarta, Basuki Tjahaja Purnama, widely known as Ahok, gave a public address decrying his opponents' partisan use of religion. An edited video soon appeared online, in which a word had been clipped from Ahok's remark, causing it to sound as if he were criticizing the Koran itself and not his opponents' appropriation thereof. An enormous public outcry followed, resulting in Ahok losing his governorship and being imprisoned on charges of blasphemy.[21]

Similarly, in 2015, the American anti-abortion pressure group Center for Medical Progress released surreptitious recordings of a "sting" meeting with a representative of Planned Parenthood, maliciously edited to make it appear that the latter admitted to profiting from the sale of fetal body parts. As a result, several state governments cut Medicare funding to Planned Parenthood. Then-candidate Donald Trump cited the recordings as grounds for ending federal funding, an ambition he fulfilled in 2019.[22]

In both cases, not everyone believed that the edited videotapes were veridical. But to those who did, the videotapes appeared to license punishment that was in fact unjust. Ahok and Planned Parenthood both suffered wrongs by being portrayed as saying something other than what they actually said.[23]

Yet even when the faked video is widely disbelieved, deepfakes could still impose illocutionary harm, along the lines of McKinney's second type. As we have already stressed, being forced to publicly *deny* an embarrassing rumor can itself be harmful, partly for the reasons McKinney identifies: it subverts the vic-

21 Soeriaatmadja, "Man Who Uploaded Controversial Video of Ex-Jakarta Governor Ahok Sentenced to Jail." The person who edited the video, a university lecturer, was *also* sent to prison, separately, on hate-crime charges.

22 Kliff, "I Watched 12 Hours of the Planned Parenthood Sting Videos"; Diamond, "Trump"; Armstrong, "Planned Parenthood Cut Off from Federal Funding Under Trump Rule."

23 Technically this is probably not "extracted speech" in McKinney's sense, since in these cases the depicted speech act never even happened (at least not as portrayed). One might call this phenomenon "imposter speech."

tim's communicative agency. There are many understandable reasons that one may not wish to publicly speak on a topic, such as tact, embarrassment, privacy, and safety. Being compelled to do so by a fabricated recording unjustly compels speech.[24]

Illocutionary harm may happen even when the fabricated speech is truthful and consistent with the (apparent) speaker's beliefs. A deepfake might make a public figure appear to say something that they *do* believe, yet for whatever reason did not wish to express publicly. In other words, the *occurrence of the speech act* might be faked, though the content it expresses might be truthful. The target may then be forced to publicly address the circulating fake, either falsely denying they believe what had been attributed to them, or openly admitting what they would rather have left unspoken.

An obvious example of this kind of situation happens when a public figure is forced out of the closet. In 2001, a tabloid published (genuine) photos of the Australian American actress Portia de Rossi with her then girlfriend. De Rossi did not publicly identify as a lesbian and struggled with how to respond. She later told *The Advocate:* "The most important thing for me was to never, ever, ever deny it. But I didn't really have the courage to talk about it."[25] For the next several years de Rossi avoided talking to reporters about the topic, until she officially came out when she began dating her future wife, Ellen DeGeneres.

In this case, de Rossi's sexuality became a sort of open secret. Hollywood people knew about the tabloid images, of course, but so long as she did not address them, respectable publications avoided bringing it up. Imagine, however, that deepfake technology had been available in 2001. Imagine someone made a deepfake video seeming to depict de Rossi saying to the camera: "I am a proud lesbian and I want the world to know." In that case, she could not have simply ignored it and counted on respectable media to cooperate; without an explicit denial, respectable media would take it as legitimate news. De Rossi would have been forced to denounce the video—and in doing so, forced to either deny or confirm the rumor, neither of which would be voluntary.

We can see similar risks already in existing technology. If a person gets their hands on your mobile phone, they can send messages to your loved ones, posing as you. A malicious or merely paternalistic acquaintance might say things you think are true but for whatever reason do not wish to say. There is a striking

24 Some philosophers argue that a person's moral interest in privacy is precisely about being able to control their own social self-presentation; see Nagel, "Concealment and Exposure"; Velleman, "The Genesis of Shame"; Marmor, "What Is the Right to Privacy?" Thanks to an anonymous referee for suggesting this connection.

25 Kort, "Portia Heart and Soul."

example in Kristen Roupenian's short story, "Cat Person." Margot is a college student in an unhappy relationship with Robert. She begins ignoring his texts, hoping he will simply disappear from her life. But he keeps trying to contact her. Finally, her friend Tamara takes her phone and sends Robert the following message: "Hi im not interested in you stop textng me." Margot is horrified by this. She imagines Robert "picking up his phone, reading that message, turning to glass, and shattering to pieces."[26] Yet Margot does not send a follow-up message disclaiming authorship or denying the content of the first text. After all, it does express her genuine feelings about Robert. She would not have chosen to say it in quite that way, bluntly and heartlessly, but it is an accurate representation of her thinking. Unable to bring herself to deny the text, she simply allows Robert to believe this is how she ended things.

This case is fictional, but surely many real people have sent imposter texts on behalf of friends (or enemies). And the rise of deepfakes will make these situations both more frequent and more compelling. A text is one thing; a voicemail or video message is much more affecting. As deepfake technology becomes powerful enough to operate in real time, it may be possible to fake a live video call.[27] The more lifelike and compelling the fabrication, the more pressure there will be for the victim to say *something* about it. And when one is illegitimately forced to say something about a topic one would rather not address at all, one has been harmed as a speaker.

These last examples bring to the forefront an important objection: Is there really anything new about deepfakes?[28] Can the same sorts of harms not be caused by already existing technology, such as imposter texts, edited videos, or even forged letters? Is there any cause for *particular* ethical concern about deepfakes?

In a strictly logical sense, the answer seems to be no. The *types* of harm we have considered so far are clearly possible without deepfakes. Rumors alone can damage a person's reputation, producing material harms. Illocutionary harm is possible through low-tech means.

But in a more practical sense, deepfakes are a distinct ethical problem. They make the possibility of these harms much easier to bring about, and therefore a much more realistic threat to ordinary lives. With deepfakes, one need not be a skilled forger, master phone thief, or expert in the dark arts of political rumor-

26 Roupenian, "Cat Person."

27 See Thies, et al., "Face2Face." You can see a demonstration at https://www.youtube.com/watch?v=ohmajJTcpNk.

28 We consider related objections in more detail elsewhere; see Rini, "Deepfakes and the Epistemic Backstop."

mongering in order to successfully compel undesired speech. All it takes is a decently powerful computer and reliable Wi-Fi connection.

Our worry is that this moral problem, while not theoretically unprecedented, will be practically unfamiliar. Ordinary people—rivalrous coworkers, jilted lovers, bored teenagers—will suddenly have the ability to generate compellingly fabricated evidence of anyone doing or saying anything. The most spectacular consequences will involve public figures, but the most morally troubling may happen on the intimate scale of ordinary enmity. How will our day-to-day relationships, the bonds of routine civility, fare when subversion of recorded reality is a realistic temptation? We have no idea, and we think that is a serious problem.

3. PANOPTIC GASLIGHTING AND EXISTENTIAL TRAUMA

There is at least one more way in which deepfakes may generate harms, one which may go beyond facilitating a newly efficient way to do ancient harms. We turn finally to the potential for deepfakes to threaten memory and the existential bases of personhood.

In most deepfake scenarios, there are at least three different participants: a creator who generates fabricated recordings, a target who is falsely represented in the fake, and an audience whose response to the fake causes difficulty for the target. But deepfakes can be troubling even when the target and the audience are the *same person*—that is, when someone views a deepfake falsely depicting *their own* past actions. A fabricated recording could be used to destabilize or even overwrite first-personal, autobiographical memories.

To see the point, imagine that you are in one of those complicated triangular friendships where everyone is a bit of a rival for everyone else's time and attention. (Maybe you are a high school student, or just someone whose life continues to feature a lot of drama.) Imagine that one of your friends claims to have heard you say terrible things about your other friend. You certainly do not remember doing that, and you are pretty sure you would never say such a thing out loud. But now your rival pulls out their phone and plays a video: there you are, at your group's favorite pub, looking and sounding just a bit tipsy. And there you go, saying those terrible things. The video is dated from a year or so ago. "I just found it last night," says your so-called friend, "while I was going through old pictures. Don't worry though. I mean, of *course* I'd never show this to you-know-who."

You would be confused, of course. You would suspect *some* sort of trick. But suppose you had never heard of deepfakes. Suppose you thought this technology was only possible for wealthy Hollywood studios, not something your petty

friend could do on their smartphone. Then what? It is hard to resist video evidence: there you are; you said it. Which should you trust more: your fallible memory or independent recordings?

Psychologists have shown that fake photographic and video evidence can be used to manipulate autobiographical memories. In one study, participants chose their favorite brands of various consumer products and were photographed with their selections. Later, they were shown fake photos (edited by the experimenters) depicting them with different brands, and many were willing to un-self-consciously claim that *these* really were their favorites. In other words, participants were more willing to spontaneously reassign their preferences than to challenge fake photographic evidence about their own choices.[29]

Consumer brand preferences are perhaps not that big a deal. But the same techniques can be used to trick people into accepting that they may have *done* things they did not. In one study, participants shown faked photos of themselves with broken pencils or unsealed envelopes were more likely to later falsely remember having made those things happen.[30] Worst of all: participants who were shown faked video of themselves cheating in a gambling game were willing to sign false confessions, with many confabulating plausible stories to explain to themselves why they might have cheated.[31]

So, if you were shown a video like the one in our story of friend-group rivalry, you very well might believe it. You might fall for it *even if* you suspect some form of trickery. After all, maybe it did happen. Maybe you did forget. Perhaps the video shows you saying things that, yes, you do sometimes think about your friend—though you try never to say them aloud! Maybe you got a bit drunk, vino did its verifying, and you had forgotten by the next morning. It was more than a year ago, after all. Are you sure you know exactly what you said in every pub conversation of years past?

So even if you know that the recording *might* be fake, you cannot be sure. And that is the core of the existential danger of deepfakes: they could be used to create effective skepticism about one's own first-personal memories.

This possibility has several serious consequences that roughly parallel the harms we have already discussed. Most obvious are material harms caused by being tricked. A fake video showing you making a disadvantageous promise or bet might induce you to give up something you should not. Highly honest people would be the most vulnerable to this sort of abuse: even if they know the

29 Hellenthal, Howe, and Knott, "It Must Be My Favourite Brand."

30 Henkel, "Photograph-Induced Memory Errors."

31 Nash and Wade, "Innocent but Proven Guilty."

video might be fake, they may err on the side of honoring even unremembered, uncertain obligations.

A more ominous possibility concerns *gaslighting*, which Kate Abramson defines as "a form of emotional manipulation in which the gaslighter tries (consciously or not) to induce in someone the sense that her reactions, perceptions, memories and/or beliefs are not just mistaken, but utterly without grounds—paradigmatically, so unfounded as to qualify as crazy."[32] Gaslighting typically involves *telling or implying to* people that things are other than as they perceive or remember. Doing so regularly and persistently can wear down their resistance.

Motives for gaslighting can be complicated. In the 1944 film, *Gaslight*, which gave us the term, Charles Boyer's character torments Ingrid Bergman's character to get access to her wealth. Driving her mad is only a means to this end. But in the real world, casual gaslighting can be motivated by social jockeying, domestic abuse, retribution, or even internet trolling. It may not even be deliberate; some manipulators are so skilled that they can gaslight without even realizing what they are doing.

The creator of a deepfake may not set out to gaslight their target. If the goal is to trick a third-party audience in order to cause reputational or illocutionary harm, then deceiving the target of the recording may be a mere side effect. Intentionally or otherwise, in at least some cases deepfakers are likely to make their targets begin to question their own memories.

And it is scarily easily to imagine the extreme case, which we will call *panoptic gaslighting*, where a vicious person sets out to *deliberately* ruin another's grip on reality through systemic use of deepfakes. The abundance of casually recorded and shared videos on social media makes this a very real possibility. Imagine a concerted campaign of just slightly changed videos on Facebook, showing a person doing things at last night's party or last week's dinner that are just a bit different than what the victim remembers. Each individual change is fairly unobtrusive by itself, *except* for its just barely noticeable (to the victim) discordance with memory. Done shrewdly and consistently, this sort of abuse could lead the victim to begin to doubt not just particular memories, but the reliability of their memory altogether.

Panoptic gaslighting would be existentially harmful. On some prominent theories of personal identity, what makes someone the same person over time is their ability to veridically recall earlier experiences.[33] A person who begins to

32 Abramson, "Turning Up the Lights on Gaslighting," 2.

33 See, for instance, Sydney Shoemaker's part of Shoemaker and Swinburne, *Personal Identity*. For further discussion of neo-Lockean "psychological relation" theories of personal identity, see Parfit, *Reasons and Persons*; Johnston, "Human Beings"; and Baker, *Persons and Bodies*.

systematically doubt their memories loses this connection to past selves. Worse still, a person who *accepts* the contents of deepfakes and develops new false memories could experience a form of identity fracture.

This is not just a point of abstruse metaphysics. The first-personal experience of being panoptically gaslit—of coming to doubt one's memories generally—would surely be terrible. It would mean helplessness, dislocation, disintegration. Losing faith in your own memories would gradually undo the foundations of self-respect and the ability to withstand pressure from others. As Trudy Govier puts it:

> To discriminate between apt and ill-founded challenges from others, one needs to trust one's own memory, judgment, and conscience. A person who has no resources to preserve her ideas, values, and goals against criticism and attack from others will be too malleable to preserve her sense that she is a person in her own right, and will therefore be unable to maintain her self-respect.[34]

This sort of vulnerability to manipulation is not incidental, in the way you can be tricked by scam email. Rather, the tension between our own internal self-concept (chiefly through memory) and the ways others perceive us is essential to how we function as social agents. Bernard Williams makes this point while arguing that, in an important sense, individual people may not even *have* determinant beliefs or desires before engaging with others. On his view, our need to make ourselves trustworthy and responsible to others is crucial to bringing our multifarious mental states into coherent order:

> We must leave behind the assumption that we first and immediately have a transparent self-understanding, and then go on either to give other people a sincere revelation of our belief … or else dissimulate in a way that will mislead them. At a more basic level, we are all together in the social activity of mutually stabilizing our declarations and moods and impulses into becoming such things as beliefs and relatively steady attitudes.[35]

Williams's point here is *not* a postmodern denial of objective truth. Rather, he is highlighting the fact that our social attitudes toward the truth—our reliance on and expectations about one another's sincerity and competence—play an essential role in determining not only what we do, but also what we end up be-

34 Govier, "Self-Trust, Autonomy, and Self-Esteem," 111. For related points, see Jones, "The Politics of Intellectual Self-Trust."

35 Williams, *Truth and Truthfulness*, 193. For an important development of this line of thought, see Fricker, *Epistemic Injustice*, 52–55.

lieving, and in an important sense who we *are*. As Williams suggests, this mutual dependence ("we are all together") is a shared predicament, one that demands solidarity and cooperation from people of good faith.

Deepfaked attacks on personal memory are a potent weapon for malefactors who seek to exploit that mutual dependence. A person who has been panoptically gaslit by systemic manipulated video depictions of their past is no longer in a symmetrically dependent relationship with their tormenter. They are instead placed at another's mercy, with not only the contents of their beliefs but also their basic capacity to stabilize their mind upon any determinate belief state, held hostage to the dubitable goodwill of a deceiver. Deepfakes are more than just dishonest; they hold the potential to truly destroy individuals.

4. CONCLUSION

Deepfakes may have valuable commercial and artistic applications. They might permit new sorts of harmless fun. Related technology has already been used to protect the identities of victims testifying about atrocities.[36] But they also might lead to new harms, and not just the obvious practical consequences of epistemic malfeasance.

We have surveyed three categories of distinctive harm: frankenporn objectification, illocutionary harm, and existentially traumatic panoptic gaslighting. We are sure that more devious minds than ours are already at work on others.

This may seem like a grab bag of distinct ethical risks, only loosely clustered around the technological vector of deepfakes. But we believe there is a more fundamental commonality to the worries we have raised. It is not an accident that all involve the use of epistemic malfeasance to achieve illicit social manipulation. Whether in objectifying frankenporn, cruel illocutionary harm, or identity-sapping panoptic gaslighting, these uses of deepfakes show the extent that human ethical life is dependent on our epistemic relations.

In recent decades, ethicists and epistemologists have recognized a growing overlap between their concerns and even their methods.[37] We think this a valuable and timely development, as current events make increasingly apparent the social implications of epistemic contention (such as doubt in scientific experts,

36 The documentary film *Welcome to Chechnya* uses a deepfake-like technique to project the faces of volunteer actors over those of real victims of homophobic persecution. As the filmmakers explain, this allowed them to preserve the emotional intensity of their informants' testimony without placing them in greater danger. See Thomson, "Digital Disguise."

37 See, for example, Cuneo, *The Normative Web*; Marušić, *Evidence and Agency*; Basu and Schroeder, "Doxastic Wronging"; Srinivasan, "Radical Externalism."

fake news, deep disagreement, "post-truth"). We believe that deepfakes are yet another facet of this worrisome convergence, and we hope thoughtful minds turn to forestalling their deep harms.[38]

York University, Toronto
rarini@yorku.ca

leahliora@gmail.com

REFERENCES

Abramson, Kate. "Turning Up the Lights on Gaslighting." *Philosophical Perspectives* 28 (2014): 1–30.

Armstrong, Drew. "Planned Parenthood Cut Off from Federal Funding Under Trump Rule." Bloomberg, February 22, 2019. https://www.bnnbloomberg.ca/planned-parenthood-cut-off-from-federal-funding-under-trump-rule-1.1218733.

Baker, Lynne Rudder. *Persons and Bodies: A Constitution View*. Cambridge: Cambridge University Press, 2000.

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.

Bauer, Nancy. *How to Do Things with Pornography*. Cambridge, MA: Harvard University Press, 2015.

Califano, Joseph A. *The Triumph and Tragedy of Lyndon Johnson: The White House Years*. New York: Simon and Schuster, 1991.

Cawston, Amanda. "The Feminist Case against Pornography: A Review and Re-evaluation." *Inquiry* 62, no. 6 (2019): 624–58.

Chesney, Robert, and Danielle Keats Citron. "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security." *California Law Review* 107, no. 6 (2019): 175.

Cole, Samantha. "AI-Assisted Fake Porn Is Here and We're All Fucked." *Motherboard*, December 11, 2017. https://motherboard.vice.com/en_us/article/gydydm/gal-gadot-fake-ai-porn.

38 We owe thanks to the editors and referees for the journal, as well as to audiences at the University of Toronto Centre for Ethics and the Department of Philosophy at the University of Waterloo. This research was supported by York University's Vision: Science to Applications (VISTA) project.

———. "Deepfakes Were Created as a Way to Own Women's Bodies—We Can't Forget That." *Vice*, June 18, 2018. https://www.vice.com/en_us/article/nekqmd/deepfake-porn-origins-sexism-reddit-v25n2.

Cox, Kate. "Deepfake Revenge Porn Distribution Now a Crime in Virginia." *Ars Technica*, July 1, 2019. https://arstechnica.com/tech-policy/2019/07/deepfake-revenge-porn-distribution-now-a-crime-in-virginia.

Cuneo, Terence. *The Normative Web: An Argument for Moral Realism*. Oxford: Oxford University Press, 2007.

Diamond, Jeremy. "Trump: I Would Shut Down Government over Planned Parenthood." CNN, August 4, 2015. https://www.cnn.com/2015/08/04/politics/donald-trump-government-shutdown-planned-parenthood/index.html.

Dworkin, Andrea. "Against the Male Flood: Censorship, Pornography, and Equality." *Harvard Women's Law Journal* 8, no. 1 (1985): 1–29.

Farid, Hany. "Digital Forensics in a Post-Truth Age." *Forensic Science International* 289 (August 2018): 268–69.

Fricker, Miranda. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford: Oxford University Press, 2007.

Govier, Trudy. "Self-Trust, Autonomy, and Self-Esteem." *Hypatia* 8, no. 1 (Winter 1993): 99–120.

Harris, Douglas. "Deepfakes: False Pornography Is Here and the Law Cannot Protect You." *Duke Law and Technology Review* 17 (2019): 99–127.

Hellenthal Maria V., Mark L. Howe, and Lauren M. Knott. "It Must Be My Favourite Brand: Using Retroactive Brand Replacements in Doctored Photographs to Influence Brand Preferences." *Applied Cognitive Psychology* 30, no. 6 (November/December 2016): 863–70.

Henkel, Linda A. "Photograph-Induced Memory Errors: When Photographs Make People Claim They Have Done Things They Have Not." *Applied Cognitive Psychology* 25, no. 1 (January/February 2011): 78–86.

Johnston, Mark. "Human Beings." *Journal of Philosophy* 84, no. 2 (February 1987): 59–83.

Jones, Karen. "The Politics of Intellectual Self-Trust." *Social Epistemology* 26, no. 2 (2012): 237–51.

Kelion, Leo. "Reddit Bans Deepfake Porn Videos." BBC News, February 7, 2018. https://www.bbc.com/news/technology-42984127.

Kliff, Sarah. "I Watched 12 Hours of the Planned Parenthood Sting Videos. Here's What I Learned." *Vox*, September 9, 2015. https://www.vox.com/2015/8/13/9140849/planned-parenthood-videos-unedited.

Kort, Michele. "Portia Heart and Soul." *The Advocate*, August 29, 2005. https://

www.advocate.com/politics/commentary/2005/08/29/portia-heart-amp-soul.

Langton, Rae, and Jennifer Hornsby. "Free Speech and Illocution." *Legal Theory* 4, no. 1 (March 1998): 21–37.

Li, Yuezun, and Siwei Lyu. "Exposing DeepFake Videos by Detecting Face Warping Artifacts." *arXiv*, 2019. https://doi.org/10.48550/arXiv.1811.00656.

Liu, Lingjie, Weipeng Xu, Michael Zollhöfer, Hyeongwoo Kim, Florian Bernard, Marc Habermann, Wenping Wang, and Christian Theobalt. "Neural Rendering and Reenactment of Human Actor Videos." *ACM Transactions on Graphics* 38, no. 5 (October 2019). https://doi.org/10.1145/3333002.

MacKinnon, Catharine. *Feminism Unmodified: Discourses on Life and Law*. Cambridge, MA: Harvard University Press, 1987.

Maitra, Ishani. "Silencing Speech." *Canadian Journal of Philosophy* 39, no. 2 (June 2009): 309–38.

Marmor, Andrei. "What Is the Right to Privacy?" *Philosophy and Publics Affairs* 43, no. 1 (Winter 2015): 3–26.

Marušić, Berislav. *Evidence and Agency: Norms of Belief for Promising and Resolving*. Oxford: Oxford University Press, 2015.

McKinney, Rachel Ann. "Extracted Speech." *Social Theory and Practice* 42, no. 2 (April 2016): 258–84.

Nagel, Thomas. "Concealment and Exposure." *Philosophy and Public Affairs* 27, no. 1 (Winter 1998): 3–30.

Nash, Robert A., and Kimberley A. Wade. "Innocent but Proven Guilty: Eliciting Internalized False Confessions Using Doctored-Video Evidence." *Applied Cognitive Psychology* 23, no. 5 (July 2009): 624–37.

Nussbaum, Martha C. "Objectification." *Philosophy and Public Affairs* 24, no. 4 (Autumn 1995): 249–91.

Parfit, Derek. *Reasons and Persons*. Oxford: Oxford University Press, 1984.

Rini, Regina. "Deepfakes and the Epistemic Backstop." *Philosophers' Imprint* 20, no. 24 (August 2020): 1–16.

Roupenian, Kristen. "Cat Person." *The New Yorker*, December 4, 2017. https://www.newyorker.com/magazine/2017/12/11/cat-person.

Schiller, Henry Ian. "Illocutionary Harm." *Philosophical Studies* 178, no. 5 (May 2021): 1631–46.

Searle, John. *The Rediscovery of the Mind*. Cambridge, MA: MIT Press, 1992.

Shead, Sam. "Facebook to Ban 'Deepfakes.'" BBC News, January 7, 2020. https://www.bbc.com/news/technology-51018758.

Shoemaker, Sydney, and Richard Swinburne. *Personal Identity: Great Debates in Philosophy*. Hoboken, NJ: Blackwell, 1984.

Simonite, Tom. "Most Deepfakes Are Porn, and They're Multiplying Fast." *Wired*, October 7, 2019. https://www.wired.com/story/most-deepfakes-porn-multiplying-fast.

Soeriaatmadja, Wahyudi. "Man Who Uploaded Controversial Video of Ex-Jakarta Governor Ahok Sentenced to Jail." *The Straits Times*, November 14, 2017. https://www.straitstimes.com/asia/se-asia/man-who-uploaded-controversial-ahok-video-sentenced-to-jail.

Srinivasan, Amia. "Radical Externalism." *Philosophical Review* 129, no. 3 (July 2020): 395–431.

Strossen, Nadine. *Defending Pornography: Free Speech, Sex, and the Fight for Women's Rights*. New York: Scribner, 1995.

Thies, Justus, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. "Face2Face: Real-Time Face Capture of Reenactment of RGB Videos." Conference on Computer Vision and Pattern Recognition, March 17, 2016. https://www.youtube.com/watch?v=ohmajJTcpNk.

Thomson, Patricia. "Digital Disguise: 'Welcome to Chechnya''s Face Veil Is a Game Changer in Identity Protection." *Documentary Magazine*, June 30 2020. https://www.documentary.org/column/digital-disguise-welcome-chechnyas-face-veil-game-changer-identity-protection.

Velleman, J. David. "The Genesis of Shame." *Philosophy and Public Affairs* 30, no. 1 (January 2001): 27–52.

Von der Burchard, Hans Joachim. "Belgian Socialist Party Circulates 'Deep Fake' Donald Trump Video." *Politico*, May 21, 2018. https://www.politico.eu/article/spa-donald-trump-belgium-paris-climate-agreement-belgian-socialist-party-circulates-deep-fake-trump-video.

Williams, Bernard. *Truth and Truthfulness: An Essay in Genealogy*. Princeton: Princeton University Press, 2002.

Journal of Ethics and Social Philosophy
VOL. 22, NO. 2 · JULY 2022
https://doi.org/10.26556/jesp.v22i2.1258

FAKE NEWS AND DEMOCRACY

Merten Reglitz

THIS PAPER offers a philosophical explanation of the moral relevance of online fake news. Fake news has become a sweeping political catchphrase that sparks worries about political manipulation. It is used by some to discredit political opponents, the free press, or any dissenting political opinion in general. However, behind this buzzword lies a serious problem, made possible by the internet and social media, for democratic institutions. The British House of Commons' Digital, Culture, Media, and Sport Committee deems fake news a potential threat "to our democracy and our values."[1] Likewise, a report of the Directorate-General for Communication Networks, Content, and Technology of the European Commission sees fake news as generating "threats to democratic political processes, including integrity of elections, and to democratic values that shape public policies."[2] Thus, at the highest level of some of the most powerful democratic institutions, online fake news has been identified as an important problem. But also in academia, prominent social scientists worry about the impact of fake news on public discourse.[3]

However, there is a problem with these concerns about fake news: the most significant empirical studies of the phenomenon conclude that fake news is quite ineffective in convincing people of the veracity of its content. It is thus not obvious *how exactly* fake news threatens the operations and values of democratic institutions. Allcott and Gentzkow's study of voters' overall exposure to fake news in the 2016 US presidential election suggests that fake news did not decisively influence this election because voters were not exposed to false news reports often enough.[4] More importantly still, their study indicates that only a

1 Digital, Culture, Media, and Sport Committee, *Disinformation and "Fake News,"* 3.

2 European Commission, *A Multi-Dimensional Approach to Disinformation*, 5.

3 Lazer et al., "The Science of Fake News."

4 Allcott and Gentzkow, "Social Media and Fake News in the 2016 Election," 232. They estimate that "the average US adult might have seen perhaps one or several [fake] news stories in the months before the election" (213).

small minority of voters actually *believed* the fake news stories they read.[5] Another study that analyzes user engagement with fake news on social media finds that the vast majority of Facebook users did not share articles from fake news websites during the 2016 US presidential elections at all.[6] A third study investigating how Twitter users engaged with fake news during this election states that "engagement with fake news sources was extremely concentrated. Only 1% of individuals accounted for 80% of fake news source exposures, and 0.1% accounted for nearly 80% of fake news sources shared."[7] A study of the prevalence of computational propaganda in the run-up to the elections for the 2019 European Parliament finds that "less than 4% of the sources circulating on Twitter during our data collection were junk news, with users sharing higher proportions of links to professional news sources overall; on Facebook, junk news outlets tended to receive more engagement per story, but are seen, shared, and liked by far less people overall."[8] Moreover, fake news is often thought to cause irreconcilable political difference among voters in liberal democracies. However, an important study finds that political polarization in the United States started to increase well before the internet was widespread.[9] Polarization, according to this study, is particularly prevalent among older US voters who are less active on the internet and thus less exposed to online fake news. This suggests that existing political polarization is a climate in which fake news can become a problem, rather than fake news itself being the cause of political polarization. Finally, political disinformation and propaganda are as old as human politics. So why exactly should liberal democracies worry about fake news?

It might appear that the philosophical inquiry into the moral relevance of fake news could stop right here. After all, fake news is believed by fewer people than is thought. Thus, fake news might seem to be a nonissue. However, the situation is not that simple. Fake news *does* present a problem because what citizens believe matters for their opinions of their democratic institutions and the moral justifiability of these institutions. And fake news is indeed believed by citizens to be influential—regardless of whether that belief is true. Surveys among voters show that there is widespread concern about fake news among democratic citizens. According to a poll conducted by *The Hill*, 65 percent of Americans

5 "After weighting for national representativeness, 15 percent of survey respondents recalled seeing the Fake stories, and 8 percent both recalled seeing the story and said they believed it" (Allcott and Gentzkow, "Social Media and Fake News in the 2016 Election," 227).

6 Guess, Nagler, and Tucker, "Less Than You Think."

7 Grinberg et al., "Fake News on Twitter during the 2016 US Presidential Election," 374.

8 Marchal et al., *Junk News during the EU Parliamentary Elections*, 6.

9 Boxell, Gentzkow, and Shapiro, "Is the Internet Causing Political Polarization?"

believe that fake news is prevalent in the mainstream media.[10] A different poll shows that American voters have very concrete concerns about the effects of fake news: 88 percent are worried that fake news has spread confusion among voters.[11] Another survey by the Pew Research Center indicates that Americans consider fake news a bigger threat than terrorism.[12] This concern about misinformation spread online is not exclusive to the United States. A survey on fake news and disinformation online conducted by the European Commission in all twenty-eight European Union member states among twenty-six thousand participants finds that "more than eight in ten respondents (85%) think that the existence of fake news is a problem in their country.... A similar proportion (83%) say that it is a problem for democracy in general."[13] And a survey by PwC reports that 71 percent of German voters were concerned about the influence of fake news on the 2019 elections for the European Parliament.[14] Perceptions, not only facts, matter in politics, and that is why fake news stories can be a problem for democracies even though their content is not widely believed. Philosophical analysis can help clarify this connection between concerns about fake news and the legitimacy of democratic institutions.

Accordingly, my argument in this paper is that online fake news threatens democratic values and processes by playing a crucial role in reducing the perceived legitimacy of democratic institutions. This decrease in perceived legitimacy is the outcome of the primary effect that fake news has on citizens: even if its content is not believed, fake news can be a major cause of a loss of citizens' epistemic trust in each other's political views and judgment. Such a loss of trust in each other is problematic for democratic institutions since these rely for their acceptance and functioning on citizens seeing them as morally justified. Critiques of fake news often focus on citizens' loss of trust in their mainstream media. While this is indeed part of the problem, I will argue that the main threat of fake news pertains to the loss of epistemic trust citizens have *in each other*. Fake news is thus a moral problem insofar as we think of democracies as a morally special, or at least a particularly valuable, form of government.[15] This paper is significant because—unlike most discussions of fake news that assume that citizens are likely to accept these falsehoods as true—it takes seriously the empirical studies

10 Easley, "Poll: Majority Says Mainstream Media Publishes Fake News."

11 Barthel, Mitchell, and Holcomb, *Many Americans Believe Fake News Is Sowing Confusion*.

12 Mitchell et al., *Many Americans Say Made-Up News Is a Critical Problem that Needs to Be Fixed*.

13 European Commission, *Fake News and Disinformation Online*, 4.

14 PricewaterhouseCoopers, *"Fake News,"* 8.

15 Christiano, *The Constitution of Equality*.

that assert that most people do not believe the content of fake news and explains why we should nonetheless consider fake news a morally significant problem.

Contemporary fake news has a debilitating effect on relatively well-functioning democracies because it is spread via social media platforms that operate according to a particular business model. Facebook, for example, generates profit by enabling third parties to pay to influence the behavior of its users by sending them advertisements, political messages, and almost any kind of information—whether factually correct or not.[16] It is ultimately online information technology and social media platforms that make fake news a threat to every democratic system's foundation—namely, citizens' belief that the system is morally justified as a whole and deserves their allegiance.

My argument is developed as follows: section 1 offers a characterization of fake news. Section 2 explains the role that epistemic trust plays for the functioning of democracy, which is the crucial resource eroded by fake news. Section 3 subsequently shows how fake news undermines epistemic trust among democratic citizens—even if it is not widely consumed, shared, or believed—and considers an important objection to this argument. Section 4 spells out the most likely ways in which the loss of epistemic trust undermines the sociological legitimacy (the perceived moral justification), and potentially the normative legitimacy (the actual moral justification), of democratic institutions. Finally, section 5 suggests a number of obligations that democratic institutions can be said to have in light of the trust-undermining effect of fake news. Ultimately, online fake news is but a symptom illustrating a larger problem: the internet has enormous effects on democratic processes that we have yet fully to understand. Nonetheless, fake news is a powerful enough influence on democratic values and processes to deserve political action and a thorough philosophical analysis of its own.

1. CHARACTERIZING FAKE NEWS

Before I analyze the effects of fake news on democracy, it is necessary to delineate the meaning of the term itself. "Fake news" has become a term that is used to denote very different things: it is employed to discredit political opponents or the respectability of particular news outlets, and it is used colloquially to simply refer to untruths in any given context. However, the phenomenon that public institutions like the European Commission and the British Parliament are concerned about most plausibly entails at least three features:

1. Fake news contains false information.

16 Zuboff, *The Age of Surveillance Capitalism*, 508–12.

2. Fake news is created with deceptive intent.
3. Fake news is presented as resembling traditional news items (even though it is not produced in accordance with editorial standards).

First, if fake news did not contain false information, it would be genuine news that would not present a concern. Second, deceptive intent necessarily is involved in the creation of fake news; otherwise, honestly but improperly researched information (e.g., inaccurate reporting) would be an instance of fake news.[17] The deceptive intent can take different forms: The creator of fake news might intend for the false information simply to be shared. Alternatively, they might want a fake news story to attract visitors to the hosting website to generate advertisement revenue. What is decisive is that the real motivation of the creator of fake news is not transparent, and that their intention is *not* to distribute accurate information.[18] Finally, the phenomenon at the heart of the concerns of institutions like the European Union or the British Parliament is not simply any kind of falsehood shared online but rather false information dressed up as a genuine news item.[19] This final feature of fake news is important for understanding its negative effects on citizens' trust in the information environment of their democratic societies.

Examples of the politically relevant phenomenon in the focus of this article include stories such as candidate Donald Trump being endorsed by the pope, candidate Hillary Clinton selling weapons to ISIS and being a member of a child pornography ring, or refugees raping women in German public baths.[20] Many of these false news stories have been produced for political reasons, others as clickbait for economic gain.[21] Fake news is not a new phenomenon.[22] However, what is new in our virtual age is that such falsehoods can be disseminated cheaply, quickly, and globally via the internet because the internet has lowered the costs of sending and receiving information and widened the potential audience of all published content. People's ability to set up websites and to post information and messages means that the traditional news sources and gatekeepers of facts,

17 Fallis and Mathieson, "Fake News Is Counterfeit News."

18 Rini, "Fake News and Partisan Epistemology," 44–45.

19 Fallis and Mathieson, "Fake News Is Counterfeit News."

20 Silverman, "This Analysis Shows How Viral Fake Election News Stories Outperformed Real News on Facebook"; *Der Spiegel*, "Is There Truth to the Refugee Rape Reports?"

21 Silverman, "This Analysis Shows How Viral Fake Election News Stories Outperformed Real News on Facebook."

22 See McKernon, "Fake News and the Public"; and Soll, "The Long and Brutal History of Fake News."

such as traditional media and public institutions, have been demoted to some among many sources of information to which citizens are exposed.[23]

It might be tempting to dismiss the relevance of these new possibilities that the internet and social media platforms offer by pointing to the aforementioned studies about the ineffectiveness of fake news. If few are exposed to and believe these falsehoods, how could such falsehoods undermine entire democratic systems? I argue that the danger of fake news lies in citizens' (incorrect) belief that fake news is actually effective in manipulating their fellow citizens. Social media platforms offer channels through which such manipulation is at least theoretically possible. However, in order to see why citizens' concerns render fake news an important problem for democratic politics, we require an understanding of those essential features of liberal democracies that are particularly susceptible to the effects of fake news.

2. DEMOCRACY AND EPISTEMIC TRUST

There are many conditions for the functioning of democracy. There are, for instance, minimal institutional requirements such as formally equal votes, equal basic freedoms (e.g., free speech and the right to run for public office), and an independent judiciary.[24] Without these elements, states are unlikely to respect the political liberties of their citizens. There are also socioeconomic conditions for the viability of democracy, such as limited material inequality.[25] If wealth and income become increasingly unevenly distributed, citizens are no longer likely to consider a democracy as working for them, and instead they may turn toward nondemocratic political options. However, for the purpose of understanding the dangers that fake news can pose to democratic processes and values, we need to look specifically at the essential element of epistemic trust.

Democracies are based on collective public decision-making that expresses the moral equality of all eligible members of the community, who have an equal say in the process. Democracies are thus special as the only form of government that realizes the moral equality of its members publicly in its collective decision-making.[26] Because of this, democracies are also unlike other forms of government in that their members are epistemically dependent on each other. That is to say that the life of every citizen is determined to a significant degree by what all others think is morally correct or factually the case, because the quality of the

23 Lazer et al., "The Science of Fake News."

24 Christiano, "Self-Determination and the Human Right to Democracy," 461.

25 Milanovic, *Global Inequality*, ch. 4.

26 Christiano, *The Constitution of Equality*, ch. 3.

laws made by the democratic body depends, to a significant extent, on the quality of the judgments of all those who elect the lawmakers. As Michael Fuerstein puts this point: "As democratic citizens, we are epistemically interdependent in the respect that our epistemic status on politically significant issues is contingent on the knowledge of others and our ability to trust them in accepting it."[27] If, for instance, a large part of the population is misinformed about the risks and benefits of vaccinations and consequently elects politicians that restrict or ban vaccinations, my options for being vaccinated become limited as well—even though I would correctly believe that the benefits of vaccinations vastly outweigh the risks. To take another example, achieving herd immunity against a novel virus might require a population vaccination rate of 90 percent. However, a significant number of citizens might be unwilling to be vaccinated against the virus because they hold implausible or unsubstantiated beliefs about the vaccine, so that herd immunity is unattainable. Additionally, there might be no democratic support for imposing vaccine mandates. In this situation, even those who are vaccinated and hold plausible beliefs regarding the vaccine remain at risk from the virus because it might mutate or infect them and cause serious illness as it continues to circulate in the population.

Because we are epistemically interdependent as democratic citizens, we need to have a significant degree of epistemic trust in our political community to accept its laws and decisions as morally justified and binding. "Epistemic trust" denotes the idea that a person accepts information and reasons offered by another "because of the belief that the speaker is sufficiently epistemically reliable, where reliability concerns both the epistemic *competence* of the speaker—the likelihood that her beliefs in some domain are true—and her *sincerity*—the likelihood that she will represent what she believes accurately."[28] Pervasive disagreement about our collective goals or about how to achieve them creates a situation in which citizens frequently are not convinced by each other's reasons. Yet voters do not have to agree to accept the democratic decision procedure as morally justified. Rather, voters must have some trust in their cocitizens' competence and sincerity to accept as politically binding their judgment and the political decisions based on it. Thus, the less I am convinced that political decisions are based on the best available evidence, the more grounds I have to reject their legitimacy.

Importantly, for citizens to accept their democratic institutions and procedures as morally justified, they need to have epistemic trust in a number of people. First, political candidates or officeholders who are deemed to be ignorant

27 Fuerstein, "Epistemic Trust and Liberal Justification," 181–82.

28 Fuerstein, "Epistemic Trust and Liberal Justification," 181.

or disingenuous normally do not carry democratic majorities. Citizens may, of course, deem knowledge and sincerity less important qualities if a political candidate promises to promote their most important goals. Yet, normally, achieving one's desired outcomes requires taking into account relevant facts. Thus, political candidates or officeholders need to be perceived to act on the basis of the best available evidence and on the objectives they profess to promote. Second, citizens must trust their primary sources of information. Politics is too complex for every individual to know all the relevant facts. Thus, it is traditionally the primary function of the free press to supply citizens with the information they require to form political opinions and make political choices. If citizens distrust most of the main information outlets, they can become distrustful of their fellow citizens who are exposed to the same outlets as well.

Third and most importantly, citizens need to have epistemic trust in *each other* because of their interdependence. If I believe that the vast majority of my fellow citizens hold beliefs that are completely factually mistaken, our conflict is not simply a moral disagreement. Rather, this disagreement is aggravated by our expectations about others' beliefs about what is factually the case. If my distrust in others' competence and sincerity reaches critical levels, I will stop trusting them to be capable of making joint decisions that fundamentally shape and determine my life. Epistemic distrust toward my fellow citizens will also affect to what extent I feel our collective decisions deserve my respect because my fellow citizens' ability to make political choices based on truth or the best available evidence seems problematically limited. It is thus the need for epistemic trust of citizens in each other that is relevant for understanding the danger that fake news poses to the viability of democratic values and structures—even if these falsehoods are not actually believed or shared by most who read them.

3. How Fake News Undermines Epistemic Trust

A widespread concern about fake news is that it is indeed believed by those who consume it and that the democratic process is therefore undermined by the choices of manipulated voters who do not act in their own best interests or for the common good. This concern is one motivation behind, for example, attempts to identify ways of "inoculating" individuals against fake news.[29] The same worry is also expressed in the concern about "echo chambers"—a term that describes individuals only communicating with like-minded persons and primarily consuming information reflecting their own views. In such echo

29 Roozenbeek and van der Linden, "Fake News Game Confers Psychological Resistance against Online Misinformation."

chambers, people are not exposed to the views and reasons of others, and thus no critical exchange of perspectives takes place. Fake news is thought to contribute to the echo chamber phenomenon by reaffirming people's suspicions about their political opponents. Such group polarization could turn the political climate among citizens from cooperative into adversarial. For that reason, Cass Sunstein, for instances, worries that "fake news is everywhere. To date, social media have not helped produce a civil war, but that day will probably come."[30]

The problem with this concern is that, as pointed out above, it is contradicted by empirical research. The more fundamental threat that fake news poses is a different one. Fake news can indeed have a serious and debilitating effect on democratic processes and values regardless of whether its content is widely believed. The main danger that fake news presents to democracies is that it destroys the epistemic trust of voters in each other.[31] Examples of such a loss of epistemic trust are documented in two recent polls conducted by the Pew Research Center. The first survey shows that a majority of US citizens have "little or no confidence in the political wisdom of the American people."[32] And according to the second poll, 54 percent of Americans have lost confidence in each other *because* of fake news.[33] Citizens' *perceptions* about the effectiveness of fake news, not its actual effectiveness, are decisive for its destructive potential. As mentioned in the introduction, multiple surveys covering the United States and the European Union show that widespread concern about the prevalence of fake news exists because it is feared to spread confusion among citizens. Thus, these surveys can be taken as empirical evidence for the main argument of this paper.

The argument also finds support in the hypothesis of the third-person effect.[34] According to this well-supported thesis, people generally believe others to be more vulnerable to media effects than themselves. This third-person effect in relation to online fake news is indirectly borne out in surveys as well.[35] According to a Pew Research Center survey from 2016, 88 percent of surveyed US Americans worry that fake news has spread confusion among the citizenry, while at the same time, 39 percent of respondents are "very confident" and 45 percent are at least "somewhat confident" in their ability to identify fake news. That is to

30 Sunstein, *#Republic*, 11.

31 Rini makes a similar point in "Social Media Disinformation and the Security Threat to Democratic Legitimacy," 12.

32 Pew Research Center, *The Public, the Political System and American Democracy*.

33 Mitchell et al., *Many Americans Say Made-Up News Is a Critical Problem that Needs to Be Fixed*.

34 Davidson, "The Third-Person Effect in Communication."

35 Jang and Kim, "Third Person Effects of Fake News," 296.

say, 84 percent of respondents worry about fake news (only or predominantly) because of others' susceptibility to these falsehoods, not because of their own.[36]

It is consequently plausible to argue that fake news can have an undermining effect on epistemic trust in democratic populations if it is believed to be widespread and effective. The knowledge that many of these falsehoods are circulating and might be accepted by others arguably is sufficient to undermine cocitizens' trust in each other. From each individual's perspective, it appears sensible to assume that the more falsehoods circulate, the more likely it is that others have read and come to believe them. Once voters are convinced of each other's epistemic unreliability, it becomes difficult for them to have a respectful exchange of views and arguments. After all, if I am convinced that my counterpart is not very competent in distinguishing facts from lies, I also have little reason to believe that their political views are generally reasonable. As Fuerstein points out, a person in this situation might even begin to act strategically by employing falsehoods as a reaction to the expected lies or confusion of others.[37] In this way, epistemic distrust can lead to a vicious circle: my expectation that others will attempt to manipulate me begets my attempt to manipulate others, and so on. This situation of mutual epistemic distrust, in turn, has grave consequences for citizens' views of the moral justifiability of the political discourse and the institutions that rest on it. Citizens might come to view particular democratically made laws, parts of the democratic system, or even the entire democratic system as morally unjustified. Citizens will then no longer consider these laws or systems as worthy of their respect since they are seen to be based on the false beliefs and bad choices of a manipulated majority of voters. This is to say that (the fear of) fake news can threaten the perceived moral justification of democratic processes and norms. And because democracy is a morally valuable form of government, the fact that fake news threatens its stable functioning makes these falsehoods a moral issue.

This argument is threatened by a possible objection. If concerns about false beliefs resulting from the consumption of fake news undermine citizens' epistemic trust in each other, why should regular political disagreement based on opposing beliefs not equally cause distrust among the populace? This objection poses an apparent dilemma for my argument.[38] Political disagreements are indeed often based on people believing different things to be true. For instance, I might be in favor of liberal immigration policies because I rightly believe that immigrants bring cultural and economic benefits, while my opponent incorrect-

36 Barthel, Mitchell, and Holcomb, *Many Americans Believe Fake News Is Sowing Confusion.*

37 Fuerstein, "Epistemic Trust and Liberal Justification," 187.

38 I thank an anonymous reviewer of the *Journal of Ethics and Social Philosophy* for suggesting this objection.

ly believes that immigrants are more criminal than natural-born citizens. Since I know my opponent's beliefs are false, should I not therefore distrust them? Thus, on the first horn of the dilemma, if political disagreement based on diverging beliefs generally foments distrust, why should we consider fake news a special danger for democracies?

Further, critics of democracy, such as Jason Brennan, endorse a generally negative view of citizens' attitudes toward each other.[39] According to Brennan, democratic processes necessarily lead to distrust and strife for two reasons. First, there is too much relevant political information for each voter to reach an informed political opinion. Voters' individual political influence, moreover, is too small to make it worth their while to spend time and effort to acquire more accurate information. They are thus likely to hold political beliefs that are based on inaccurate information. Second, even individuals who possess somewhat more accurate political information are ultimately corrupted by their own biases and partisan tendencies. These biases push them to accept political beliefs that fit with the views of other members of their group. Both tendencies increase political polarization and disagreement about politically relevant facts, which in turn foster a lack of trust toward political opponents. Brennan's proposed solution is to transition from democratic governance to an epistocracy where those who know better have privileged power to make political decisions. For Brennan, the regular workings of democracies, and not fake news, are sufficient for undermining trust among the populace.

Alternatively, on the second horn of the dilemma, if we assume that political disagreements based on opposing beliefs do *not* generally generate epistemic distrust, it is unclear why fake news (that generates fears about others having false beliefs) should lead to such distrust. After all, it seems unlikely that the mere fear of others holding inaccurate beliefs can cause distrust when actual knowledge of others holding opposing beliefs does not.

However, this dilemma does not pose a problem for my argument because there is a more plausible third option. There are two factors that explain why fake news is indeed a special problem for democracies. First, not all political disagreements give reasons to epistemically distrust one's opponents, for several reasons. Not all political disagreements involve disagreements about basic facts. Some of these disputes are about the right ways to achieve agreed upon goals. We can disagree, for instance, about the best way to tackle climate change, to fight poverty, or to promote gender equality. Only if the basic facts involved (climate change, blameless poverty, equality between genders) are disputed is there a reason to

39 Brennan, *Against Democracy*, ch. 2.

distrust an opponent's ability to grasp facts. Other disagreements that do not lead to distrust are about facts where the evidence is either inconclusive (i.e., too complex) or indeterminate (i.e., insufficient) to arrive at definite conclusions.[40] An example of the first is whether more free trade is always preferable to less; an example of the latter is whether God exists. Yet other disagreements that do not lead to distrust are of a normative kind—for example, whether the state should require citizens to confront the issue of organ donation by mandating a one-time decision about opting into such a scheme.[41] In all these cases, disagreement is not unreasonable and does not have to lead to questions about opponents' epistemic competence, which shows that not all political disagreements are grounds for epistemic distrust.

Second, the problem of fake news has to be seen within the information context in which it arises. The internet has changed the information environment in which democratic citizens find themselves. Before the internet became widespread, citizens faced more relevant information than they were able to take into account. But in the offline society, there was a diverse but relatively limited range of information sources—for example, newspapers and TV stations. All citizens used a much smaller pool of information shortcuts, such as journalistic media.[42] This relatively limited range of news sources meant that there was greater information overlap among citizens, and citizens were aware of this fact. These common news sources also allowed for more "shared experiences" and knowledge that in turn gave less reason to worry about the beliefs of others.[43] Fake news was a much less likely occurrence within this more limited pool of information sources because there were fewer channels that could reach all citizens. Mainstream channels often competed over their reputations for presenting accurate rather than false information to their audiences. This situation dramatically changed once the internet became widespread. Now, citizens are confronted with even more information, as almost everyone with the equipment and relevant technical knowledge is able to produce and distribute their own "news." Due to the increasing number of sources of information, citizens now have fewer shared experiences and less shared knowledge.[44] Moreover, citizens have become aware that others might receive information through entirely different channels and share fewer beliefs with them.

Thus, in our digitalized information environment, disputes that involve con-

40 Gaus, *Justificatory Liberalism*, 152–53.

41 See Talisse, *Democracy and Moral Conflict*, ch. 1.

42 Christiano, "Democracy, Participation, and Information."

43 Sunstein, *#Republic*, 140.

44 Sunstein, *#Republic*, 144.

troversy about basic facts are the kind of political disagreement especially likely to generate epistemic distrust among citizens. And here fake news indeed presents a special danger because, as explained above, fake news masquerades as genuine news. Thus, if I am worried about others believing fake news, I am directly worried about their epistemic competence in a way that I am not in the case of other disputes. After all, it is not obviously unreasonable to arrive at different conclusions about the moral status of fetuses or the existence of God when considering the same (biological and cosmological) basic facts. However, there are no two similarly reasonable beliefs about whether most immigrants are criminals or whether our top politicians are members of a globally operating ring of pedophiles. Thus, if I am worried that others believe fake news, I am essentially worried that they are *incapable* of distinguishing veritable from untrustworthy information sources. I cannot thus epistemically empathize with them (i.e., put myself in their shoes) or epistemically respect their beliefs. This explains why not all forms of political disagreement have to lead to epistemic distrust among citizens but why fake news necessarily sows such distrust. There is thus no dilemma threatening the present argument: fake news is a problem for epistemic trust in democracies. To recognize this, we do not have to deny that there are some political disagreements that also generate such distrust or that (as Brennan argues) many political disputes involve strong emotions that challenge public debate.

4. FAKE NEWS AND PERCEIVED DEMOCRATIC LEGITIMACY

How does the loss of epistemic trust among citizens affect democracies? Any answer to this question has to be speculative to some degree because most of the time it will be extremely difficult to determine with certainty that particular political developments are primarily or significantly the result of fake news stories. What can be identified, though, are the most likely effects that fake news can have on democratic systems, some of which are observable today. Most likely, the loss of epistemic trust caused by fake news affects democratic processes in the following ways:

1. Citizens can come to reject particular democratically made decisions if they think these decisions have been motivated by, or justified on the basis of, false information.
2. Fake news can exacerbate and deepen existing political polarization and public distrust in democratic institutions, thereby promoting system compromise and gridlock.
3. Fake news can negatively affect the moral justification of democratic in-

stitutions if this justification is taken to rest on democracy's tendency to produce epistemically better outcomes than other types of government.

4. In the worst case, fake news can contribute to complete system collapse if political divisions fanned by fear of manipulation and epistemic uncertainty undermine support for democratic systems that are no longer able to contain civil distrust and strife.

Each of these likely effects of fake news on the perceived justification of democratic institutions works in similar ways but has more or less severe consequences for democratic legitimacy.

4.1. Fake News Prompting Resistance to Democratically Made Decisions

If it is known that intentional falsehoods are widely circulating and it is feared that factually incorrect beliefs have informed major democratic decisions, citizens on the losing side of those decisions can become convinced that they have reasons not to follow the results. Knowledge of widespread circulation of fake news stories could then be seen as a "countervailing consideration" against the legitimacy of a democratically made decision.[45] Citizens on the losing side might support or engage in passive or nonviolent civil disobedience.

To take an analogous example, a 2018 Berkeley IGS poll found that in California a majority of voters supported their state's decision to provide sanctuary to undocumented immigrants, thereby resisting stricter immigration policy imposed by the Trump administration.[46] The majority of Californian voters did not vote for Donald Trump in the presidential elections in 2016 and 2020, and polls consistently showed that a majority also rejected Trump's claims that immigrants are criminals and a burden to society. Rather, most Californians considered immigrants as strengthening their country.[47] Trump's negative comments on immigrants were not a case of fake news but an example of a disagreement about basic facts. However, this example indicates how political decisions that are perceived to be based on contested claims can lead to public opposition. If fake news stories rise to prominence and are perceived to influence public decisions, it is therefore plausible to assume that they, too, have the potential to generate strong resistance.

45 Christiano, *The Constitution of Equality*, 262.

46 DiCamillo, "While the Statewide Law Providing Sanctuary to Undocumented Immigrants Is Supported by a Majority of California Voters, the Issue Is Highly Divisive."

47 "California Survey on Othering and Belonging: Views on Identity, Race and Politics," Othering and Belonging Institute, April 18, 2018, https://belonging.berkeley.edu/california-survey-othering-and-belonging.

4.2. Fake News Contributing to System Compromise and Gridlock

Citizens' awareness of widespread fake news can also affect their views and acceptance of part or all of the democratic system because fake news contributes to an information environment that can promote political polarization and distrust in democratic institutions. Political polarization began to increase decades ago.[48] Moreover, the recent loss of trust in democratic institutions in Western democracies was more likely caused by the 2008 global financial crisis than by the emergence of Twitter and Facebook.[49] However, in a political climate in which fake news significantly contributes to citizens' distrust of each other, it is plausible to assume that fake news is likely to exacerbate existing political polarization and the loss of trust in democratic institutions. In this way, these falsehoods make it more difficult to overcome political division and decreased support for public institutions.[50]

There are two likely responses of citizens feeling insecure about an information environment characterized by an abundance of information and widespread fear of manipulation. Neither of these responses is predicated on citizens actually believing the content of fake news. Such citizens might become politically apathetic. They might retreat from public debates requiring a degree of certainty about facts, given that they do not have the time, expertise, or trusted resources to acquire the required knowledge.[51] This kind of disengagement is likely to be accompanied by a loss of trust in the democratic process. Or these citizens might instead try to reduce the complexity of the problematic information environment by using certain heuristics, such as sticking to the political group that reflects their identity—even if they do not believe all of that group's claims or condone its entire agenda.[52] This prevents political dialogue and entrenches po-

48 Boxell, Gentzkow, and Shapiro, "Is the Internet Causing Political Polarization?"

49 Bennett and Livingston, "The Disinformation Order," 127.

50 Mitchell et al., *Many Americans Say Made-Up News Is a Critical Problem that Needs to be Fixed*. A survey found that in the United States, trust in the federal government has steadily declined over the past fifty years, and in 2019 only 17 percent of Americans said they can trust the government ("Public Trust in Government: 1958–2019," Pew Research Center, April 11, 2019, https://www.pewresearch.org/politics/2019/04/11/public-trust-in-government-1958-2019/). And in a survey conducted in October 2019 by the US Associated Press and NORC at the University of Chicago, 60 percent of respondents agreed that "political division in the United States [is the] result of Americans having different beliefs about how to address major problems facing the country" (Associated Press-NORC Center for Public Affairs Research, "State of the Facts 2019," survey data, USAFacts, November 13, 2019, https://static.usafacts.org/public/resources/2019_topline_final.pdf, 9).

51 Beckett and Livingstone, *Tackling the Information Crisis*.

52 Gottfried et al., *Trusting the News Media in the Trump Era*; Talisse, *Overdoing Democracy*;

litical divisions even if fake news by itself does not cause the divisions. These divisions, in turn, make it more difficult for different political factions to compromise and work together to solve problems.[53] Fake news then contributes to citizens adopting the confrontational stance that Brennan diagnoses in democratic politics. Thus, fake news can play a significant part in compromising the operations of democratic states, leading to political gridlock.

4.3. Fake News's Impact on the Normative Legitimacy of Democratic Institutions

So far, we have seen that fake news can undermine the epistemic trust of citizens in each other and the democratic process. From this *sociological* perspective, fake news is thus a problem for the stability of democratic institutions. With this danger in mind, we can also anticipate how fake news might undermine the moral justification (or *normative* legitimacy) of democratic institutions.[54] The difference between the sociological and the normative perspectives of legitimacy is that if a political authority possesses normative legitimacy, it is morally justified in wielding power—irrespectively of the perceptions and opinions of its subjects. There are a number of theories about what morally justifies democratic power. One is the epistemic theory of democracy, which itself comes in different versions. All of these epistemic views hold that democratic authorities are instrumentally justified because they produce better outcomes than other forms of government. Epistemic views of democracy are particularly relevant for our purposes because fake news threatens to undermine precisely the epistemic trust among citizens that is practically essential for producing the epistemically better outcomes that (according to these kinds of democratic theories) justify democratic authority.

As Hélène Landemore explains, for some advocates of the epistemic view, democracies generate better outcomes than other forms of government "because including more people in the decision-making process naturally tends to increase what has been shown to be a key ingredient of collective intelligence in the contexts of both problem solving and prediction—namely, cognitive diversity."[55] The larger the number of people, with their own perspectives and knowledge, that join together to make collective decisions, the better the results. But better outcomes are not achieved by simply adding up the views of individuals. Rather, the idea here is that democratic citizens deliberate together to arrive at better conclusions. Deliberating together has a number of advantages:

Mason, "Ideologues without Issues."

53 McCoy, Rahman, and Somer, "Polarization and the Global Crisis of Democracy," 24.

54 I thank one of the anonymous referees of the journal for encouraging me to clarify this point.

55 Landemore, *Democratic Reason*, 3.

it increases the pool of information and ideas all citizens can draw on; everyone can offer their interpretation of facts and reasons, which can help others better understand the matter at hand; and reasoning together helps to overcome individual biases in favor of better arguments, which in turn promotes better collective decisions.[56]

According to a different version of the epistemic defense of democratic authority, democracy's unique "wisdom of crowds" is based on the idea that under certain conditions, the larger the group involved in making a majoritarian decision democratically, the more likely this decision is to be the correct one. According to the Condorcet Jury Theorem, this is the case if the choice involved is binary, if there is a correct answer available, and, crucially, if each voter on average has a better than even chance to select the correct answer.[57] In such a situation, the larger the number of voters involved, the more likely it is that the majority decision will be the right choice.[58]

However, by undermining epistemic trust of citizens in each other, fake news can be expected to threaten the processes at the heart of epistemic accounts of democracy. If citizens believe that their counterparts are misinformed by fake news, it is plausible to presume that they will question their opponents' epistemic competence and doubt the meaningfulness of entering into a dialogue with them. As suggested above, this lack of trust might manifest in the form of political polarization or disengagement from the political process. If this is the case, there are a number of possible ways in which the epistemic quality of the outcomes of democratic decisions might be lessened, which in turn would negatively affect the decisions' normative legitimacy, on epistemic views of democracy. I can only hint here at the empirical connections between a loss of epistemic trust caused by fake news and a loss of epistemically grounded normative legitimacy. For these illustrative purposes, I limit myself to the two aforementioned versions of the epistemic defense of democracy.

First, as Landemore argues, on some versions of the epistemic theory, democratic decisions are better because they are informed by a diversity of viewpoints. On this picture, if fake news deters citizens from deliberating together, the diversity of perspectives involved in the decision-making process will be reduced. This will, in turn, render the resulting decisions less well informed and thus worse, which weakens the argument that democratic power is justified because it leads to better outcomes. Second, if we assume that the Condorcet Jury Theo-

56 See Landemore, *Democratic Reason*, 99.

57 For additional conditions of the Condorcet Jury Theorem, see Landemore, *Democratic Reason*, 148.

58 There are other versions of the epistemic view that I will not go into here for reasons of space.

rem provides a sound normative defense of democracy, citizens' disengagement from the political process due to epistemic distrust caused by fake news weakens the epistemic defense because it lowers the number of voters involved in the decision. The more citizens are deterred from voting due to fake news, the less likely the majority is to arrive at the right decision. And the less likely the majority is to arrive at the correct decision, the weaker this version of the epistemic defense of democracy becomes.

Despite the limited nature of these explanations, we can identify a crucial problem for epistemic theories of democracy when fake news leads to a loss of citizens' epistemic trust in each other. Since these views rely on citizens' participation in democratic processes, fake news threatens the moral justification of democratic institutions and decisions insofar as it undermines or changes precisely the political participation that is supposed to deliver better outcomes. This suggests that fake news not only threatens the sociological legitimacy (i.e., stability) but also the normative legitimacy (i.e., moral justification) of democratic authorities.

4.4. Fake News Contributing to System Collapse

Returning to the sociological legitimacy issue, we have seen that fake news is a threat to the stability of democratic institutions. The most extreme case of the pernicious influence of fake news would be one in which political divisions—stoked by informational uncertainty and fear of false information—lead to a critical loss of support for the democratic state overall. Citizens would then no longer see their state as morally justified and instead believe that it has lost its "right to issue and enforce laws without interference."[59] The ensuing active and violent resistance might take the form of intimidation of or attacks on political candidates, officials, and the supporters of opposing parties. At this point, though, the viability of democratic processes (i.e., the election of officeholders) and democratic values (i.e., acceptance of the moral justification of collectively made decisions) is truly in jeopardy, and political dissent might turn into open rebellion.

No democratic state has yet collapsed because of online fake news. Since fake news is not the sole source of false beliefs and political divisions, it is also unlikely ever to be the sole cause of democratic system collapse. Instead, financial and economic crises have frequently been conditions contributing to democratic system collapse in recent history.[60] However, the end of the German Weimar Republic and the role that mass media played in its end are instructive. The Weimar Republic did not collapse due to a public revolt but instead was dismantled

59 Adams, "Institutional Legitimacy," 98.

60 Funke, Schularick, and Trebesch, "Going to Extremes."

by the Nazis that were democratically elected as the largest party in the November 1932 elections. False or misleading political information significantly shaped the political context of the last years of the republic. As Bernhard Fulda argues, propaganda, a fragmented press, and dramatized reporting mixing information and entertainment "contributed significantly to the polarization of Weimar society and the escalation of political conflict."[61] In the end, public distrust in the state and among various factions was so pervasive that when the Nazis began to dismantle the Weimar democracy, there was no crucial mass of citizens left to defend it. The case of the Weimar Republic thus demonstrates that manipulative and false information can significantly contribute to the downfall of democracies. As Jason Stanley points out, the destruction of citizens' trust in each other and the state is a crucial aim of fascist forces aiming to eliminate democratic rule:

> Spreading general suspicion and doubt undermines the bonds of mutual respect between fellow citizens, leaving them with deep wells of mistrust not just toward institutions but also toward one another. Fascist politics seeks to destroy the relations of mutual respect between citizens that are the foundation of a healthy liberal democracy, replacing them ultimately with trust in one figure alone, the leader.[62]

Today, online fake news offers a new possibility for (external and internal) enemies of democracy to undermine the public's confidence in each other and in their democratic institutions. Crucially, as in the two other cases, the loss of perceived legitimacy of democratic values and processes can be brought about merely by the fear of others being manipulated by fake news, without citizens believing the content of fake news itself.

5. OBLIGATIONS TO TACKLE FAKE NEWS

Nonetheless, democracy remains morally important because if the conditions are right, it can produce better outcomes than other forms of government and it can be a unique way of respecting everyone's moral equality in the political decision-making process. This explains why democratic institutions have not only prudential reasons but, more importantly, also moral obligations to fight the spread of fake news and its deleterious effects on democratic stability. After all, we ought to make better rather than worse political decisions, and we ought to respect the genuine views and interests of others in our collective decision-making process.

61 Fulda, *Press and Politics in the Weimar Republic*, 223.

62 Stanley, *How Fascism Works*, 71.

What exactly then are the particular obligations for combating fake news that arise for democratic institutions? Several plausible suggestions have been made. Given that fake news today is mostly spread via social media, Regina Rini has suggested that social media companies ought to be incentivized through regulation to be transparent about who pays them to distribute information on their platforms.[63] Rini also proposes that social media companies introduce reputation scores for individual users as a mechanism for attaching social costs to the spreading of fake news.[64] In light of fake news's negative impact on the general information environment in which democratic citizens find themselves (see section 3.1 above), my own focus here will be on how generally to strengthen the information environment. More concretely, this implies at least three urgent obligations for democratic institutions.

First, as we saw, an important problem of the information environment of democratic societies in the digital age is the decreasing number of common information sources. Everyone can create and disseminate information online. Democratic institutions thus have a duty to offer a publicly funded news source that is politically independent and operates fully transparently. Public funding can be provided, for example, through a compulsory license fee paid by all eligible members of the public. Independence can be guaranteed by a supervisory board that is sufficiently independent from government influence and that consists proportionally of representatives of all major groups of society. Promoting public awareness of the independence and neutrality of such public broadcasters can motivate significant public trust in these news sources. For instance, in the United Kingdom, the BBC is the most widely trusted source of news, and in Germany 80 percent of the population trusts public broadcasters.[65]

Second, to combat fake news, democratic institutions ought to promote education and digital media literacy in order to empower their citizens to distinguish trustworthy from untrustworthy news.[66] These efforts should also be publicized to foster epistemic trust among citizens. Initiatives of this kind have been shown to be successful in various countries.[67] As a corollary, states might

63 Rini, "Social Media Disinformation and the Security Threat to Democratic Legitimacy," 13.

64 Rini, "Fake News and Partisan Epistemology," 57.

65 Daniel Marshall, "BBC Most Trusted News Source 2020," Ipsos MORI, May 22, 2020, https://www.ipsos.com/ipsos-mori/en-uk/bbc-most-trusted-news-source-2020; "News Media and Public Attitudes in Germany," Pew Research Center, May 17, 2018, https://www.pewresearch.org/global/fact-sheet/news-media-and-political-attitudes-in-germany/.

66 Rini, "Social Media Disinformation and the Security Threat to Democratic Legitimacy," 13.

67 Guess et al., "A Digital Media Literacy Intervention Increases Discernment between Mainstream and False News in the United States and India."

also consider, as part of their public education, disincentivizing a focus on fake news, as some have suggested, because giving prominence to fake news might increase unfounded concerns about it among those being educated.[68] The duty to increase the digital media literacy and general education of citizens is thus in stark contrast to the conclusions that Brennan draws from the dangers of partisanship and polarization in democratic politics.[69] Insofar as the problem is citizens' ignorance of accurate information, democratic institutions, rather than disenfranchising part of the population, should promote citizens' information literacy and strengthen their information environment. Public education and public broadcasts are public goods. The fact that public goods promoting the functioning of democracy may be underfunded or nonexistent does not show that democracy cannot work. It also does not show that we should give up on democracy and adopt epistocracy. It rather shows that we should properly fund these public goods that both empower individuals and improve the information environment in democratic societies.

Finally, public officeholders must not use the label "fake news" to discredit their opponents or, in particular, the free press. The most famous example is Donald Trump, who is sometimes credited with popularizing the term "fake news" itself. The labeling of certain press outlets as fake is particularly damaging because, as explained above, fake news masquerades as real news. If citizens believe that others accept the labeling of regular news sources as fake, their view that their fellow citizens are unable to distinguish real news from fake news is exacerbated. In such a situation, the democratic system itself (through its representatives) fans the flames of epistemic distrust and confusion. For this to happen, it is sufficient that citizens are convinced that others believe the accusations that certain information is false.[70] The use of the "fake news" label by public officials is thus a direct attack on the sociological and normative legitimacy of democratic institutions and is correspondingly insidious. Instead, public officials ought to work toward fulfilling their obligations to safeguard and promote epistemic trust among democratic citizens.

6. CONCLUSION

Major democratic institutions have correctly identified fake news as a threat to

68 Habgood-Coote, "Stop Talking about Fake News!"

69 Brennan, *Against Democracy*.

70 The view that holds that fake news poses a threat because its content is actually believed has difficulties accounting for this problem. I thank an anonymous reviewer of the *Journal of Ethics and Social Philosophy* for stressing this point.

their values and processes. However, the danger posed by these online falsehoods does not primarily lie in their power to convince readers of the veracity of their factually incorrect content. Rather, the primary danger fake news poses to democratic values and institutions lies in the corrosive effect it has on trust among citizens and thus on citizens' trust in their democracy.

It would be unreasonable to expect that the magnitude of fake news's threat to democracies can be quantified with any precision because fake news is unlikely to be the sole cause of civil resistance to particular political decisions, of system gridlock, of arrival at suboptimal decisions via democratic processes, or of system collapse. Political conflicts and polarization that put pressure on the perceived legitimacy of democratic institutions are not the effect of, but predate, fake news and create the circumstances in which fake news can thrive. However, increasing political polarization, political gridlock, and a growing lack of trust in democratic institutions are well-documented trends in liberal democracies. Moreover, reputable polling evidence shows that fake news leads to a loss of citizens' trust in each other, which is a major cause of the destabilization of democratic processes and of the erosion of the benefits that morally justify democratic institutions.

Online fake news is of course not the sole source of false beliefs (and thus of public concern about false beliefs) in democratic societies. However, given that in Western liberal democracies most voters today obtain political information via the internet, online fake news has become a major threat to epistemic trust among cocitizens.[71] Democratic institutions are therefore rightly worried about the spread of these falsehoods. Democratic processes are only viable when citizens have a sufficient degree of epistemic trust in their main sources of political information and in each other's epistemic competence. Fake news is a major threat to both of these conditions even when it is not straightforwardly believed. Primarily for this reason, fake news presents a threat to democratic processes and values and is rightly a matter of concern for democratic institutions.[72]

University of Birmingham
m.reglitz@bham.ac.uk

71 Shearer and Matsa, *News Use across Social Media Platforms 2018*.

72 I am grateful to Heather Widdows, Wouter Peeters, Jonathan Parry, Regina Rini, Kristin Reglitz, Art Held, and Scott Lucas for their useful comments. The paper has also benefited from feedback at the 9th Braga Meetings on Ethics and Political Philosophy. Finally, I thank the anonymous reviewers of the *Journal of Ethics and Social Philosophy* for their comments, which have helped improve this paper.

REFERENCES

Adams, Nathan. "Institutional Legitimacy." *Journal of Political Philosophy* 26, no. 1 (March 2018): 84–102.

Allcott, Hunt, and Matthew Gentzkow. "Social Media and Fake News in the 2016 Election." *Journal of Economic Perspectives* 31, no. 2 (Spring 2017): 211–36.

Barthel, Michael, Amy Mitchell, and Jesse Holcomb. *Many Americans Believe Fake News Is Sowing Confusion*. Pew Research Center, December 15, 1016. https://www.pewresearch.org/journalism/2016/12/15/many-americans-believe-fake-news-is-sowing-confusion/.

Beckett, Charlie, and Sonia Livingstone. *Tackling the Information Crisis: A Policy Framework for Media System Resilience*. London School of Economics and Political Science, 2018. https://www.lse.ac.uk/law/news/2018/truth-trust-technology.

Bennett, Lance W., and Steven Livingston. "The Disinformation Order: Disruptive Communication and the Decline of Democratic Institutions." *European Journal of Communication* 33, no. 2 (April 2018): 122–39.

Boxell, Levi, Matthew Gentzkow, and Jesse M. Shapiro. "Is the Internet Causing Political Polarization? Evidence from Demographics." Working Paper 23258. National Bureau of Economic Research, March 2017.

Brennan, Jason. *Against Democracy*. Princeton, NJ: Princeton University Press, 2016.

Christiano, Thomas. *The Constitution of Equality: Democratic Authority and Its Limits*. Oxford: Oxford University Press, 2008.

———. "Democracy, Participation, and Information: Complementarity between Political and Economic Institutions." *San Diego Law Review* 56, no. 4 (December 2019): 935–60.

———. "Self-Determination and the Human Right to Democracy." In *Philosophical Foundations of Human Rights*, edited by Rowan Cruft, Matthew Liao, and Massimo Renzo, 459–80. Oxford: Oxford University Press, 2015.

Davidson, W. Phillips. "The Third-Person Effect in Communication." *Public Opinion Quarterly* 47, no. 1 (Spring 1983): 1–15.

DiCamillo, Mark. "While the Statewide Law Providing Sanctuary to Undocumented Immigrants Is Supported by a Majority of California Voters, the Issue Is Highly Divisive." Institute of Governmental Studies of the University of California, Berkeley, April 27, 2018. https://escholarship.org/uc/item/443790rm.

Digital, Culture, Media, and Sport Committee. *Disinformation and "Fake News": Interim Report*. United Kingdom House of Commons 2017–2019, 363. Lon-

don, 2018. https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/363/363.pdf.

Easley, Jonathan. "Poll: Majority Says Mainstream Media Publishes Fake News." *The Hill*, May 24, 2017. https://thehill.com/homenews/campaign/334897-poll-majority-says-mainstream-media-publishes-fake-news.

European Commission. *Fake News and Disinformation Online*. Flash Eurobarometer 464 Report. European Union, 2018. https://europa.eu/eurobarometer/surveys/detail/2183.

———. *A Multi-Dimensional Approach to Disinformation*. Report of the Independent High-Level Group on Fake News and Online Disinformation. Luxembourg: Publications Office of the European Union, 2018. https://op.europa.eu/en/publication-detail/-/publication/6ef4df8b-4cea-11e8-be1d-01aa75ed71a1/language-en.

Fallis, Don, and Kay Mathiesen. "Fake News Is Counterfeit News." *Inquiry* (forthcoming). Published ahead of print, November 6, 2019. https://doi.org/10.1080/0020174X.2019.1688179.

Fuerstein, Michael. "Epistemic Trust and Liberal Justification." *Journal of Political Philosophy* 21, no. 2 (June 2013): 179–99.

Fulda, Bernhard. *Press and Politics in the Weimar Republic*. Oxford: Oxford University Press, 2009.

Funke, Manuel, Moritz Schularick, and Christoph Trebesch. "Going to Extremes: Politics after Financial Crises, 1870–2014." *European Economic Review* 88 (September 2016): 227–60.

Gaus, Gerald F. *Justificatory Liberalism: An Essay on Epistemology and Political Theory*. Oxford: Oxford University Press, 1996.

Gottfried, Jeffrey, Galen Stocking, Elisabeth Grieco, Mason Walker, Maya Khuzam, and Amy Mitchell. *Trusting the News Media in the Trump Era*. Pew Research Center, December 12, 2019. https://www.pewresearch.org/journalism/2019/12/12/trusting-the-news-media-in-the-trump-era/.

Grinberg, Nir, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. "Fake News on Twitter during the 2016 US Presidential Election." *Science* 363, no. 6425 (January 2019): 374–78.

Guess, Andrew, Jonathan Nagler, and Joshua Tucker. "Less Than You Think: Prevalence and Predictors of Fake News Dissemination on Facebook." *Science Advances* 5, no. 1 (January 2019): 1–8.

Guess, Andrew M., Michael Lerner, Benjamin Lyons, Jacob M. Montgomery, Brendan Nyhan, Jason Reifler, and Neelanjan Sircar. "A Digital Media Literacy Intervention Increases Discernment between Mainstream and False

News in the United States and India." *Proceedings of the National Academy of Science of the United States of America* 117, no. 27 (July 2020): 15536–45.

Habgood-Coote, Joshua. "Stop Talking about Fake News!" *Inquiry* 62, no. 9–10 (2019): 1033–65.

Jang, S. Mo, and Joon K. Kim. "Third Person Effects of Fake News: Fake News Regulation and Media Literacy Interventions." *Computers in Human Behaviour* 80 (March 2018): 295–302.

Landemore, Hélène. *Democratic Reason: Politics, Collective Intelligence, and the Rule of the Many*. Princeton, NJ: Princeton University Press, 2013.

Lazer, David M. J., Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, et al. "The Science of Fake News: Addressing Fake News Requires a Multidisciplinary Effort." *Science* 359, no. 6380 (March 2018): 1094–96.

Marchal, Nahema, Bence Kollanyi, Lisa-Maria Neudert, and Philip N. Howard. *Junk News during the EU Parliamentary Elections: Lessons from a Seven-Language Study of Twitter and Facebook*. Data Memo 2019.3. Oxford Internet Institute, 2019. https://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2019/05/EU-Data-Memo.pdf.

Mason, Lilliana. "Ideologues without Issues: The Polarizing Consequences of Ideological Identities." *Political Opinion Quarterly* 82, no. S1 (2018): 866–87.

McCoy, Jennifer, Tahmina Rahman, and Mura Somer. "Polarization and the Global Crisis of Democracy: Common Patterns, Dynamics, and Pernicious Consequences for Democratic Polities." *American Behavioral Scientist* 62, no. 1 (January 2018): 16–42.

McKernon, Edward. "Fake News and the Public." *Harper's*, October 1925.

Milanovic, Branko. *Global Inequality. A New Approach for the Age of Globalization*. Cambridge, MA: Harvard University Press, 2016.

Mitchell, Amy, Jeffrey Gottfried, Galen Stocking, Mason Walker, and Sophia Fedeli. *Many Americans Say Made-Up News Is a Critical Problem that Needs to Be Fixed*. Pew Research Center, June 5, 2019. https://www.pewresearch.org/journalism/2019/06/05/many-americans-say-made-up-news-is-a-critical-problem-that-needs-to-be-fixed/.

Pew Research Center. *The Public, the Political System and American Democracy*. Pew Research Center, April 26, 2018. https://www.pewresearch.org/politics/2018/04/26/the-public-the-political-system-and-american-democracy/.

PricewaterhouseCoopers. *"Fake News": Ergebnisse eine Bevölkerungsbefragung*. PwC Communications, 2019. https://www.pwc.de/de/technologie-medien

-und-telekommunikation/PwC%20190420%20Berichtsband%20Fake%20News.pdf.

Rini, Regina. "Fake News and Partisan Epistemology." *Kennedy Institute of Ethics Journal* 27, no. 2 (June 2017): 43–64.

———. "Social Media Disinformation and the Security Threat to Democratic Legitimacy." In *Disinformation and Digital Democracies in the 21st Century*, edited by Joseph McQuade, Tiffany Kwok, and James Cho, 10–14. Toronto: NATO Association of Canada, 2019.

Roozenbeek, Jon, and Sander van der Linden. "Fake News Game Confers Psychological Resistance against Online Misinformation." *Palgrave Communications* 5, no. 65 (2019): 1–10.

Shearer, Elisa, and Katerina Eva Matsa. *News Use across Social Media Platforms 2018*. Pew Research Center, September 10, 2018. https://www.pewresearch.org/journalism/2018/09/10/news-use-across-social-media-platforms-2018/.

Silvermain, Craig. "This Analysis Shows How Viral Fake Election News Stories Outperformed Real News on Facebook." BuzzFeed News, November 16, 2016. https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook.

Soll, Jacob. "The Long and Brutal History of Fake News." *Politico*, December 19, 2016. https://www.politico.eu/article/fake-news-elections-trump-media/.

Der Spiegel. "Is There Truth to the Refugee Rape Reports?" January 17, 2018. https://www.spiegel.de/international/germany/is-there-truth-to-refugee-sex-offense-reports-a-1186734.html.

Stanley, Jason. *How Fascism Works: The Politics of Us and Them*. New York: Random House, 2018.

Sunstein, Cass. *#Republic: Divided Democracy in the Age of Social Media*. Princeton, NJ: Princeton University Press, 2017.

Talisse, Robert. *Democracy and Moral Conflict*. Oxford: Oxford University Press, 2009.

———. *Overdoing Democracy: Why We Must Put Politics in Its Place*. Oxford: Oxford University Press, 2019.

Zuboff, Shoshana. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. London: Profile Books, 2019.

Journal of Ethics and Social Philosophy
VOL. 22, NO. 2 · JULY 2022
https://doi.org/10.26556/jesp.v22i2.1578

DISCURSIVE INTEGRITY AND THE PRINCIPLES OF RESPONSIBLE PUBLIC DEBATE

Matthew Chrisman

POLITICAL COMMENTATORS often lament the "brokenness" of political discourse. Rather than constructive and thoughtful contributions to debate about how we should live together in a community of equals who sometimes disagree, contemporary political discourse is said to involve a lot of speaking out of both sides of one's mouth, pandering to the base, identity posturing, gaslighting, bullshitting, flipflopping, dog whistling, mudslinging, name calling, and outright lying. Many of these accusations trade on the impression that people speaking as part of public political discourse are not engaging with their interlocutors in a sincere way but rather trying to manipulate other people to advance some covert agenda. This is widely thought to be true of politicians who are often regarded as untrustworthy liars, willing to say whatever they have to in order to advance their political aims. But I suspect many people have this impression of at least some political commentators in the media and even disagreeable fellow citizens encountered in the new public square of social media.

Celebrating honesty, authenticity, and sincerity in public political life is a natural response. We see evidence of this response in the way some politicians have recently become successful through eschewing political correctness and appearing unafraid to "say it like [they think] it is" regardless of potential offense. There is also an increasing tendency for political figures to lean into crude and unsophisticated ways of speaking or slips in judgment in their private lives, seeking to own these as part and parcel of their authentic engagement with public life. Moreover, the recent philosophical literature contains several prominent defenses of sincerity and honesty as a very important interpersonal moral value at the heart of our public and private lives.[1]

1 For instance, Shiffrin writes, "Reliable, sincere speech enables sophisticated forms of self-understanding, knowledge of others and of the world, moral agency, and personal relations of trust. The relation between communication and these foundational compulsory ends explains the strong presumption of sincere communication as well as our responsibility to strive for accuracy" (*Speech Matters*, 186). See also Hawley's argument that assertion

I doubt, however, that failures of honesty, authenticity, and sincerity are the main obstacles to constructive political debate in contemporary democracies. As a first step to explaining this doubt, I want to argue here that, although expressive sincerity is valuable, we should not ignore discursive integrity in thinking about how to address problems with contemporary political debate.[2] The first task for this paper then is to explain the difference, as I see it, between expressive sincerity and discursive integrity and to argue that they are two importantly different communicative ideals that we risk running together in thinking about what makes public political debate go better or worse. As an initial gloss on the distinction, expressive sincerity is about whether someone really thinks what they say, whereas discursive integrity is about whether someone takes responsibility for what they say in a way that warrants reliance on it.[3] These are related because liars and hypocrites do not typically take responsibility for what they say, so we usually should not rely on them. However, as I am thinking of it, discursive integrity is different from expressive sincerity. It is in large part about whether the claims someone makes as part of public political debate manifest recognition of the attending responsibility to back up, justify, motivate, or modulate what they say in the face of unconvinced audiences.

The second task for this paper is to explain why discursive integrity is important for public political debate. Once one sees how it is different from expressive sincerity, it will be pretty obvious that it matters in interpersonal communica-

---

should be understood in terms of promising to speak truthfully (*How to Be Trustworthy*, ch. 3); and Miller's argument that honesty is a widely valued important virtue but understudied by philosophers ("Honesty"). To be clear, these philosophers are concerned with a much broader phenomenon than specifically public political debate, which is my concern here, and none of them claim that sincerity and honesty are the only important communicative virtues. Indeed, Hawley argues that trustworthiness requires sincerity *and* competence in getting things right, which can be related to the ideal of discursive integrity I will discuss. Williams offers a sustained discussion and defense of the value of sincerity, which he conceives in much richer terms than I deploy here, though saying what one thinks is a core element of his conception (*Truth and Truthfulness*). In connection to public political discourse, Williams also argues that stressing the value of sincerity (in the sense of non-deceitfulness) for government officials is crucial to the ability of liberal democracies to overcome the tyranny of corrupt governments (*Truth and Truthfulness*, ch. 9).

2 Markovits argues that obsession with sincerity was also a feature of ancient Greek politics, and argues on some similar grounds to me that this is not the most important communicative value for a thriving democracy (*The Politics of Sincerity*).

3 There are important conceptual distinctions between honesty and sincerity that I am here suppressing, and the precise analysis of both is a controversial topic that I am not going to address here, as I am more interested in the contrast between this nexus of concepts and discursive integrity. For useful discussion, see Eriksson, "Straight Talk"; and Ozar, "Sincerity, Honesty, and Communicative Truthfulness."

tion. However, I intend the theoretical articulation of the distinction I develop below to illuminate some more specific ways in which discursive integrity is not just a nice trait to manifest when we talk to each other but actually crucial for most constructive debate among political equals who disagree. To make this case, I shall develop three arguments from within the resources of political philosophy and epistemology, for thinking that advances toward an ideal of discursive integrity is an important step for repairing public debate in contemporary democracies. The thrust of these arguments is that our concern in contemporary political culture should not be entirely or even mainly with whether political actors are dishonest, insincere, or hypocritical; we should also be concerned with a different question: whether they participate in public political debate in a way that demonstrates an ability and willingness to back up, justify, motivate, or modulate what they say in the face of disagreement.

The final task of this paper is to consider a practical strategy for better incorporating the ideal of discursive integrity into contemporary public political debate. Appealing to a recent example of what I see as a good case of public political discussion of controversial issues, I will argue that public political discourse should include not only debate about first-order issues but also second-order debate about the "rules of the game" in an attempt to coordinate on a shared understanding of which laws and public policies are legitimate but also of how we should talk about controversial issues of mutual interest in our political community. It is an empirical question outside the scope of this paper how effective this strategy will be, but I do want to explain how the potential of this practical suggestion becomes easier to appreciate in light of my articulation of the theoretical difference between expressive sincerity and discursive integrity.

1. TWO MODELS OF COMMUNICATION

In this section, I distinguish two lines of thought about linguistic communication and meaning.[4] I do not intend to defend either view but rather to use the distinction as a lens through which we can view practices around public political debate, suggesting different aspects to focus on in our attempts to understand what makes debate among political equals who disagree go better or worse.

The first conception of communication traces back at least to Locke, who wrote that "words in their primary or immediate signification signify nothing

4 I discuss this distinction much more in the context of metaethical debates about the meaning of normative claims and the metasemantic interpretation of formal semantic models of compositionality in Chrisman, *The Meaning of "Ought,"* ch. 6. See also Chrisman, "Two Nondescriptivist Views of Normative and Evaluative Statements."

but the ideas in the mind of him that uses them" capturing a natural intuition about what is going on when someone says something meaningful.[5] The basic claim is that the core function of meaningful speech is to express what is antecedently in one's mind as part of communicating with others. If this is right, it is natural to think the meaningfulness of words and sentences should be understood primarily in terms of their ability to serve this function, and the communicative norms governing speech should be understood fundamentally in terms of how individuals express what is antecedently in their minds to others. Or, put slightly differently, we can say that Lockeans in the philosophy of language place a lot of theoretical weight on the idea that a speaker intentionally expresses thoughts, ideas, and attitudes to an audience by articulating these mental states in words whose meaning derives from their usability to get an audience to know what thoughts, ideas, and attitudes are in the speaker's mind.[6]

The second conception of communication I want to highlight traces back at least to Peirce, who wrote, "An act of assertion supposes that, a proposition being formulated, a person performs an act which renders him liable to the penalties of the social law (or, at any rate, those of the moral law) in case it should not be true, unless he has a definite and sufficient excuse."[7] The basic idea, as I understand it in this context, is that the norms governing language use should be understood primarily in terms of collections of agents self-consciously performing mutually recognized linguistic moves in a social space of responsibilities and commitments, and the meaningfulness of words and sentences should be understood primarily in terms of their capacity to mediate such moves. There is controversy among Peirceans about what exactly one commits to in saying something, but the version of the view that is relevant for what follows is one that focuses on contexts where speakers are expected at least to some extent to try to justify or modulate what they have said in the face of disagreement.[8] Accordingly, on this Peircean view, communication is analyzed most fundamentally in terms of the

5 Locke, *An Essay Concerning Human Understanding*, III.2.2.

6 This gets picked up and modified in various important ways in the expression-based tradition in theorizing about meaning. See especially Grice, *Studies in the Way of Words*; Schiffer, *Meaning*; Bennett, *Linguistic Behaviour*; Davis, *Meaning, Expression and Thought*.

7 Hartshorne, Weiss, and Burks, *Collected Papers of Charles Sanders Peirce*, 2.315.

8 Although endorsing the Peircean picture, Hawley argues that the commitment to speak truthfully does not entail a commitment to justify what one says or to retract what one has said when unable to justify what one says (*How to Be Trustworthy*, 50–57). Her main target is Robert Brandom's idea about "taking the commitment involved in asserting to be the undertaking of justificatory responsibility for what is claimed" (Brandom, "Asserting," 641). For discussion of modulation and retraction, see also MacFarlane, "What Is Assertion?" I side with Brandom and MacFarlane here for the specific context of making claims in public

mutually recognized undertaking of a specific kind of commitment rather than in terms of the expression of an antecedently held mental state in the mind of the speaker.[9]

In sum, the Lockean conception stresses the attitude-expressing function of communicative speech whereas the Peircean conception stresses its commitment-undertaking function. I suspect they both capture important parts of the phenomenon of communication, which is why I do not propose to argue that one or the other is correct.[10] Rather, I have extracted these two lines of thought from discussions in the philosophy of language so I can next explore the different sorts of communicative ideals that attach to them.

2. TWO COMMUNICATIVE IDEALS AND TRUSTWORTHINESS

Much of what we take ourselves to know comes from the word of others, but people do not always tell the truth. So how can the fact that a speaker says something justify the audience in believing what the speaker said. This is known as the problem of testimony in epistemology (where "testimony" is being used broadly to include ordinary conversations as well as making statements in formal settings such as law courts).[11]

More colloquially, we can put the question in terms of *trust*: When should we trust what other people say? And we can generalize this question of trust beyond

political debate but leave it open whether that is the correct Peircean view of assertion more generally.

9 This gets picked up and modified in various ways in the commitment-based tradition in theorizing about meaning. See especially Sellars, "Some Reflections on Language Games" and "Meaning as Functional Classification"; Searle, *Speech Acts*; Brandom, *Making It Explicit*; Alston, *Illocutionary Acts and Sentence Meaning*; Kukla and Lance, *"Yo!" and "Lo!"*

10 Lance argues that it is wrong to separate the attitude-expressing and commitment-undertaking aspects of speech ("Some Reflections on the Sport of Language"). His idea is that a *mere* "game" of giving and asking for reasons could lack the sort of psycho-physical connection with agents' dispositions to action that are a crucial part of why we talk with each other in the first place. He illustrates with a degenerate case, highly relevant here, of a politician who is good at justifying what he says and retracting things in the face of good objection but who lacks the integrity to act on what he says because he does not really believe it. This is related to the dual aspect of what Grice calls the "Maxim of Quality" for cooperative speech, which subsumes both not saying what one believes to be false and not saying that for which one lacks adequate evidence. See Grice, "Logic and Conversation."

11 Much of the debate in this area concerns whether testimonial knowledge can be reduced to other kinds of knowledge (e.g., perceptual and inferential knowledge) or whether testimony should be treated as a *sui generis* source of knowledge. For an overview, see Adler, "Epistemological Problems of Testimony." Below I discuss and cite some reductionist and nonreductionist views in more detail.

belief by noticing that many of the nonbelief attitudes we have stem in large part from adopting the attitudes of others because we trust them. This kind of reliance is a normal and pervasive feature of human communication.

Because of this reliance, it is very natural to prize trustworthiness in our interlocutors. This is of course true in private personal communication, but it is also true for my topic here: public political discourse. So, I now want to use the distinction between Lockean and Peircean accounts of communication and meaning to generate two importantly different articulations of the ideal of trustworthiness for the public political sphere. This will provide a *prima facie* case for distinguishing two communicative ideals in thinking about what is important for constructive political debate.

You probably see where this is going. Lockeans focus on the relation between what people say and what they think. This focus encourages investigating possible mismatches that would undercut reasons we otherwise have for accepting something because someone else says it. Although such reliance is a pervasive feature of human communication, there are cases where evidence suggests that the speaker does not really think what they say—evidence, that is, of *expressive insincerity*. Sometimes people do not believe what they say, sometimes they do not have the positive or negative attitude toward something that their words convey, and other times they do not really intend to do what they say they are going to do. I am counting all of these as cases of expressive insincerity. They undercut the normally good inference from *S said that p* to *I should accept that p*.[12] Accordingly, the Lockean ideal of trustworthiness turns on how individuals treat others in speaking—whether they really think what they purport to think. When they do not or we have enough inductive evidence to warrant doubt, this pervasive route for expanding one's own views by relying on the testimony of others breaks down.

Peirceans focus instead on the kinds of commitments that are mutually acknowledged in normal communication. As we saw above, their core idea is that a speaker undertakes responsibility for what they say, and I proposed to focus on communicative contexts where this responsibility includes a commitment to back up or modulate what one says in the face of disagreement. This is a key way that the speaker and others become entitled to take what was said as a premise in collective reasoning toward further views about what we collectively should think, feel, and do. This Peircean focus encourages investigating potential short

12 It may not be initially clear how this applies to the case of trusting people to do what they say they are going to do, but I think of that as a case where the speaker says (in effect) "I'm going to ϕ," but lacks the intention conventionally conveyed by these words, which means that the audience should not accept that the speaker is going to ϕ, despite what has been said.

circuits, where a speaker does not acknowledge or cannot meet their responsibility to back up or modulate what was said. This lack of *discursive integrity* is bad for communication because it undermines reasoning collectively through mutually acknowledged responsibilities and entitlements to shared views about what to think, feel, and do. Accordingly, the Peircean ideal of trustworthiness turns on how speakers relate to each other in taking up what is said in collective reasoning. When we think a speaker cannot be relied upon to live up to the discursive commitments carried by their words, downstream collective reasoning about what to think, feel, or do can break down.

So far, I have used the distinction between Lockean and Peircean models of communication to articulate two different ideals of trustworthiness in communication: expressive sincerity and discursive integrity. Speakers may often live up to these ideals to similar degrees, but they can come apart. Someone can be sincere in what they say without meriting trust that they are committed to back up what they say in the face of disagreement. For example, we all know people who concede, "Well, it's just my opinion," in response to any disagreement, or people who seem to meet any challenge to what they say with distracting ad hominem attacks. Likewise, although it is less common, someone can manifest discursive integrity without meriting trust that they are being sincere. For example, I certainly know philosophers who seem prepared to offer arguments ad nauseum for what they say but who comport themselves in a way that leaves me feeling unsure whether they really believe what they say.

Now let us apply this distinction to public political debate. When someone accuses a politician of being an untrustworthy liar, who speaks out of both sides of his mouth, saying whatever he has to say to get elected, often two importantly different ideals are being evoked. On the one hand, the concern may be with the politician's sincerity: he does not really think what he says he thinks. On the other hand, the concern may be with the politician's integrity: he is not really acknowledging responsibility for backing up or modulating what he is saying in the face of disagreement, at least not in a way that entitles himself and others to rely on it in further collective reasoning.[13] In most cases, the critic is probably thinking vaguely about a bit of both of these failings. But my point in marking

13 In a related vein, Richardson argues that an important problem with Donald Trump's public political speech is often not insincerity but rather the fact that he speaks in ways that undermine attempts to attribute specific assertoric content to what he says ("Noncognitivist Trumpism"). Richardson suggests that Trump often fails to put forward such content because of the way his political speech tends to foreground insults, engage in explicit but unexplained contradictions, and float polarizing suggestions approvingly but seemingly unseriously. In my view, these are failures of discursive integrity.

the distinction is that these are different concerns, and ameliorating suspicion of insincerity will not always address suspicion of lack of integrity, and vice versa.

Similarly, when we lament a polarized media landscape, worrying that certain prominent political commentators are mainly advancing covert agendas rather than engaging in good-faith discussion of matters of public intertest, often two importantly different ideals are implicitly evoked. On the one hand, the concern might be with whether we can trust the person in the Lockean sense of assuming that they really think what they say. On the other hand, the concern might be with whether we can trust the person in the Peircean sense of expecting that they will acknowledge and meet responsibilities for backing up what they say. As before, it is probably often a bit of both that we are vaguely worried about, but they are separable concerns, and strategies for mollifying or avoiding them are going to be different.

Finally, think about the all-too-regular occurrences on social media of someone gaslighting a person they disagree with, posting memes with offensive presuppositions, or spreading an ideologically convenient but ill-sourced account of how some newsworthy event happened. Why does behavior like this undermine constructive communication in the public sphere? One concern is that it is dishonest because its motivations are something other than conveying what the speaker really thinks. Another concern is that such communication is irresponsible because it does not involve undertaking commitments to justify or modulate what is said in the face of disagreement, and so any support provided for further views about what to collectively think, feel, and do is mostly arational and non-justificatory. In the case of social media, I suspect the latter concern is usually more pressing, but my point at this stage is only to highlight the differences between such concerns, where one is based on the ideal of expressive sincerity and the other is based on the ideal of discursive integrity.

My first aim in this paper has been to explain the difference between expressive sincerity and discursive integrity, conceived as two importantly different communicative ideals. There is a risk of running these together in thinking about why contemporary political debate is "broken." This is especially true where we take trustworthiness to be an important ideal in constructive political discussion. Hence, in this section, I have sought to explain how the Lockean and Peircean orientations toward communication yield two important but different conceptions of trustworthiness in public political discourse. In the next section, I turn to the second main aim of this paper, which is to consider some arguments for thinking that discursive integrity is crucial for constructive political debate, and that we will not repair the political debate of contemporary democracies by exclusively or mainly focusing on expressive sincerity.

3. DISCURSIVE INTEGRITY AND PUBLIC POLITICAL DELIBERATION

At the beginning of this paper, I noted that it is natural to call for greater honesty among political agents in response to the way contemporary political discourse seems broken, which motivates politicians to highlight their own authenticity and drives the media to focus on insincerity and hypocrisy. We are now in a position to appreciate how this response turns on a communicative ideal of expressive sincerity. It is about expecting people to communicate in ways that reveal what they really think. Expressive sincerity is important to all sorts of interpersonal communication, but I suspect it is most important in the sphere of personal communication with family and friends, where our dependence on others is deep, repeated, reciprocal, and multidimensional. In these relationships, it is important to know that we can regularly and easily learn what people think about things by listening to what they say, since we depend on this knowledge in iterated coordination of collective action, not to mention our mutual and ongoing emotional well-being and moral development.

It is tempting to extend this ideal to the public political context, wanting to view political agents as an extension of our friends and family. And there is certainly some place for the ideal in this context. However, as we saw above, there is another kind of trust based on the distinct communicative ideal of discursive integrity. Demanding increased sincerity from political agents is important, but when it comes to the sort of large-scale collective deliberation that politics ideally manifests, we want people not only to say what they think but also to take responsibility for backing up what they say and to tread carefully in reasoning together to further conclusions about what to think, feel, and do.[14] Of course, expressive sincerity facilitates discursive integrity, since one will not usually take responsibility for what one says unless one really believes it (see below for some potentially important exceptions). But these ideals are not identical, and there tends to be too much focus on sincerity over integrity in the political culture of contemporary democracies. So, next, I want to advance three more specific arguments for the importance of the ideal of discursive integrity in thinking about what facilitates constructive political debate.

14 Shiffrin argues that it is important for our mutual moral development in community that we sincerely express our thoughts (even when these are vicious or offensive) (*Speech Matters*, ch. 3,). She uses this idea to generate a moral argument for honesty and moral-political argument for robust protection of free speech. Although it may not initially seem to be, this argument is consistent with what I go on to argue here. As I see things, our moral development occurs primarily in our interpersonal and often private interactions with friends and family, whereas I am focused here on public interactions among political equals who disagree deeply but who intend to live together peacefully in a political community.

The first argument stems from debates about what makes legitimate the sort of majority rule characteristic of a democracy. Deliberative democrats provide a prominent case for thinking that decisions about public policy will be legitimate only under certain circumstances that give those in the minority reason to go along with the majority. The circumstances are ones (roughly) where decision making has been pursued in a way that facilitates genuine collective deliberation rather than mere aggregation of preferences.[15] Such deliberation is conceived (again, roughly) as an opportunity to not only express one's view but also to give reasons for one's view and to consider reasons for opposing views in a way that improves those views with the integration of new information and the refinement of our mutual understanding of the relevant issues.[16]

I do not want to rest the argument of this paper on the success of the deliberative democrats' case for the legitimacy of democracy based in collective deliberation, but I do think their theory represents one of the main ways to address skepticism about the legitimacy of majority rule. And to the extent that it does, the legitimacy of majority rule depends implicitly on valuing not just expressive sincerity but also discursive integrity.[17] According to the deliberative democrats, it is not enough for political legitimacy that people are free to (sincerely) express their opinions in the course of making collective decisions in a democracy. Those opinions must also be subject to a fair process of giving and taking reasons in pursuit of a refined collective view about what to think, feel, and do.[18] And my claim is that achieving legitimacy in this way requires more of political agents than mere sincerity; it requires that people take responsibility for what they say,

15 This is partially inspired by Habermas's discourse approach to grounding moral principles in the constitutive preconditions of engaging in a specific sort of public dialogue about collective decisions (*Moral Consciousness and Communicative Action* and *Between Facts and Norms*). My aim in this paper is not to add to the discussion of whether anything like a universal morality can be derived from these conditions, but I do argue below that reflecting on what we are doing when engaged in public political discussions of contentious issues might help us identify some principles of responsible public debate that I think should count as constitutive preconditions for engaging in such debate in a democracy.

16 This is a controversial and widely discussed view in democratic theory. For articulation and defense see especially Cohen, "Deliberation and Democratic Legitimacy"; Dryzek, *Deliberative Democracy and Beyond*; Gutmann and Thompson, *Why Deliberative Democracy?*; Peter, *Democratic Legitimacy*; and Fishkin, *When the People Speak*.

17 For a similar point, see Markovits, "The Trouble with Being Earnest."

18 One might worry that this stress on justification makes democratic deliberation more susceptible to confirmation bias. However, see Landemore for an argument that public justification is a corrective to confirmation bias (*Democratic Reason*). See also Mercier and Landemore, "Reasoning Is for Arguing."

seeking to back it up with reasons when appropriately challenged or modulate it when met with plausible counterarguments.[19]

Moreover, and more intuitively, failures of discursive integrity threaten pursuit of mutually acceptable and practically actionable collective decisions in a way that is distinct from failures of expressive sincerity. Consider, for example, the hypocrisy involved in saying people should behave in some way while failing to manifest in one's own actions a belief that this injunction applies to oneself. Or consider the persistent devil's advocate who raises critical points that they do not really believe. Or consider the public official who justifies state use of force on grounds she does not really believe but that she knows will convince those who are affected.[20] These examples of insincerity can certainly be annoying, and they may even be morally bad in certain circumstances.[21] However, that does not mean that the claims made cannot be supported with convincing reasons

19 Cass Sunstein argues that public discussion within groups sharing common viewpoints tends to harden opinions and lead people to adopt more extreme versions of their original views ("The Law of Group Polarization"). It is not my intention here to assess whether this is a problem for deliberative democrats' account of legitimacy. But there is a natural explanation of this phenomenon having to do with expressive sincerity and the sorts of affective alignment that go hand in hand with political polarization. In choosing whom to hang out with, whom to be friends with, whom to live with, people are possibly more concerned with alignment of feelings toward things than whether someone takes justificatory responsibility for what they say. If so, expressive sincerity in personal relationships provides a means for increasing affective alignment within these relationships. My suggestion, then, in emphasizing discursive responsibility, is to highlight some common ground between deliberative democrats and their critics on this point. We all agree that collective decision making improves when we move beyond affective alignment; there is just disagreement about whether robust public deliberation facilitates that.

20 Brian Carey argues that a weaker honesty principle should replace a stronger sincerity principle in characterizing the ideal of public reason justifications of state coercion ("Public Reason—Honesty, Not Sincerity"). Roughly speaking, he thinks that we should be honest about whether we believe the reasons we offer in public justification of state use of force, which is consistent with not sincerely endorsing those as one's own personal reasons for viewing such force as justified on a particular occasion. I see this as an interesting middle ground between the ideals of expressivist sincerity and discursive integrity, and I am sympathetic to the claim that, insofar as we hold certain state decisions to a standard of public reason, an honesty principle is more realistic than a sincerity principle. My topic here, however, is the broader category of public political assertions, and honesty about one's reasons for one's views is a step toward taking discursive integrity—but it does not get us all the way there.

21 For discussion and critique of the moral condemnation of hypocrisy, see Dover, "The Walk and the Talk"; and O'Brien and Whelan, "You're Such a Hypocrite." For a defense of the democratic value of attempts at persuasion even when they are not aimed at consensus, see Garsten, *Saving Persuasion*.

or that better collective decisions cannot be reached by engaging with these reasons. On the flipside, however, if these people were to also refuse to try to justify or modulate what they say in the face of disagreement, it would be hard to know how to take the discussion forward in pursuit of some mutually acceptable collective decision about how to behave.

So, in short, my first argument for the importance of discursive integrity for fixing contemporary political debate is that efforts to live up to this ideal seem crucial for achieving the sorts of conditions for collective decision making that deliberative democrats and others have long been advocating as improving and possibly legitimating democratic governance. To be sure, contemporary democracies often fall short of the ideals of deliberative democrats, but these ideals reveal how the legitimacy of majority rule might depend on achieving certain levels of discursive integrity, conceived as a separate ideal from expressive sincerity.

The second argument stems from debates in epistemology of self-knowledge about when it is appropriate to attribute a mental state to someone, including oneself. On the one hand, in everyday discourse we tend to accord people a significant degree of first-personal authority about the contents of their own minds, deferring by default to statements about what they believe, want, and feel.[22] On the other hand, attributions of such mental states figure in explanations of people's behaviors and various psychological studies suggest that our minds are far from transparent to first-personal introspection.[23] Consider for example the idea that fossil fuel consumption is causing dangerous climate change or the view that a diet including meat causes extreme pain and cruelty to animals. Many people profess belief in these propositions, yet they act in ways that seem to be inconsistent with such belief. What do these people really think (believe, feel, intend) about climate doom or animal cruelty?

I will not venture an answer on this vexed question. My point in raising it is simply to highlight the fact that it is not always clear that people can even follow the injunction to be sincere.[24] For it is not always clear, even to speakers themselves, what exactly they think about controversial and complicated issues.

22 For discussion, see Burge, "Our Entitlement to Self-Knowledge"; Wright, "Self-Knowledge"; Moran, *Authority and Estrangement*; and Bar-On, *Speaking My Mind*.

23 For discussion, see Bem, *Beliefs, Attitudes, and Human Affairs*; Lycan, "Tacit Beliefs"; Gopnik and Meltzoff, "Minds, Bodies and Persons"; Carruthers, "How We Know Our Own Minds"; and Mandelbaum, "Thinking Is Believing."

24 For similar reasons, Ridge argues for a weaker form of the sincerity norm for assertion, which requires only that people assert what they believe that they believe, rather than more simply what they believe ("Sincerity and Expressivism"). For an even stronger conception, see Eriksson, who argues that sincere communication requires the speaker to be justified in what they think in order to communicate sincerely ("Straight Talk"). I think Ridge's version

This is especially true in the context of public political discourse. In personal communication about uncontroversial matters with friends and loved ones, it is perfectly fine to assume that people generally know what they think about something, and we evaluate people as honest or dishonest in communicating to the extent that they say what they really think (even if what they think is sometimes that they are not quite sure what to think about something). In other contexts, however, especially when communicating about broad and important societal values or complex matters of fact, I suspect there is often no completely cut and dry answer about what someone really thinks.

Dispositionalist and functionalist accounts of belief can support this idea. These views identify psychological states as complex sets of cognitive and behavioral functions or dispositions.[25] And this opens up the possibility that, when it comes to complex or controversial issues, people often have some kind of in-between state.[26] That is to say, they have some of the dispositions associated with thinking that *p* but not others, or they manifest the relevant dispositions in some circumstances but not all of the circumstances we would normally expect. Because of this, there may simply be no fact of the matter about whether they really think *p*, and so no fact of the matter about whether they are being sincere when asserting *p*.

I do not want to rest my case here on any particular theory of folk psychological states or on the controversial idea that there are "in-between" states. My point in referring to these ideas is simply to make vivid a difficulty with implementing the ideal of expressive sincerity when it comes to discussions of complex and controversial issues. This difficulty does not arise (or at least does not arise as starkly) for the ideal of discursive integrity. Living up to that ideal requires taking responsibility for what you say, being prepared to either offer considerations in its favor or to revise your stance in light of countervailing arguments. To be sure, this process of collective deliberation requires some degree of self-awareness of what one is inclined to think in various circumstances, and one must

---

is too weak and Eriksson's version is too strong, but their motivations for weakening or strengthening the norm support my point above.

25 For relevant dispositionalist theories, see Audi, "Dispositional Beliefs and Dispositions to Believe"; and Schwitzgebel, "A Phenomenal, Dispositional Account of Belief." For relevant functionalist theories, see Loar, *Mind and Meaning*; and Leitgeb, *The Stability of Belief*. Although they do not identify belief as one of these kinds of mental states, similar points apply to the sorts of interpretationist accounts defended by Dennett, *The Intentional Stance* and "Real Patterns"; and Davidson, "Rational Animals."

26 See especially Schwitzgebel, "In-Between Believing" and "Acting Contrary to Our Professed Beliefs, or the Gulf between Occurrent Judgment and Dispositional Belief."

have some idea of the reasons on which one is basing one's view.[27] But some of the apparent indeterminacy or in-betweenness of our views about broad and controversial values or complex matters of fact can be explained by our intuitive appreciation of the tenuousness of our reasons in favor of these views and our natural tendency to modulate in light of ongoing debate.

So, my second argument for the importance of discursive integrity is that this ideal, unlike the ideal of expressive sincerity, is less vulnerable to failures of self-understanding and indeterminacy about what one really thinks. If our focus in public political debate is on finding mutually acceptable views by collectively considering reasons for and against various views and the consequences of any particular view, then what individual contributors antecedently think and whether they sincerely express it is not going to be the main thing that matters. Instead, the most important issue will be whether we can achieve mutual recognition of what we are collectively entitled to assume for further democratic reasoning toward conclusions about what to collectively do, think, and feel.

The third argument that discursive integrity deserves more attention when thinking about fixing contemporary public political debate has to do with how testimony is normally thought to transmit epistemic justification but sometimes seems to fail to transmit justification in discussions of complex and normative issues. As mentioned above, we get all sorts of everyday knowledge from the testimony of others, and theories of the epistemology of testimony seek to account for the way that epistemic justification can transmit from speakers to their audiences via testimony. Reductive views seek to explain the transmission of justification in terms of other kinds of epistemic justification, whereas nonreductive views claim that the kind of justification that testimony transmits from speakers to audience is not reducible to other forms of justification. Whichever way one goes on the question of reduction, the Lockean view of communication combines with the epistemology of testimony to suggest a picture whereby access to other people's thoughts via sincere testimony is one of the main processes by which groups of people increase their mutual understanding of the world. When someone says something ordinary, we are usually automatically justified in believing what they said. Of course, we have to employ various strategies for avoiding gullibility, but very many of the beliefs we share with other people are acquired through uncritical acceptance of what other people say. This is, for example, why Wikipedia is so useful.

This picture may be accurate to large swaths of justified beliefs held within a group of people. However, the picture is plainly inadequate for most of the sorts

27 For a useful discussion of some of the main views and a proposed alternative view of knowledge of our own reasons for attitudes, see Keeling, "Knowing Our Reasons."

of views we seek to form in public political debate. Political matters are, unlike most ordinary cases of learning something through testimony, controversial and often morally tinged. For anything someone says about a matter of public policy, complex scientific fact, or deep moral value, we can usually find someone else who would say the opposite; and even when we do not know someone who disagrees, we can usually easily imagine the sort of fellow citizen who would disagree. Because of this, it is not nearly as plausible that testimony—whether from politicians, political commentators, or friends and neighbors—can automatically transmit justification about such matters, and we generally need to form our own views about such matters in a more careful and critical way.[28] One cannot look up answers to controversial questions of public policy on Wikipedia.

That is not to say that we should ignore what other people say when thinking about political matters. Far from it, what I want to suggest is that, when it comes to public policies, complex scientific facts, or moral values, it is more important than in quotidian cases of testimony to listen to people's *reasons* for their views. And this is why discursive integrity is so important in public political debate. If people are prepared to say only, "This is how I see things; it's my sincere opinion that *p*," then when it comes to controversial and important matters, their testimony is not going to be very helpful for forming well-justified collective views. Public political debate also needs people who are prepared to say, "The reason I think that *p*, is *q*, *r*, and *s*." And even "To the objections *a*, *b*, and *c*, I would try to respond with *e*, *f*, and *g*, but I might need to modulate my original position somewhat, in which case I'd still endorse *p**."

There are epistemic reasons that such inferential intercourse with fellow members of a polity is more likely to lead individuals to form robustly justified views.[29] In my view, however, it is equally important that discussions proceeding under a mutually acknowledged ideal of discursive integrity also promote a kind of mutual understanding and shared values that are not otherwise available. In justifying our views to one another in public political debate, ideally we are not primarily trying to convince other people of our views by giving our personal reasons for these views or giving reasons we believe our audience should accept

28 For discussion of the related issue of the strangeness of deference to moral testimony, see Nickel, "Moral Testimony and Its Authority"; McGrath, "The Puzzle of Pure Moral Deference" and "Skepticism about Moral Expertise as a Puzzle for Moral Realism"; and Fletcher, "Moral Testimony."

29 See Landemore, *Democratic Reason*, 124–30, for an argument that participation in collective democratic reasoning as part of a large group with diverse views and abilities has ameliorative effects on failures of individual reasoning. For further discussion of the essential collectivity of the pursuit of knowledge, see Chrisman, "Believing as We Ought and the Democratic Route to Knowledge."

as their personal reasons for sharing the views, we are also seeking to legitimate our collective decisions for a potentially broader group. To this end, what matters in justifying myself is not just what convinces me, nor what I think will convince you, but what should convince (all of) "us."[30]

Of course, in reality, nothing convinces everyone, and maybe nothing should. But if we conceive of a public justificatory standard as the regulative ideal of collective reasoning, then taking on the sort of discursive responsibility I am suggesting we expect each other to take on when making contributions to public political debate will not just move the relevant "us" toward broader common knowledge but also serve to constitute and expand the domain of our shared concerns and our mutual understanding.

In summary, then, my third argument for the importance of discursive integrity is that such debate is often about controversial, complicated, and moral issues where we should not take other people's word for things but need to investigate reasons collectively. For that, we need something other than expressive sincerity; we need discursive integrity. The complexity and controversy can extend to other domains such as scientific or philosophical debate (where I also think discursive integrity is an important ideal), but in the political realm I have suggested there is also a moral reason that this is important having to do with the way collective reasoning aims to not only explain and convince but also to constitute shared concern and mutual understanding capable of fostering cooperative collective action.

Earlier, I sought to distinguish the communicative ideals of expressive sincerity and discursive integrity. Once distinguished, I think it is fairly clear that discursive integrity is important to communication, but the relative importance of discursive integrity for good public political debate is not widely appreciated in the popular political culture of contemporary democracies. So, in this section, I have developed three independent arguments for thinking that discursive integrity deserves significantly more attention than it presently gets. Even if these arguments convince you of their conclusion, they may leave you wondering

30 I do not mean here to endorse the so-called public reason requirement on justification of state policy, but debate about it is relevant to our appreciation of discursive integrity. For discussion, see especially Rawls, *Political Liberalism*, 36–37, 55–57; Gaus, *The Order of Public Reason*; and Quong, "What Is the Point of Public Reason?" I am inspired here by Postema, who makes a similar point to the one in the text above arguing for a robust conception of public justification, which he contrasts with two thinner conceptions having to do with articulating one's own personal reasons and having to do with articulating reasons one hopes to use to convince specific interlocutors ("Public Practical Reason"). According to him, neither of these is ideal public justification because of the way they fail to treat reasoning as a collective project.

what anyone can do about failures of discursive integrity among political agents, especially in a political-media landscape where scandal sells and reasoned arguments never go viral. In the final section I venture speculation about how things might be improved.

4. SECOND-ORDER DEBATE ABOUT THE RULES OF PUBLIC POLITICAL DEBATE

Public political debate is a social practice. And social practices are partially constituted by the mutual (even if only implicit and often imperfect) acknowledgement of a set of norms by participants in the practice. This sort of mutual acknowledgement partially constitutes individual actions as part of some collective activity. For example, we queue to get into a concert, applaud when the song is finished, chant "encore" if it was good, go silent when the band returns to the stage. Or we sit down together for a meal, pass the food around, use the correct utensils, say "cheers" while clinking glasses. These are social practices that lack explicit rules, but they seem to involve mutual acknowledgment of norms, and individual actions seem to make little sense absent reference to the collective activity partly constituted by this mutual acknowledgement. In a similar fashion, we should view the diffuse activity of public political debate in liberal democracies as a social practice with implicit norms whose mutual acknowledgment is crucial for understanding individual speech acts as part of the relevant practice.

With examples of "broken" political discourse in mind, however, one might worry about whether there actually are any mutually acknowledged norms of public political debate. I suspect the best way to answer this question is to ask participants to try to make more explicit their understanding of rules of the practice, and then to engage in debate about the second-order issue of how best to treat each other when discussing controversial issues of mutual concern.[31] My expectation is that this would foster more explicit discussions of principles for responsible political debate, both among regular citizens and among politicians and political reporters. For instance, if we can insist that political agents tell us not only what they propose regarding first-order issues but also that they suggest and seek to justify views about how public political debate should be conducted, we will be positioned to develop a more constructive mutual understanding

31 This is not meant to deny the political importance of expressions of emotions such as anger, self-respect, and feelings of solidarity. For important discussion, see Boxill, "Self-Respect and Protest"; Bell, "A Woman's Scorn"; Shelby, *Dark Ghettos*, ch. 7; Delmas, *A Duty To Resist*, ch. 6; Pasternak, "Political Rioting"; and Srinivasan, "The Aptness of Anger." I see these sorts of expression as first-order contributions to public political discussion (and sometimes debate), and think they highlight the need for more second-order reflection on when and how such expressions are legitimate.

of responsible political debate—one that will improve operationalization of the ideal of discursive integrity.

So much of public political debate is about first-order questions, such as whether there should be a wealth tax, how we should reform the criminal justice system, and which energy policies are supported by the best climate science. There is also a lot of discussion of particular politicians and their parties, for example: whether Donald Trump is hiding something in his tax returns, whether the British Labour party is protecting anti-Semites, whether the rise of the Conservatives in Canada will block increased environmental protection laws, and whether Narendra Modi implicitly supports attacks on Muslims in India. We can view this discussion of representatives as a sort of second-order debate, one level removed from the issues that matter to shared understanding and collective action in a polity. However, there is another sort of second-order discussion that we do not often have, which is more important; and when it does happen, it seems to get relegated to places such as subcommittees on party convention or conference rules, backroom dealings about the structure of TV debates, and arcane procedural debates in our legislative bodies. This is public political debate about what we might think of as the "rules of the game" in public political discourse. If we could have public debate about the principles of responsible public debate, I think first-order debate would manifest more discursive integrity.

To make this suggestion vivid, I want to close this paper by considering the example of the subreddit r/changemyview, which is an online space self-conceived as "a place to post an opinion you accept may be flawed, in an effort to understand other perspectives on the issue."[32] Users are encouraged to "enter with a mindset for conversation, not debate." A post on the website is titled "Religious institutions should have the right to refuse to marry LGBT+ couples if homosexuality goes against the religion's beliefs."[33] What follows is a six-paragraph explanation of the view and why the person posting thinks it is correct. Then there is a tree of responses and counter-responses that are rated and displayed according to an internal "delta system" whereby users can acknowledge when a response or counter-response changed their view about something related to the original post at least to some degree. The sorts of free democratic debate hosted by this website are often highly informative for those without enough knowledge about an issue to have an opinion. And even when one already has a strongly held opinion about an issue, the back-and-forth discussion often helps to understand reasons for another view and to appreciate nuances of the issue

32 https://www.reddit.com/r/changemyview.

33 https://www.reddit.com/r/changemyview/comments/nwvk7g/cmv_religious_institutions_should_have_the_right, accessed June 10, 2021.

and elements where one's own confidence might be modulated without fully changing one's original view.

It is fairly clear how the site manages such constructive discussions of controversial issues: It has an explicit set of rules that the designers seek to explain and justify and that are enforced through informal reputational policing and explicit moderation. Moreover, within these, it is revelatory of an implicit commitment to the ideal of discursive integrity that the site's first rule for responding to a post is that "Direct responses to a submission must challenge or question at least one aspect of the submitted view." This means the site's designers want to avoid responses that are mere "likes," restatements, and expansions of the original post—however sincere these may happen to be. Noting that they want to avoid echo chambers, they explain this as follows: "If we allowed responses that reinforced the [original poster's] view as top level, [Change My View] would quickly become an echo chamber where only popular opinions were allowed. It would also increase the likelihood that people would come here to soapbox rather than take a critical look at their own viewpoint."[34] So, in the designers' view, it is through respectful challenges that constructive discussion of contentious issues best starts.

Of course, it is not realistic to expect most public political discussion to be designed from the outset with a hyperlink to the rules for engagement or for there to be formal moderators to enforce those rules. And, to be clear, I do not think such explicit reflection on the "rules of the game" will overcome problems with lying and dishonesty. Nevertheless, as long as we understand public political debate as a shared practice with implicitly acknowledged norms, we might begin to improve it by getting participants to reflect more on what they think the norms are and should be and to discuss this second-order issue more explicitly.[35] On the backdrop of the assumption that public political debate in liberal democracies should increase shared understanding of complex and controversial issues and use this to pursue widely acceptable collective decisions about how to live together, we—ordinary citizens, political commentators, and politicians—

34 https://www.reddit.com/r/changemyview/wiki/rules#wiki_rule_1, accessed June 10, 2021.

35 Peter defends a congenial interpretation of Rawls's idea of public reason in *Democratic Legitimacy*, ch. 6. According to her preferred "pure proceduralist" version of a public reason requirement on democratic legitimacy, it is not substantive governmental policy that must be justified by public reasons (i.e., reasons not dependent on substantive and disputed conceptions of what is good) but rather the rules or principles that frame the decision-making process by which political equals who potentially disagree about substantive goods engage in debate about public policy. Although not focused here on the idea of public reason, Peter's kind of pure proceduralism encourages just this sort of second-order public debate about the principles of responsible debate.

might all usefully try to develop and agree to a charter for responsible debate, i.e., a set of rules, which we seek to explain and justify on the model of r/changemyview, for norm-governed contributions to that social practice. For the reasons explained in the previous section, although norms concerning expressive sincerity are going to prove important, norms concerning discursive integrity are going to be even more important to promoting the purpose of this activity.[36]

University of Edinburgh
matthew.chrisman@ed.ac.uk

REFERENCES

Adler, Jonathan. "Epistemological Problems of Testimony." *Stanford Encyclopedia of Philosophy* (Winter 2017).

Alston, William. *Illocutionary Acts and Sentence Meaning*. Ithaca, NY: Cornell University Press, 1999.

Audi, Robert. "Dispositional Beliefs and Dispositions to Believe." *Noûs* 28, no. 4 (December 1994): 419–34.

Bar-On, Dorit. *Speaking My Mind: Expression and Self-Knowledge*. Oxford: Oxford University Press, 2004.

Bell, Macalester. "A Woman's Scorn: Toward a Feminist Defense of Contempt as a Moral Emotion." *Hypatia* 20, no. 4 (November 2005): 80–93.

Bem, Daryl. *Beliefs, Attitudes, and Human Affairs*. Belmont, CA: Brooks/Cole, 1970.

Bennett, Jonathan. *Linguistic Behaviour*. Cambridge: Cambridge University Press, 1976.

Boxill, Bernard. "Self-Respect and Protest." *Philosophy and Public Affairs* 6, no. 1 (Autumn 1976): 58–69.

Brandom, Robert. "Asserting." *Noûs* 17, no. 4 (November 1983): 637–50.

———. *Making It Explicit*. Cambridge, MA: Harvard University Press, 1993.

Burge, Tyler. "Our Entitlement to Self-Knowledge." *Proceedings of the Aristotelian Society* 96, no. 1 (June 1996): 91–116.

36 For helpful feedback on previous versions of this material, I would like to thank James Brown, Michael Cholbi, Rowan Cruft, Guy Fletcher, Peter Graham, Graham Hubbs, Alice König, Peter McColl, Filipa Melo Lopes, John O'Connor, Michael Ridge, Piers Turner, anonymous reviewers for this journal, and participants in the Ethics Work in Progress series at the University of Edinburgh.

Carey, Brian. "Public Reason—Honesty, Not Sincerity." *Journal of Political Philosophy* 26, no. 1 (March 2018): 47–64.

Carruthers, Peter. "How We Know Our Own Minds: The Relationship between Mindreading and Metacognition." *Behavioral and Brain Sciences* 32, no. 2 (April 2009): 121–82.

Chrisman, Matthew. "Believing as We Ought and the Democratic Route to Knowledge." In *The Ethics of Belief and Beyond: Understanding Mental Normativity*, edited by Sebastian Schmidt and Gerhard Ernst, 47–70. Abingdon, UK: Routledge, 2020.

———. *The Meaning of "Ought": Beyond Descriptivism and Expressivism in Metaethics*. Oxford: Oxford University Press, 2016.

———. "Two Nondescriptivist Views of Normative and Evaluative Statements." *Canadian Journal of Philosophy* 48, nos. 3–4 (2018): 405–24.

Cohen, Joshua. "Deliberation and Democratic Legitimacy." In *The Good Polity: Normative Analysis of the State*, edited by Alan Hamlin and Philip Pettit, 17–34. Oxford: Blackwell, 1989.

Davidson, Donald. "Rational Animals." *Dialectica* 36, no. 4 (December 1982): 317–27.

Davis, Wayne. *Meaning, Expression and Thought*. Cambridge: Cambridge University Press, 2003.

Delmas, Candice. *A Duty To Resist: When Disobedience Should Be Uncivil*. New York: Oxford University Press, 2018.

Dennett, Daniel. *The Intentional Stance*. Cambridge, MA: MIT Press, 1987.

———. "Real Patterns." *Journal of Philosophy* 88, no. 1 (1991): 27–51.

Dover, Daniela. "The Walk and the Talk." *Philosophical Review* 128, no. 4 (October 2019): 387–422.

Dryzek, John. *Deliberative Democracy and Beyond: Liberals, Critics, Contestations*. Oxford: Oxford University Press, 2002.

Eriksson, John. "Straight Talk: Conceptions of Sincerity in Speech." *Philosophical Studies* 153, no. 2 (March 2011): 213–34.

Fishkin, James. *When the People Speak: Deliberative Democracy and Public Consultation*. Oxford: Oxford University Press, 2009.

Fletcher, Guy. "Moral Testimony: Once More with Feeling." In *Oxford Studies in Metaethics*, vol. 11, edited by Russ Shafer-Landau, 45–74. Oxford: Oxford University Press, 2016.

Garsten, Bryan. *Saving Persuasion: A Defense of Rhetoric and Judgment*. Cambridge, MA: Harvard University Press, 2009.

Gaus, Gerald. *The Order of Public Reason: A Theory of Freedom and Morality in a Diverse and Bounded World*. Cambridge: Cambridge University Press, 2011.

Gopnik, Alison, and Andrew N. Meltzoff. "Minds, Bodies and Persons: Young Children's Understanding of the Self and Others as Reflected in Imitation and 'Theory of Mind' Research." In *Self-Awareness in Animals and Humans*, edited by Sue Taylor Parker, Robert W. Mitchell, and Maria L. Boccia, 166–86. New York: Cambridge University Press, 1994.

Grice, Paul. "Logic and Conversation." In *Syntax and Semantics*. Vol. 3, *Speech Acts*, edited by Peter Cole and Jerry L. Morgan, 41–58. New York: Academic Press, 1975.

———. *Studies in the Way of Words*. Cambridge, MA: Harvard University Press, 1989.

Gutmann, Amy, and Dennis Thompson. *Why Deliberative Democracy?* Princeton: Princeton University Press, 2004.

Habermas, Jürgen. *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*. Translated by William Rehg. Cambridge: Polity, 1996.

———. *Moral Consciousness and Communicative Action*. Cambridge, MA: MIT Press, 1990.

Hartshorne, Charles, Paul Weiss, and Arthur W. Burks, eds. *Collected Papers of Charles Sanders Peirce*. Cambridge, MA: Harvard University Press, 1931–1958.

Hawley, Katherine. *How to Be Trustworthy*, Oxford: Oxford University Press, 2019.

Keeling, Sophie. "Knowing Our Reasons: Distinctive Self-Knowledge of Why We Hold Our Attitudes and Perform Actions." *Philosophy and Phenomenological Research* 102, no. 2 (March 2021): 318–41.

Kukla, Rebecca, and Mark Lance. *"Yo!" and "Lo!": The Pragmatic Topography of the Space of Reasons*. Cambridge, MA: Harvard University Press, 2009.

Lance, Mark. "Some Reflections on the Sport of Language." *Philosophical Perspectives* 12 (1998): 219–40.

Landemore, Hélène. *Democratic Reason: Politics, Collective Intelligence, and the Rule of the Many*. Princeton: Princeton University Press, 2013.

Leitgeb, Hannes. *The Stability of Belief*. Oxford: Oxford University Press, 2017.

Loar, Brian. *Mind and Meaning*. Cambridge: Cambridge University Press, 1981.

Locke, John. *An Essay Concerning Human Understanding*. 1690. Edited by Peter Nidditch. Oxford: Clarendon, 1979.

Lycan, William. "Tacit Beliefs." In *Belief: Form, Content, and Function*, edited by Radu Bogdan, 61–82. Oxford: Oxford University Press, 1986.

MacFarlane, John. "What Is Assertion?" In *Assertion: New Philosophical Essays*, edited by Jessica Brown and Herman Cappelen, 79–96. Oxford: Oxford University Press, 2011.

Mandelbaum, Eric. "Thinking Is Believing." *Inquiry* 57, no. 1 (2014): 55–96.

Markovits, Elizabeth. *The Politics of Sincerity: Plato, Frank Speech, and Democratic Judgment*. University Park, PA: Pennsylvania State University Press, 2008.

———. "The Trouble with Being Earnest: Deliberative Democracy and the Sincerity Norm." *Journal of Political Philosophy* 14, no. 3 (September 2006): 249–69.

McGrath, Sarah. "The Puzzle of Pure Moral Deference." *Philosophical Perspectives* 23 (2009): 321–44.

———. "Skepticism about Moral Expertise as a Puzzle for Moral Realism." *Journal of Philosophy* 108, no. 3 (March 2011): 11–137.

Mercier, Hugo, and Hélène Landemore. "Reasoning Is for Arguing: Understanding the Successes and Failures of Deliberation." *Political Psychology* 33, no. 2 (April 2012): 243–58.

Miller, Christian. "Honesty." In *Moral Psychology*. Vol. 5, *Virtue and Character*, edited by Walter Sinnott-Armstrong and Christian Miller, 237–74. Cambridge, MA: MIT Press, 2017.

Moran, Richard. *Authority and Estrangement: An Essay on Self-Knowledge*. Princeton: Princeton University Press, 2001.

Nickel, Philip. "Moral Testimony and Its Authority." *Ethical Theory and Moral Practice* 4, no. 3 (2001): 253–66.

O'Brien, Maggie, and Alexandra Whelan. "You're Such a Hypocrite." Unpublished manuscript.

Ozar, Anne. "Sincerity, Honesty, and Communicative Truthfulness." *Philosophy Today* 57, no. 4 (Winter 2013): 343–57.

Pasternak, Avia. "Political Rioting: A Moral Assessment." *Philosophy and Public Affairs* 46, no. 4 (Fall 2018): 384–418.

Peter, Fabienne. *Democratic Legitimacy*. New York: Routledge, 2008.

Postema, Gerald. "Public Practical Reason: An Archeology." *Social Philosophy and Policy* 12, no. 1 (Winter 1995): 70–74.

Quong, Jonathan "What Is the Point of Public Reason?" *Philosophical Studies* 170, no. 3 (September 2014): 545–53.

Rawls, John. *Political Liberalism*. New York: Columbia University Press, 1996.

Richardson, Henry. "Noncognitivist Trumpism: Partisanship and Political Reasoning." *Journal of Social Philosophy* 50, no 4 (Winter 2019): 642–43.

Ridge, Michael. "Sincerity and Expressivism." *Philosophical Studies* 131, no. 2 (November 2006): 487–510.

Schiffer, Stephen. *Meaning*. Oxford: Clarendon, 1972.

Schwitzgebel, Eric. "Acting Contrary to Our Professed Beliefs, or the Gulf between Occurrent Judgment and Dispositional Belief." *Pacific Philosophical Quarterly* 91, no. 4 (December 2010): 531–53.

———. "In-Between Believing." *Philosophical Quarterly* 51, no. 202 (January 2001): 76–82.

———. "A Phenomenal, Dispositional Account of Belief." *Noûs* 36, no. 2 (June 2002): 249–75.

Searle, John. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge: Cambridge University Press, 1969.

Sellars, Wilfrid. "Meaning as Functional Classification: A Perspective on the Relation of Syntax to Semantics." *Synthese* 27, nos. 3–4 (July–August 1974): 417–37.

———. "Some Reflections on Language Games." *Philosophy of Science* 21, no. 3 (July 1954): 204–28.

Shelby, Tommie. *Dark Ghettos: Injustice, Dissent, and Reform*. Cambridge, MA: Harvard University Press, 2016.

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

Srinivasan, Amia. "The Aptness of Anger." *Journal of Political Philosophy* 26, no. 2 (June 2018): 123–44.

Sunstein, Cass. "The Law of Group Polarization." *Journal of Political Philosophy* 10, no. 2 (June 2002): 175–95.

Williams, Bernard. *Truth and Truthfulness: An Essay in Genealogy*. Princeton: Princeton University Press, 2002.

Wright, Crispin. "Self-Knowledge: The Wittgensteinian Legacy." *Royal Institute of Philosophy Supplement* 43 (1998): 101–22.

Journal of Ethics and Social Philosophy
VOL. 22, NO. 2 · JULY 2022
https://doi.org/10.26556/jesp.v22i2.1599

EPISTEMIC TRESPASSING AND EXPERT WITNESS TESTIMONY

Mark Satta

NATHAN BALLANTYNE recently coined the term *epistemic trespassing* to refer to the act of judging matters outside one's field of expertise.[1] Mikkel Gerken has examined a specific type of epistemic trespassing, *expert trespassing testimony* (i.e., epistemic trespassing via testimony).[2] Based on his conclusion that expert trespassing testimony can be both morally and epistemically problematic, Gerken offers the following guideline: "When *S* provides expert trespassing testimony in a context where it may likely and/or reasonably be taken to be expert testimony, *S* should qualify her testimony to indicate that it does not amount to expert testimony."[3]

In this paper, I assess Gerken's guideline—which he calls the "Expert Trespassing Guideline"—as applied to expert witness testimony in a court of law. I conclude that, depending on how it is interpreted, Gerken's guideline either fails to give relevant guidance or gives the wrong guidance when applied to expert witness testimony in court. I argue instead for the following:

> *No Courtroom Trespassing Principle*: Those participating as expert witnesses in legal trials should not make any claim outside their area of expertise if the claim is of the type that normally could only be offered by a properly qualified expert witness.[4]

1 Ballantyne, "Epistemic Trespassing," 367. This leads to the question of what counts as "judging matters." I assume that reaching a conclusion by accepting another's testimony does not count as judging the matter for purposes of epistemic trespassing. Rather, epistemic trespassing seems, at its core, to be about relying on oneself to reach a conclusion in an expert domain where one lacks the epistemic foundation for such self-reliance.

2 Gerken, "Expert Trespassing Testimony and the Ethics of Science Communication," 299, 300.

3 Gerken, "Expert Trespassing Testimony and the Ethics of Science Communication," 299, 301, 310.

4 The qualification "of the type that normally could only be offered by a properly qualified expert witness" is necessary because in laying the foundation for expert testimony, an ex-

Following Gerken, the "should" claim in my principle can be read both epistemically and morally. Thus, I will argue that the No Courtroom Trespassing Principle (hereafter sometimes referred to simply as the principle) is true on both an epistemic and a moral reading. As a precursor, I will also argue that the principle is true if its "should" is interpreted legally. I make this legal argument because the legal impermissibility of epistemic trespassing by expert witnesses contributes to the specific social and epistemic conditions of trials that make epistemic trespassing by expert witnesses epistemically and morally impermissible. The legal impermissibility is part, but not all, of what grounds the epistemic and moral impermissibility.

In section 1, I provide relevant background on the US legal system. In section 2, I examine Gerken's guideline. I then argue for my first conclusion, which is that, for both epistemic and moral reasons, my principle rather than Gerken's guideline should be applied to expert witness testimony in court. With that first conclusion in mind, I then argue for the following additional conclusions.

My second conclusion is that judges, litigators, and jurors can often reliably identify at least moderate and severe forms of epistemic trespassing using information standardly provided about expert witnesses, such as their credentials and track record. Here I appeal to the philosophical literature that addresses more general questions about a layperson's ability to identify experts.

Third, judges and litigators should take epistemic trespassing seriously. Judges should deny admission as expert witnesses to those whose testimony would constitute significant expert trespassing testimony and should not permit expert witnesses to make claims that constitute epistemic trespassing. Litigators should not request the admission of epistemic trespassers as expert witnesses and should object when epistemic trespassing occurs on the witness stand. When epistemic trespassers are admitted to testify in court as expert witnesses, litigators for the opposing side should vigorously cross-examine the epistemic trespasser with the aim of revealing their trespassing. Uses of "should" in this paragraph indicate practical advice about what judges and lawyers ought to do when they recognize epistemic trespassing. But this practical advice derives from a more general normative claim: even if a judge were to fail to recognize an expert witness as an epistemic trespasser, it remains the case that, in virtue of the epistemic and moral badness of epistemic trespassing as an expert witness, the witness should not have epistemically trespassed.

Finally, identifiable epistemic trespassing provides jury members with defea-

pert witness may be asked to testify about general background information. Such testimony serves to provide the foundation for later expert testimony but may not itself require expertise to give.

sible evidence against the truth of the claims made by epistemic trespassers. At least three reasons can be given for this: (a) epistemic trespassing counts against the reliability of an expert witness, (b) epistemic trespassing counts against the trustworthiness of an expert witness, and (c) epistemic trespassing counts in favor of the conclusion that no qualified expert may have been willing to make at least some of the trespasser's claims. When I say that something counts against a view, all I mean is that it can serve as a rational ground to lower one's credence or confidence in the view. This is compatible with one believing or having a high credence in the view, so long as other evidence sufficiently counts in favor of the view. Similarly, when I say something counts in favor of a view, all I mean is that it can serve as a rational ground for raising one's credence or confidence in the view. This is compatible with one disbelieving or having a low credence in the view, so long as other evidence sufficiently counts against the view.

For historically contingent reasons, in recent decades scholarship on expert witness jurisprudence in the United States has focused mostly on the reliability and nature of the *methods* used by expert witnesses.[5] There has been less focus on the expertise of expert witnesses themselves. Assessing the reliability and nature of an expert's method is important, but focusing on method alone misses some important issues related to expert witness testimony. In assessing potential expert witnesses, we should look at both the reliability of the methods they rely on and the nature of their expertise. The examination of one need not, and should not, come at the exclusion of the other. This paper is written primarily with the US federal legal system in mind, but the epistemic and moral spirit of the conclusions reached applies to US state judiciaries and many other national judiciaries as well.

1. EXPERT WITNESSES IN US LAW

Before turning to the philosophical questions, it will be useful to have in mind some information about US federal law concerning expert witnesses. While US evidence law is imperfect, my goal here is not to critique US evidence law. Rather, current evidence law forms part of the backdrop from which I address how expert witnesses and other individuals should behave, both epistemically and morally. Current evidence law affects what one epistemically should do in court because current evidence law helps shape our social and epistemic expectations during trials. If the law were different (and, as a result, our social and epistemic expectations during trials were different), then perhaps what actors epistemically should do in court might be different too. Thus, my paper should be under-

5 For more information on the history of US evidence law, see Haack, "The Expert Witness."

stood as advocating a set of behaviors in court that are rooted in epistemic and moral reasons that account for the way the law is.

Federal law permits two types of witnesses to testify at trial: lay witnesses and expert witnesses.[6] Lay witnesses are admitted to testify about relevant knowledge acquired by personal perception (e.g., having witnessed someone running from a crime scene).[7] In contrast, expert witnesses are admitted to testify about relevant matters by virtue of their expertise acquired through "knowledge, skill, experience, training, or education."[8] Expertise includes scientific, technical, and other specialized knowledge.[9] In order to be qualified to testify as an expert witness, an expert's testimony must (a) help the trier of fact understand the evidence or determine a fact at issue, (b) be based on sufficient facts or data, (c) be the product of reliable principles and methods, and (d) be an instance of the expert reliably applying the principles or methods to the facts in the case.[10]

In recent times, US federal evidence law for the admission of expert witness testimony has been guided by two key precedents. The first, the *Frye* test, held that a scientific technique is admissible only when the technique is generally accepted as reliable in the relevant scientific community.[11] This test was superseded by the US Supreme Court's 1993 decision in *Daubert v. Merrell Dow Pharmaceuticals Inc.*, which enacted a more "flexible" test that focused "solely on principles and methodology."[12] Factors to be considered under *Daubert* include whether the relevant theory or technique has been tested and whether it has been vetted through peer-reviewed publication.[13]

Much of the scholarship about US evidence law has focused on interpreting the older *Frye* test and the newer *Daubert* standard. Neither test is directly about what makes someone a relevant expert. Rather, both tests are about the methods used as the basis for expert witness testimony. As a result, US expert witness law scholarship has tended to focus on an expert witness's methodology rather than their purported expertise. But this attention to techniques and methods does not mean the law does not require that expert witnesses be experts in the relevant fields. It does.

6 Fed. R. Evid. 701, 702.

7 Fed. R. Evid. 602, 701.

8 Fed. R. Evid. 702; cf. Civ. Pro. R. 35 for the standards for expert witness testimony in England and Wales.

9 Fed. R. Evid. 702; see also, Kumho Tire Co. v. Carmichael, 526 U.S. 137 (1999).

10 Fed. R. Evid. 702 (a)–(d).

11 Frye v. United States, 293 F. 1013, 1014 (D.C. Cir. 1923).

12 Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579, 594–95 (1993).

13 Daubert, 509 U.S. at 593.

We can see this first by looking at how various courts have interpreted and applied the *Daubert* standard. For example, on remand after the Supreme Court's holding in *Daubert*, the US Court of Appeals for the Ninth Circuit held that "one very significant fact to be considered is whether the experts are proposing to testify about matters growing naturally and directly out of research they have conducted independent of the litigation."[14] The Ninth Circuit then clarifies as follows: "That an expert testifies based on research he has conducted independent of the litigation provides important, objective proof that the research comports with the dictates of good science."[15] Here the Ninth Circuit put forward a standard even higher than that the expert simply be an expert in the relevant field. Rather, the court takes as a "very significant fact" that an expert's research be directly related to the relevant matters. If this higher standard is a very significant factor to be considered, than *a fortiori* so is the more general requirement that one be an expert in the relevant field to begin with.

Similar ideas can be seen in the US Court of Appeals for the Seventh Circuit's ruling that an expert witness should be "as careful as he would be in his regular professional work outside his paid litigation consulting" and the Supreme Court's statement that before admitting an expert witness the trial court judge should be assured that the expert "employs in the courtroom the same level of intellectual rigor that characterizes the practice of an expert in the relevant field."[16]

The need for an expert witness to be an expert in the relevant field is stated even more clearly in the Advisory Committee Notes for the 2000 amendment to the Federal Rules of Evidence. The notes state that "the expert's testimony must be grounded in an accepted body of learning or experience *in the expert's field*, and the expert must explain how the conclusion is so grounded."[17]

To help show how high the stakes can be when expert witnesses engage in epistemic trespassing, consider an example.[18] In 1988, Curtis Weeks was being transferred from one Texas prison to another. Weeks, who was HIV positive,

14 Daubert v. Merrell Dow Pharmaceuticals, Inc., 43 F.3d 1311, 1317 (9th Cir. 1995).

15 Daubert, 43 F.3d at 1317.

16 Sheehan v. Daily Racing Form, Inc. 104 F.3d 940, 942 (7th Cir. 1997); Kumho Tire Co. v. Carmichael, 526 US 137, 152 (1999).

17 Mueller and Kirkpatrick, *Federal Rules of Evidence with Advisory Committee Notes and Legislative History*, 175; emphasis added.

18 In the example that follows, *Daubert* would not have been binding precedent both because *Daubert* had not yet been decided at the time of the trial and because Texas evidence law rather than federal evidence law would have been applied. But the philosophical arguments in this paper do not apply solely to trials that are subject to *Daubert* or *Frye*, so the inapplicability of *Daubert* is not a shortcoming in using this trial as an example of epistemic trespassing in action.

tried to escape en route. During his attempted escape he declared that he was "going to take somebody with him" and spat twice on the face of a prison guard. For this, Weeks was convicted of attempted murder by the state of Texas.[19]

At the time of Weeks's trial, as now, attempted murder in Texas requires that the perpetrator commit an act that "could have caused the death of the complainant but failed to do so."[20] Saliva does not transmit HIV.[21] And no one has ever seroconverted because someone living with HIV spat on them.[22] Thus, Weeks could not have committed attempted murder simply by spitting on the prison guard because his action could not have caused the death of the complainant. Yet, a jury convicted Weeks, and his conviction was upheld by a Texas Appeals Court, US District Court, and US Appellate Court.[23]

Enough was known about the transmission of HIV at the time of Weeks's conviction that the American Civil Liberties Union and the Texas Human Rights Foundation requested on appeal that the court take judicial notice of the fact that "it is impossible to transmit the virus which causes AIDS by spitting."[24] They made their request under the Texas Rules of Criminal Evidence, which defined a judicially noticed fact as "one not subject to reasonable dispute in that it is … capable of accurate and ready determination by resort to sources whose accuracy cannot reasonably be questioned."[25]

The appeals court denied the request and upheld the conviction on the grounds that "the jury chose to believe the [expert] witnesses who testified that HIV could be transmitted through saliva."[26] The court's ruling attests to the influence that expert witness testimony can have on the outcome of trials. It also attests, as will be shown shortly, to the ability of expert witnesses to abuse their privilege of testifying via epistemic trespassing.

At Weeks's trial, four individuals were admitted to testify as "experts on HIV."[27]

19 See Weeks v. State, 834 S.W.2d 559, 560-61 (Tex. App. Eastland 1992).

20 See Flanagan v. State, 675 S.W.2d 734 (Tex. Cr. App. 1984).

21 See, e.g., Centers for Disease Control and Prevention, "HIV Transmission," https://www.cdc.gov/hiv/basics/transmission.html.

22 Cresswell, Ellis, Hartley, et al., "A Systematic Review of Risk of HIV Transmission through Biting or Spitting."

23 Weeks v. State, 834 S.W.2d 559, 560–61 (Tex. App. Eastland 1992); Weeks v. Collins, 867 F.Supp. 544 (SD Tex. 1994); Weeks v. Scott, 55 F.3d 1059 (5th Cir. 1995).

24 Weeks, 834 S.W.2d at n2.

25 Weeks, 834 S.W.2d at n2. The appeals court acknowledged that "Many of the AIDS experts express the opinion that it is impossible to transmit HIV through saliva," but declined to take judicial notice claiming that the matter had not been "conclusively established."

26 Weeks, 834 S.W.2d at 565.

27 Weeks, 834 S.W.2d at 562. This citation and the following are from the descriptions of the

Three were for the prosecution and one for the defense. The first witness for the prosecution, Mark E. Dowell, holds an MD and was at the time an infectious diseases physician at Baylor College of Medicine. Dowell testified that "the medical community is uncertain as to whether HIV could or could not be transmitted through saliva."[28] He also testified that in a study that tested whether HIV would grow in saliva found that "the virus developed in 3 out of 55 instances," and that he had "seen one report which indicated that there was some 'inhibitor' effect of saliva to HIV."[29] Still, he concluded that "the possibility is low but certainly not zero" that HIV could be transmitted by spitting.[30]

The prosecution's second witness, Paul Drummond Cameron, holds a PhD in psychology and works (both at the time of the trial and now) at the Family Research Institute, a nonprofit that he founded. Cameron, who had been expelled from the American Psychological Association in 1983, does not have training in medicine or the life sciences.[31] The basis for Cameron's admission as an expert witness is unclear, aside from his own testimony that "a goodly amount" of his time was devoted to literature research, including literature on HIV/AIDS.[32] Cameron made claims much stronger than, and contradictory to, Dowell's. Cameron testified that "most experts agree that there has been approximately ten cases of transmission [of HIV] through saliva" and that in his opinion "a person could become infected with HIV by being spit on."[33]

The prosecution's third witness, Lorraine Day, holds an MD and at the time was an orthopedic surgeon at San Francisco General Hospital. Day "admitted that she has not specialized in internal medicine and has not had any formal training in infectious diseases."[34] The grounds for her admission as an "expert on HIV," aside from general medical training, seem to be that she had "begun researching AIDS on her own."[35] Like Cameron, Day's testimony contradicted

witnesses provided in the Texas Court of Appeals's opinion upholding Weeks's conviction.

28 Weeks, 834 S.W.2d at 562.

29 Weeks, 834 S.W.2d at 562.

30 Weeks, 834 S.W.2d at 562. Given his assertion that "the medical community is uncertain whether HIV could or could not be transmitted through saliva," the only reading of this claim that would seem to allow for consistency is if Dowell meant the *epistemic* possibility was low but not zero. The record provides no clarification on this point.

31 Siegel, letter from Max Siegel on behalf of the American Psychological Association to Paul Cameron.

32 Weeks, 834 S.W.2d at 562.

33 Weeks, 834 S.W.2d at 563.

34 Weeks, 834 S.W.2d at 564.

35 Weeks, 834 S.W.2d at 564.

Dowell's claim that "the medical community is uncertain as to whether HIV could or could not be transmitted through saliva." Day claimed that there were "documented cases of saliva transmission of the HIV virus [*sic*]" and that "it was possible that the complainant could contract HIV if appellant spit his saliva onto the complainant's face."[36]

The fourth witness, Richard B. Pollard, was the only expert witness for the defense. Pollard holds an MD and at the time was a professor of internal medicine and microbiology at the University of Texas Medical Branch at Galveston and director of the Diagnostic Virology Laboratory at the University of Texas.[37] Pollard testified that he was board certified in internal medicine, specialized in infectious disease, and had eleven years of "training and experience in the area of virology."[38] As part of his work, Pollard "directed a research program active in both clinical and scientific research focused upon infections with HIV" and "sat on a national panel which looks at all the drug studies conducted by the National Institute [*sic*] of Health for the treatment of AIDS infections."[39]

Pollard stated that it had never been shown that HIV could be transmitted by saliva and that he was "unaware of anyone acquiring HIV as the result of contact with only saliva."[40] He provided rebutting explanations for the instances of HIV transmission by saliva put forward by Cameron and Day. He also provided information that gave greater context for Dowell's testimony, stating that "the HIV which is present in saliva is actually inactive."[41]

Weeks's case provides us with a paradigmatic example of epistemic trespassing by expert witnesses to various degrees. There is the particularly egregious epistemic trespassing of Cameron, whose credentials give him no special ability to make judgments about HIV whatsoever. There is the slightly less blatant epistemic trespassing of Day, who by virtue of her medical training is better positioned to assess the relevant facts than your average layperson, but who went

36 Weeks, 834 S.W.2d at 563, 564. Given how blatant Day's and Cameron's epistemic trespassing was, one may think that the real issue here was their testifying in bad faith, not their being epistemic trespassers. It may be the case that Day and Cameron were testifying in bad faith, but that is compatible with their being epistemic trespassers. Regardless of what beliefs or intentions Day and Cameron had when testifying, the claims they asserted qualify them as issuing judgments beyond their competence. The issuing of these judgments outside their areas of expertise while on the witness stand as expert witnesses is an irreducible part of what went wrong here.

37 Weeks, 834 S.W.2d at 562.

38 Weeks, 834 S.W.2d at 564.

39 Weeks, 834 S.W.2d at 564.

40 Weeks, 834 S.W.2d at 564.

41 Weeks, 834 S.W.2d at 564.

far beyond her area of specialization as an orthopedic surgeon in offering judgments about the transmission of HIV. There is also perhaps a subtle form of epistemic trespassing on the part of Dowell. He has medical training and, unlike Day, specializes in infectious diseases. But there is no evidence in the record that he had ever done scientific research on HIV/AIDS. This may have resulted in gaps in his knowledge, such as Pollard's claim that HIV in saliva is inactive. Finally, there is Pollard, who is a good example of someone with the relevant kind of targeted expertise that an expert witness should have. His credentials are narrowly tailored to provide him with the relevant kind of expertise and he has an active research agenda in the relevant area.

As Weeks's case shows, jurors do not always identify epistemic trespassing or make sound judgments that follow from such recognition. Given this, I will argue that the No Courtroom Trespassing Principle provides better guidance for expert witness testimony than does Gerken's guideline, which only advocates that the testifier qualify their trespassing claims to indicate that they are nonexpert claims. I will also later argue that the judge, litigators, and jurors had the tools needed to identify the epistemic trespassing in this case. As a result of this epistemic trespassing, the judge should have denied Cameron and Day the opportunity to testify. If there was epistemic trespassing on the part of Dowell, it was subtle enough that he should have been allowed to testify as an expert witness. However, it would have been prudent for opposing counsel to cross-examine Dowell with the goal of highlighting where his expertise fell short in comparison to Pollard's.

2. GERKEN'S EXPERT TRESPASSING GUIDELINE

As stated earlier, Mikkel Gerken offers the following guideline.

> *Expert Trespassing Guideline*: When *S* provides expert trespassing testimony in a context where it may likely and/or reasonably be taken to be expert testimony, *S* should qualify her testimony to indicate that it does not amount to expert testimony.[42]

Gerken states that this guideline "articulates a *prima facie* moral obligation."[43] This obligation flows from Gerken's assumptions that it is "morally problematic to put someone in an epistemically inhospitable circumstance, if this could easily have been avoided" and that expert trespassing testimony in the circumstances

42 Gerken, "Expert Trespassing Testimony and the Ethics of Science Communication," 310.

43 Gerken, "Expert Trespassing Testimony and the Ethics of Science Communication," 310.

the guideline addresses are likely to create epistemically inhospitable circumstances—i.e., circumstances that undermine a subject's ability to form truth-conducive beliefs.[44] Thus, for Gerken the moral problem with epistemic trespassing is rooted, at least in part, in an epistemic problem with epistemic trespassing.[45]

Assuming that the "when" in Gerken's guideline can be treated as equivalent to "if," Gerken's guideline is a conditional that outlines something one must do *if* one offers expert trespassing testimony. Gerken's guideline does not advocate for expert trespassing testimony; it only comments on what should occur when it happens. Even recognizing this, there is an ambiguity in Gerken's guideline between what I will call the *absolution reading* and the *mitigation reading*.

On the absolution reading, so long as a subject qualifies their testimony to indicate that it does not amount to expert testimony, then the subject is absolved of moral or epistemic wrongdoing. On this reading, qualification either renders the trespassing permissible or renders it no longer trespassing to begin with.

In contrast, on the mitigation reading, a subject's testimony can remain morally and epistemically impermissible even if they qualify their testimony. On this reading, the guideline's consequent merely articulates a new obligation that one incurs in virtue of committing an epistemic trespass. The trespasser can remain guilty of the epistemic trespass, but their guilt may be mitigated (or at least not increased) by the ameliorative action of qualifying their testimony.[46]

I need not settle the question of which interpretation Gerken intended. If one adopts an absolution reading of Gerken's guideline, the No Courtroom Trespassing Principle entails that Gerken's guideline gets matters wrong in the case of expert witness testimony. Qualification does not absolve the expert witness who engages in epistemic trespassing. If one adopts a mitigation reading of Gerken's guideline, the No Courtroom Trespassing Principle entails that Gerken's guideline is largely irrelevant in cases of expert witness testimony because there should be no epistemic trespassing by expert witnesses in a court of law to begin with.[47]

44 Gerken, "Expert Trespassing Testimony and the Ethics of Science Communication," 307, 304.

45 Gerken, "Expert Trespassing Testimony and the Ethics of Science Communication," 301.

46 The guideline "If *S* engages in a hit and run, then S should call 911 as they drive away from the scene of the accident" is an example of a mitigation guideline rather than an absolution guideline. Calling 911 while driving away from an accident does not absolve a hit-and-run driver of guilt for leaving the scene, but it remains the case that *if* the driver is going to do so, at the very least they should call 911 to get first responders to the scene. This is very different from a principle like, "If you are going to go onto another's private property, then you should secure permission from the owner in advance." This latter principle is an absolution principle. If you get permission from the owner, going onto their private property is no longer a wrong at all.

47 I specify *largely* irrelevant because there are select circumstances under which the guideline could still be of use. For example, if an expert witness catches themselves accidentally of-

Thus, in arguing for the No Courtroom Trespassing Principle, I simultaneously am arguing that in the context of epistemic trespassing by expert witnesses Gerken's guideline is either wrong or largely irrelevant, depending on how it is interpreted.

3. EXPERT WITNESSES SHOULD NOT COMMENT ON MATTERS OUTSIDE THEIR EXPERTISE

In this section, I argue that instead of Gerken's guideline, the primary operative principle in cases of epistemic trespassing by expert witnesses in a court of law is the following.

> *No Courtroom Trespassing Principle*: Those participating as expert witnesses in legal trials should not make any claim outside their area of expertise if the claim is of the type that normally could only be offered by a properly qualified expert witness.

I begin by looking at Gerken's arguments for his guideline in more detail in order to shed light on why the guideline is largely inapplicable in cases of epistemic trespassing by expert witnesses in court.

To motivate his position, Gerken considers two hypothetical cases. In the first case, a meteorologist with a specialization in atmospheric composition is interviewed about global warming by a local news station. During the interview, she answers a question about the impact of global warming on marine life, which is a topic outside her area of specialization. In the second case, a cognitive psychologist specializing in color vision is serving as an expert witness in court. During cross-examination, she makes claims about the significance of the defendant's troubled social environment, which is a topic outside her area of specialization.[48] Gerken's guidelines suggests that the proper remedy in both cases would have been for the speakers to qualify their statements to indicate that they were not speaking in their capacity as experts (e.g., "This is outside the bounds of my expertise, but I think that . . .").

Gerken's guideline may be a sensible remedy for the meteorologist faced with an interview question that covers ground beyond her area of expertise; this seems to be the kind of case Gerken is most concerned with. But his guideline is insufficient when it comes to what the cognitive psychologist should do when tempted to speak about matters outside her area of expertise while on the stand. What the

fering trespassing testimony, their best course of action is to qualify their testimony in the way Gerken's guideline advocates. But the relevance here is in rectifying a mistake, not in deciding what testimony to offer in the first place.

48 Gerken, "Expert Trespassing Testimony and the Ethics of Science Communication," 300–1.

speaker in this second case must do is avoid making any claims that (i) normally could only be made by a qualified expert witness, and (ii) fall outside of her area of expertise. This is what the No Courtroom Trespassing Principle requires.

Various epistemic and moral considerations support the No Courtroom Trespassing Principle. Here I offer two epistemic reasons and three moral reasons for the principle, but because some of those reasons depend in part on the legal status of expert trespassing testimony and the social expectations bound up with that status, I begin with a legal observation about Gerken's guideline.

One can see that Gerken's guideline does not satisfy the legal requirements for expert witness testimony by noticing that as soon as an expert witness tries to qualify a claim in the way that Gerken suggests ("I'm not an expert on this matter, but …"), opposing counsel ought to object to the witness continuing their statement because the testimony would be inadmissible. Expert trespassing testimony is simply not the kind of thing that the witness is permitted to say.[49] This legal observation is relevant to my epistemic and moral arguments in two ways. First, the rules of evidence—which include the rules about expert witness testimony—serve epistemic and moral goals. Such rules are meant to provide jurors with information that will help them reach accurate conclusions.[50] Seeking to provide jurors with information that will help them reach accurate judgments in turn is meant to serve the moral aim of making trials fair and just.[51] Second, an understanding of the rules of evidence may naturally influence how one interprets evidence, including testimony. If one understands that judges should allow only relevant evidence before jurors, one may naturally treat a judge's choice to permit testimony as evidence that it is relevant.[52] If one understands that someone is testifying as an expert witness, one may naturally conclude that an expert witness's testimony reflects expertise and as a result should be given deference.[53]

My two epistemic reasons for the No Courtroom Trespassing Principle correspond to these two legal observations. The first reason is a first-order consideration about the reliability of the witness's statements. The second reason is a second-order consideration about the epistemic value attached to expert witness testimony via the social conditions of trials.

Concerning reliability, recall that in order for an expert witness's testimony

49 This is the rule for substantive claims made for the truth of the matter asserted. This does not extend to things such as answering questions required to lay the foundation for later testimony or to answering questions aimed at assessing the expert's credibility.

50 See Fed. R. Evid. 102.

51 See Fed. R. Evid. 102.

52 See Fed. R. Evid. 402.

53 See Fed. R. Evid. 702.

to be admissible the testimony must be "the product of reliable principles and methods" and the expert must have "reliably applied the principles and methods to the facts of the case." This is an epistemic standard set by the Federal Rules of Evidence. Reliability comes in degrees. Thus, there is a need to establish a threshold at which expert witness testimony is expected to be reliable enough that it is admissible. The threshold set by the Federal Rules of Evidence is that the expert witness must be at least as reliable a testifier as a competent expert within the field of expertise being called upon. After all, it needs to be the case that the witness "employs in the courtroom the same level of intellectual rigor that characterizes the practice of an expert in the relevant field."[54]

This threshold gets things right epistemically. If the criteria for testifying as an expert witness were higher, too few experts would be allowed to testify in court. This would come at the cost of jurors missing out on expert testimony that is likely to be relevant and reliable. But if the criteria for testifying as an expert witness were lower, it would be too easy for unreliable testimony to be presented to jurors. This would come at the cost of jurors gaining misleading testimony from unreliable witnesses. Applying this standard to the example of Weeks's prosecution, Cameron and Day fail to meet this threshold as a psychologist and orthopedic surgeon, respectively. Pollard and Dowell meet the standard as practicing MDs with training in infectious diseases. Epistemically, this seems like the correct result.

There is a second reason why it is epistemically problematic to allow an expert witness to testify beyond the scope of their expertise. This reason is rooted in the social conditions of a trial and the role that expert witnesses play within that structure. Because expert witnesses are put on a pedestal (sometimes literally) as experts, jurors are rational in giving substantial deference to the claims made by those testifying as expert witnesses. This can include a juror's subordination of their own intuitions or judgments in favor of those made by the expert witness. However, when an expert witness engages in epistemic trespassing, this sort of deference is no longer deserved. When expert witnesses engage in expert trespassing testimony on the stand, they *mislead* their audience about the lack of authority for their claims. This creates misleading higher-order evidence for jurors.[55]

At this point, a reader might agree that the social conditions of trials create a problematic risk of an expert witness's trespassing testimony being mistaken for true expert testimony. They might also agree that this creates misleading higher-order evidence for jurors that may cause them to give unwarranted deference

54 Kumho Tire Co. v. Carmichael, 526 U.S. 137, 152 (1999).

55 The same may apply to a judge in a bench trial. In this paper, I have generally glossed over this distinction and treated "jurors" as synonymous with "triers of fact."

to the testifier. But they might disagree that this requires the No Courtroom Trespassing Principle. They may think instead that Gerken's guideline that an expert should qualify their testimony to indicate that it does not amount to expert testimony is a sufficient remedy. After all, if the expert witness clarifies on the particular point that they are not an expert, should not that be sufficient to avoid the creation of misleading higher-order evidence?

While this is a reasonable suggestion, ultimately I think it fails for the following reasons. First, for this suggestion to be effective, jurors must properly take note of the qualification and its scope. Given the hours of sustained attention that trials often require of jurors, it would be easy for jurors to fail to register, process, and later account for such a qualification. That someone was presented as an expert witness is easier to remember than where and when that witness qualified their testimony to indicate that they were speaking beyond their area of expertise. Thus, there is a substantial risk that qualifications will not stop jurors from concluding that trespassing testimony delivered by expert witnesses was genuine expert testimony.

Second, and perhaps more importantly, an expert witness who (1) testifies about a matter that normally requires expertise, (2) qualifies to indicate that their testimony is not expert testimony, but (3) is allowed to proceed with the testimony anyway will *always* create conflicting higher-order evidence by virtue of the rules of expert witness testimony. This is because to testify as an expert witness about a matter that normally requires expertise is to create higher-order evidence that one is a relevant expert. Qualifying one's claims in an attempt to indicate otherwise is at best only partial refutation of that evidence, because the rules of the activity require that if all the relevant parties (which include the judge and litigators) act on such a qualification, the testifier would be prohibited from offering the trespassing testimony. This is an important way in which the expert witness case differs from the television interview case. While in the television interview case, the expert may have been invited to speak because they have expertise, there is no requirement that the expert speak only about matters within their expertise. But in the expert witness case, there are such requirements. The model for admitting evidence—including testimony—during a trial is an opt-in model. That is to say, evidence is presumed inadmissible unless it meets the relevant criteria. In the case of expert witness testimony, those criteria include that the evidence is relevant and that the testimony is delivered by a properly qualified expert. For a juror who is aware of this, the testimony of an expert witness on a matter that requires expertise will always be evidence that the testimony is that of an expert. Such evidence of expertise cannot be completely eliminated by qualification in court, even if it can be in many other contexts.

In sum, there are strong epistemic reasons to ban even qualified epistemic trespassing by expert witnesses. These reasons include that such testimony is more likely to be unreliable than testimony delivered by actual experts and that such testimony will generate misleading higher-order evidence about the strength of the testimony, even if the testifier qualifies their testimony in order to try to indicate that they do not have expertise about the matter for which they are testifying. The ineffectiveness of such qualification is rooted in the rules about when testimony is admissible at trials and the social conditions such rules and knowledge of them creates.

These epistemic reasons not to epistemically trespass as an expert witness give rise to moral reasons not to trespass. First, given the significant stakes that typically accompany the outcome of a trial, the creation of misleading evidence at trial can have severe consequences. Think, for example, of Weeks's conviction for attempted murder based on expert trespassing testimony. Second, epistemic trespassing by expert witnesses shows disrespect to the epistemic agency of the jurors who are tasked with the weighty challenge of issuing just and epistemically defensible verdicts. Epistemic trespassing by expert witnesses undermines the ability of jurors to perform their civic duty as well as possible. Third, as noted earlier, the admission of testimony at trial occurs on an opt-in basis. It is a privilege, not a right, to testify as an expert witness in a trial. It is an abuse of that privilege to engage in epistemic trespassing while on the stand as an expert witness. It is also fundamentally unfair because it bypasses the normal restrictions on who is given the privilege of testifying before the jury.

In summary, while some cases of expert trespassing testimony may be permissible so long as the expert appropriately qualifies their statement, this is not true in the context of expert witness testimony. Rather, there are important epistemic and moral reasons why expert witnesses should not make even qualified claims that go beyond the scope of their expertise.

4. How Judges and Litigators Can Identify and Address Epistemic Trespassing

In a perfect world, experts would closely monitor themselves and scrupulously avoid epistemic trespassing. But in a legal system like that in the United States, where expert witnesses are paid by the parties for their testimony, experts may be tempted to engage in epistemic trespassing in order to provide testimony to the liking of those paying for the testimony. As a result, judges, litigators, and jurors should watch for epistemic trespassing.

In this section, I aim to do three things. First, I provide more detail about

what I mean by expertise. Second, I appeal to the philosophical literature on layperson recognition of experts to suggest strategies that judges, lawyers, and jurors can use to more reliably identify epistemic trespassing. Third, I offer some recommendations for how judges and litigators should respond to epistemic trespassing. I then consider what jurors ought to do about epistemic trespassing.

4.1. Describing Expertise and Epistemic Trespassing

Because epistemic trespassing is defined in terms of judging matters outside one's field of expertise, we need an understanding of what constitutes a field of expertise.[56] The first thing to notice is that no one is an expert writ large. Expertise is always indexed to a particular area of expertise. Following Gerken, I will generally refer to an area of expertise as a domain. Thus, the relevant question in determining whether someone is a relevant expert is never simply "Is *S* an expert?" but always "Is *S* an expert in domain *D*?"

For our purposes, the boundaries of a domain are determined by a fixed set of related questions or topics.[57] Sometimes a domain's scope will be coextensive with the scope of an academic discipline (e.g., physics, sociology, Russian studies). Often it will be narrower. For example, in Gerken's case of the cognitive psychologist specializing in color vision, her expertise in that area of psychology does not provide her with the requisite level of expertise to testify as an expert witness on matters within the domains of social or developmental psychology.[58] In other cases, the relevant domain of expertise may cut across related disciplinary lines. For example, specialized knowledge regarding the best soil conditions for growing crops is a domain of expertise whose experts may occupy positions within the academic disciplines of either agronomy or botany.[59]

Treating "field" and "domain" as interchangeable, I follow Ballantyne in holding that someone counts as an expert only if they possess "first, enough relevant evidence to answer reliably or responsibly their field's questions; and, second, enough relevant skills to evaluate or interpret the field's evidence well."[60] Here

56 See Ballantyne, "Epistemic Trespassing," 370–71.

57 See Ballantyne, "Epistemic Trespassing," 370.

58 See Gerken, "Expert Trespassing Testimony and the Ethics of Science Communication," 301.

59 Ballantyne makes a similar point about unclear boundary lines between disciplines using the example of biochemistry and molecular biology ("Epistemic Trespassing," 370–1).

60 Ballantyne, "Epistemic Trespassing," 371. This description of expertise in terms of relevant (i) evidence and (ii) skills aligns well with Elizabeth Anderson's description of assessments of expertise as about whether purported experts "have access to the evidence and the skills to evaluate it" ("Democracy, Public Policy, and Lay Assessments of Scientific Testimony," 145).

are three points to note about this definition. First, this description of expertise aligns with the Federal Rules of Evidence in that it incorporates a reliability condition into the necessary criteria for expertise. Second, in order to possess the kinds of evidence and skills required for expertise, one will, as a practical matter, typically require specialized experience, training, or education. Such experience, training, or education is usually verifiable using documentary evidence. Third, epistemic trespassing can come in degrees because one can be more or less reliable or responsible in answering a domain's questions and can have more or less relevant skill in evaluating a domain's evidence.[61]

For example, say that a jury would benefit from expert witness testimony about chronic diseases of the intestinal tract. Having an MD who practices family medicine testify rather than a gastroenterologist might constitute a small degree of epistemic trespassing. Having a brain surgeon testify might constitute a greater degree of epistemic trespassing than having either the family medicine doctor or the gastroenterologist testify. But having someone with no medical training at all testify would, all else equal, constitute a far greater degree of epistemic trespassing than either the brain surgeon's testimony or the family medicine doctor's testimony.

We can think of the severity of epistemic trespassing as falling on a spectrum. At one end of the spectrum is subtle epistemic trespassing. Subtle epistemic trespassing occurs when the trespasser, while lacking expertise in the most relevant domain, has expertise in a related domain that is likely to make the trespasser more reliable than the average layperson while still less reliable than the average true expert. At the other end of the spectrum is egregious epistemic trespassing. Egregious epistemic trespassing occurs when any expertise that the trespasser may have is so far removed from the relevant area of expertise that the trespasser is no more reliable than the average layperson. Moderate cases of epistemic trespassing fall in between subtle and egregious cases of trespassing.

As a general matter, the more egregious the trespass, the more likely it is that the testimony will fail to be reliable. Consider that while both the brain surgeon and I may be engaging in epistemic trespassing if testifying as an "expert witness" about complex medical facts regarding a plaintiff's colon, the brain surgeon is still a far more reliable source of information than I am, all else equal, given that I have no training in medicine. The good news is that it is also the case that, as a general

61 There is a reading of Ballantyne's description of epistemic trespassing whereby the presence of the word "enough" in both clauses transforms the definition into an in-or-out category that does not come in degrees. But I do not think such a reading is best suited to thinking about expertise in the context of expert trespassing testimony. It seems to me that "enough" is likely a context-sensitive property. But even if "enough" is not taken to be context sensitive, one can be closer or further from having enough of property *X*, and thus closer or further from being an expert.

matter, the more egregious the epistemic trespass, the easier it will be to spot. Thus, even if the suggestions I offer in the next subsection make identifying subtle instances of epistemic trespassing challenging, I argue that they are often sufficient to catch most instances of moderate and egregious epistemic trespassing, which will generally be the instances where the most unreliable testimony is offered.

4.2. Guidelines for Identifying Epistemic Trespassing

In the last few decades, multiple scholars have theorized about laypeople's abilities to assess expert testimony.[62] Of concern across these readings is the issue of how optimistic we should be about lay assessments of expert testimony. The level of optimism shown by these writers is mixed.[63] But when it comes to applying the methods put forward by these authors to the issue of epistemic trespassing by expert witnesses, we have some reason to be optimistic. This is because the question one needs to ask in order to spot an epistemic trespasser—namely, is *S* an expert in domain *D*?—is much simpler to answer than questions about assessing the truth or other qualities of a purported expert's testimony that these other writers are primarily concerned with.

Assessing whether someone is an epistemic trespasser does not require that one assess whether the epistemic trespasser's claims are true. Nor does it require that one assess which expert to trust in the face of competing expert testimony. This paper does not offer answers to questions such as Alvin Goldman's "Can novices, while remaining novices, make justified judgments about the relative credibility of rival experts?" or Elizabeth Anderson's "Given that ordinary citizens cannot directly assess [complex scientific] reasoning, does this call the democratic legitimacy of technical public policies in question?"[64] But it does appeal to the methods put forward by Goldman and Anderson for determining whether someone is an expert in a given domain.

In the remainder of this subsection, I argue that in many cases all one needs to determine whether or not a purported expert witness is engaging in, or plans

62 See, e.g., Brewer, "Scientific Expert Testimony and Intellectual Due Process"; Goldman, "Experts"; Anderson, "Democracy, Public Policy, and Lay Assessments of Scientific Testimony"; Lane, "When the Experts Are Uncertain"; Guerrero, "Living with Ignorance in a World of Experts"; Brennan, "Can Novices Trust Themselves to Choose Trustworthy Experts?"; and Watson, *Expertise*.

63 For example, Lane writes: "Can ordinary citizens in a democracy evaluate the claims of scientific experts? While a definitive answer must be made case by case, some scholars have sharply opposed general answers: a skeptical 'no' (e.g. Scott Brewer) versus an optimistic 'yes, no problem' (e.g. Elizabeth Anderson)" ("When the Experts Are Uncertain," 97).

64 Goldman, "Experts," 89; Anderson, "Democracy, Public Policy, and Lay Assessments of Scientific Testimony," 144.

to engage in, moderate or egregious epistemic trespassing on the witness stand is information about, in Goldman's terms, the purported expert's (1) credentials (as a form of "appraisals" by "meta-experts") and (2) track record.[65] Examples from these same categories show up in Anderson's suggestions of how to identify experts as well. Let us look at what each has to say in turn.

Regarding credentials, Goldman writes that credentials in the form of "academic degrees, professional accreditations, work experience and so forth (all from specific institutions with distinct reputations) reflect certification by other experts [of an expert's] demonstrated training or competence."[66] Thus, credentials such as academic degrees and professional certifications provide laypeople with something akin to testimonial evidence from other experts that the person with the credential is also an expert. This is attested to by phrases found on many university diplomas stating that degrees are conferred upon the recommendation of the faculty.

Goldman also suggests that we can use a purported expert's track record to gain information about their expertise. Goldman's primary concern is with adjudicating competing claims by experts, but the general idea of a track record as a metric of assessment can be transferred to determining whether one is an expert. The following is a non-exhaustive list of aspects of someone's track record that can be used to test for expertise:

- Has the purported expert published peer-reviewed research in a relevant area?
- Who is on the editorial board for the journals where the purported expert has publications? Are the editorial board members credentialed experts in relevant domains?
- Has the purported expert's work been relied on by other experts in the field (including by being cited by other experts)?
- If the purported expert is conducting scientific research, has their research been replicated or otherwise verified?
- Has the purported expert served on review panels or as part of professional committees?
- Is the purported expert a member in good standing of a relevant professional organization?
- Who, if anyone, has funded the purported expert's research?
- Has the purported expert received awards or other forms of recognition for their work?

65 Goldman, "Experts," 93.

66 Goldman, "Experts," 97.

In the context of a trial, a judge is able to require that evidence be provided about the purported expert's credentials and track record, and opposing counsel is able to question the purported expert about many of these things under oath. As a result, a trial is a context in which this information can be obtained and used as the basis for determining a purported expert's expertise.

Elizabeth Anderson brings together considerations of credentials and track record to construct a "hierarchy of expertise" as follows:

(a) Laypersons.
(b) People with a B.S. degree, a B.A. science major, or a professional degree in an applied science specialty far removed from the field of inquiry in question.
(c) Ph.D. scientists outside the field of inquiry.
(d) Ph.D. scientists outside the field, but with collateral expertise (for example, a statistician who is judging the use of statistics in the field).
(e) Ph.D. scientists trained in the field.
(f) Scientists who are research-active in the field (regularly publish in peer-reviewed scientific journals in the field).
(g) Scientists whose current research is widely recognized by other experts in the field, and whose findings they use as the basis for their own research. This can be determined by considering such factors as citation counts, the impact factors of the journals in which they publish, and record in winning major grants.
(h) Scientists who are *leaders* in the field—who have taken leading roles in advancing theories that have won scientific consensus or opened up major new lines of research, or in developing instruments and methods that have become standard practice. In addition to the factors cited in (g), leadership is indicated by election to leadership positions in the professional societies of the field, election to honorary scientific societies, such as the National Academy of Science, and receipt of major prizes in the field, such as the Nobel Prize.[67]

Anderson writes that "in general, the weight people should accord to others' testimony about a field increases as they go down the list, increasing especially steeply for categories (f), (g), and (h)."[68]

If we treat an MD as a credential similar to a PhD, Anderson's hierarchy provides a useful way of explaining the different levels of epistemic trespassing in

67 Anderson, "Democracy, Public Policy, and Lay Assessments of Scientific Testimony," 146–47.
68 Anderson, "Democracy, Public Policy, and Lay Assessments of Scientific Testimony," 147.

Weeks's trial for attempted murder. As a PhD in psychology, Paul Cameron is at level (c), although his expulsion from the American Psychological Association and that it is unclear whether his work at the time of the trial constituted scientific research count as marks against even this limited level of expertise. If we treat different medical specialties as equivalent to different fields, Lorraine Day is also at level (c). Mark Dowell, as a practicing MD in infectious diseases, is at level (e), but the record does not provide evidence that he was doing the kind of medical research that would bring him up to level (f). It is only the defense's witness, Richard Pollard, who reaches the levels of expertise that Anderson picks out as most significant. Like Dowell, Pollard is certified as a specialist in infectious diseases, which qualifies him for level (e). But Pollard's "research program active in both clinical and scientific research focused upon infections with HIV" satisfies level (f). In addition, that Pollard sat on a national panel which looks at all the drug studies conducted by the National Institutes of Health for the treatment of AIDS infections indicates that he is a leader in his field who may satisfy level (g) or (h).

Weeks's case is a good example of how the information needed to assess the scope and depth of a purported expert's expertise is provided as a matter of course during a trial. In a properly run trial, information about the credentials and track record of an expert witness should be available to judges, attorneys, and jurors. This information can be and should be used to assess whether expert witnesses are engaging in epistemic trespassing.

Weeks's case is also a good example of how more egregious epistemic trespassing is easier to identify than more subtle forms. If Dowell trespassed at all, it was fairly subtle epistemic trespassing. Such trespassing may not be easily identifiable by judges, attorneys, or jurors—most of whom will be laypeople in reference to the relevant area of expertise most of the time. The more closely related different areas of expertise are, the harder it will be for nonexperts to distinguish them. But one need not have any kind of special skills to recognize the difference between orthopedic surgery and infectious diseases or between psychology and medicine. Thus, while it may remain challenging for laypeople to identify subtle epistemic trespassing, in the context of a trial they are often well-positioned to get the information they need to identify moderate and severe forms of epistemic trespassing.

The good of using credentials and track records to identify epistemic trespassers comes with a potential downside. If these methods are taken too far and the standards set too high, qualified experts may be denied the opportunity to testify in court as an expert witness. If courts set the bar too high, we may encounter an access-to-justice problem whereby it becomes too difficult for many litigants and defendants to obtain the help of qualified experts. Thus, vigilance is required in both directions. The Supreme Court offers a useful guiding prin-

ciple that can help navigate between these two problematic extremes. This principle is the one we saw earlier—namely, that an expert should employ "in the courtroom the same level of intellectual rigor that characterizes the practice of an expert in the relevant field."[69] In order to testify as an expert witness, one need not be at the top of their field. They need not be exceptional among the experts. Rather, they need to be able to apply the *same* level of intellectual rigor as *an* expert in the relevant field. This standard does not eliminate the need for good judgment in application, but no legal standard does. It is enough to guide a competent decision maker in avoiding blatant epistemic trespassing and blatant elitism in determining who can testify as an expert witness at trial.

4.3. The Roles of Judges and Trial Lawyers in Preventing Epistemic Trespassing

Supreme Court precedent holds that judges serve a "gatekeeping role" in determining who is allowed to testify as an expert witness at trial.[70] This means, among other things, that judges have a responsibility to determine whether those who are requesting to testify as expert witnesses are in fact experts about the matters for which they seek to testify. Importantly, the judge's gatekeeping role is not about assessing an expert's conclusions.[71] Thus, the distinction made in the previous subsection—between determining whether someone is an expert and determining whether to trust or believe the expert's testimony—is relevant. When acting as gatekeepers, judges should do only the former, not the latter.

Appellate judges are also charged with a gatekeeping role of sorts. In *General Electric Co. v. Joiner*, the Supreme Court held that trial judges' decisions about whom to allow to testify as an expert witness are subject to an abuse of discretion standard.[72] On this standard, trial judges' decisions about whom to admit as expert witnesses are given a great deal of deference. But this deference is not complete. In cases where the appellate court finds that the trial judge's ruling about whether to admit an expert witness was clearly erroneous, the appellate court can reverse the trial judge's decision.[73]

Judges, like other laypeople, have limited knowledge from which to determine whether someone is engaging in epistemic trespassing. However, as argued in the previous subsections, many laypeople are capable of identifying moderate and egregious cases of epistemic trespassing just by looking at credentials and

69 Kumho Tire Co. v. Carmichael, 526 U.S. 137, 152 (1999).

70 Daubert, 509 U.S. at 597.

71 Daubert, 509 U.S. at 594–95.

72 See General Electric Co. v. Joiner, 522 U.S. 136 (1997).

73 For discussions of the abuse of discretion standard see Graham, "Abuse of Discretion, Reversible Error, Harmless Error, Plain Error, Structural Error"; and Ryan "Backfire."

track record. If we return to the case of Curtis Weeks, a judge who was faithfully carrying out the role of gatekeeper should easily have been able to determine that allowing Cameron and Day to testify as "experts on HIV" would be to permit egregious epistemic trespassing. As such, Cameron and Day should have been denied access to the witness stand.

As Weeks's trial shows, trial judges sometimes fail in their role as gatekeepers for expert testimony. But there are steps that litigators can take to try to combat epistemic trespassing as well. First, litigators should avoid seeking to have epistemic trespassers admitted as experts in court. Had the prosecution in Weeks's case been more scrupulous about whom it put forward as an expert, the issue of epistemic trespassing would never have arisen in the first place.

Second, if an opposing party procures a spot on the witness stand for an epistemic trespasser, litigators should seek to make clear to the jury that the purported expert is out of their depth and, as a result, both less reliable and less trustworthy. This should be done by vigorous cross-examination of the expert about the scope of their expertise. This is also something that perhaps at times can be usefully woven into the narrative framing of the case.

Finally, attorneys should treat epistemic trespassing as grounds for appeal. If, as in Weeks's case, epistemic trespassing was required to provide evidence for an otherwise undefendable aspect of the case, attorneys ought to appeal for that reason.

In this section, we covered what epistemic trespassing is, how one can identify it, and what judges and attorneys should do in response. In the final section, I consider the perspective of jurors. Specifically, I discuss three ways in which identifiable epistemic trespassing ought to shape a juror's view of the evidence presented.

5. JURIES AND THE EVIDENTIAL SIGNIFICANCE OF EPISTEMIC TRESPASSING

In this section, I argue that recognizable instances of epistemic trespassing by an expert witness provide jurors with higher-order evidence that epistemically weakens the case of the trespassing party. More specifically, I argue that recognizable instances of epistemic trespassing provide jurors with evidence that:

1. The trespasser is unreliable.
2. The trespasser is untrustworthy.
3. At least some of the assertions of the epistemic trespasser are assertions that genuine experts are unwilling to make.

The notion of evidence I am using here is one where evidence counts in favor of

that for which it is evidence, all else equal. Such evidence can be undermined, rebutted, or outweighed. Thus, one can recognize that there is evidence for a position while still rationally rejecting that position. With that notion of evidence in mind, let us examine each of the three claims.

5.1. Epistemic Trespassing as Evidence of Unreliability

This first principle follows from considerations already covered in this paper and in Ballantyne's initial argument against epistemic trespassing. When we trespass, we become less reliable. This is because when we trespass, we move into an area where we do not have the skills or experience needed to assess evidence and generally will not reach conclusions as accurately as true experts would. If as a juror you encounter an instance of recognizable epistemic trespassing, you gain higher-order evidence that the trespasser is less likely to be a reliable source of information.

5.2. Epistemic Trespassing as Evidence of Untrustworthiness

There is a relationship between reliability and trust. Those who are unreliable are, all else equal, less trustworthy than those who are more reliable. Thus, jurors already gain some reason to distrust epistemic trespassers because they are less reliable than true experts. But the unreliability of epistemic trespassers is only part of the evidence jurors gain that trespassing expert witnesses are untrustworthy.

In addition, jurors can reason from the starting point that those most deserving of our trust will be careful to avoid epistemic trespassing. Epistemic trespassers show that they either lack self-awareness about the limits of their own expertise or lack a commitment to honest representation of their own expertise. Both of those criteria give us a reason to downgrade the trustworthiness of the trespasser. If they trespass because they lack self-awareness of their limits, we gain evidence that the trespasser may be confident about other conclusions for which they should not be confident. If they trespass because they lack a commitment to honest representation of their expertise, this provides us with evidence about their moral character. If they are willing to lie or exaggerate about their expertise, we can rationally conclude that they are more likely to lie or exaggerate in giving testimony. Thus, epistemic trespassing is a sign not only of the unreliability of the trespasser, but of their untrustworthiness as well.

5.3. Epistemic Trespassing as Evidence of a Weak Position

The first two kinds of evidence I have argued jurors gain by observing recognizable epistemic trespassing are about the credibility of the expert witness. The third kind of evidence comes from the trespasser's role as part of the larger legal

strategy of one of the parties. This third kind of evidence rests on two assumptions. First, attorneys generally seek to present stronger cases rather than weaker ones. Second, the expert testimony of a true expert in domain *D* better contributes to the strength of a case than the "expert testimony" of an epistemic trespasser in domain *D*, all else equal.

If we think about what a strategic attorney would do in selecting an expert witness, we can see how this can provide the jury with evidence about the strength of a party's case. Attorneys will want to select witnesses that the jury will find convincing. Given the potential for jurors to spot clear cases of epistemic trespassing, attorneys are reasonable in concluding that a true expert will, all else equal, be a more convincing witness than a trespasser. Thus, attorneys will prefer true experts to trespassers, all else equal.

From the perspective of a juror, if an attorney presents you with an epistemic trespasser rather than a true expert, this provides you with some reason to think that the attorney failed in obtaining their preferred outcome of having a true expert assert the things that the trespasser is asserting instead. But if the attorney failed to find a true expert who would be willing to assert the claims that the trespasser is asserting, this makes it more likely that true experts are unwilling to assert at least some of what the trespasser is asserting. And if true experts are unwilling to assert at least some of what the trespasser is asserting, this is evidence that the trespasser's claims are not shared (or at least not shared widely) by true experts. This gives jurors a reason to think that the case of the party with the epistemic trespasser is weak concerning at least some matters attested to by the trespasser.

These ways in which expert trespassing testimony provides jurors with rational reason to view the trespassing party's case as weak would, if widely recognized by jurors, provide attorneys with reasons not to bring in expert witnesses who engage in epistemic trespassing. The fact that attorneys sometimes do put epistemic trespassers on the stand suggests that jurors' wariness of epistemic trespassing is not so strong as to consistently disincentivize the behavior at present. But that incentive can be strengthened if jurors become more confident in identifying and penalizing epistemic trespassing by expert witnesses. That said, it is easy to understand why jurors might be hesitant to identify expert witnesses as epistemic trespassers given that such trespassers have been treated by the court as genuine experts. As noted earlier, courts create misleading higher-order evidence when they permit "expert witnesses" to engage in epistemic trespassing. Thus, in comparison to judges and litigators, jurors are at an epistemic disadvantage when it comes to identifying and accounting for epistemic trespassing. As a result, judges and litigators are generally better positioned to identify and prevent epistemic trespassing.

6. CONCLUSION

In this paper I have examined the issue of epistemic trespassing by expert witnesses in a court of law. I have argued for each of the following. Expert witnesses should avoid making any trespassing claims. Merely qualifying such claims to indicate one's lack of expertise is not enough. Judges, litigators, and jurors can identify many instances of moderate and egregious epistemic trespassing by examining a purported expert's credentials and track record. Jurors have reason to treat recognizable instances of epistemic trespassing as counting against the strength of the trespassing party's position, but this can be offset by the misleading evidence that the epistemic trespasser has been permitted by the court to trespass as an expert witness. Judges and lawyers are better positioned than jurors to address and prevent epistemic trespassing by expert witnesses. Judges should not permit epistemic trespassers to testify as expert witnesses. Litigators should expose epistemic trespassers during cross-examination. Such steps would increase our ability to use trials as a way of achieving justice.[74]

Wayne State University
mark.satta@wayne.edu

REFERENCES

Anderson, Elizabeth. "Democracy, Public Policy, and Lay Assessments of Scientific Testimony." *Episteme* 8, no. 2 (June 2011): 144–64.

Ballantyne, Nathan. "Epistemic Trespassing." *Mind* 128, no. 510 (April 2019): 367–91.

Brennan, Johnny. "Can Novices Trust Themselves to Choose Trustworthy Experts? Reasons for (Reserved) Optimism." *Social Epistemology* 34, no. 3 (2020): 227–40.

Brewer, Scott. "Scientific Expert Testimony and Intellectual Due Process." *Yale Law Journal* 107, no. 6 (Aril 1998): 1540–1681.

Cresswell, F. V., J. Ellis, J. Hartley, C. A. Sabin, C. Orkin, and D. R. Churchill. "A

74 Thank you to Jonathan Beever, Ken Boyd, Mikkel Gerken, Will Ortman, Chad Sinckler, Greg Stoutenburg, Kate Thoreson, James Toomey, two anonymous referees for this journal, and audience members at a talk given to the Philosophy Department and the Center for Ethics at the University of Central Florida for helpful feedback on earlier versions of this paper.

Systematic Review of Risk of HIV Transmission through Biting or Spitting: Implications for Policy." *HIV Medicine* 19, no. 8 (September 2018): 532–40.

Gerken, Mikkel. "Expert Trespassing Testimony and the Ethics of Science Communication." *Journal for General Philosophy of Science* 49 (2018): 299–318.

Goldman, Alvin. "Experts: Which Ones Should You Trust?" *Philosophy and Phenomenological Research* 63, no. 1 (July 2001): 85–110.

Graham, Michael H. "Abuse of Discretion, Reversible Error, Harmless Error, Plain Error, Structural Error: A New Paradigm for Criminal Cases." *Criminal Law Bulletin* 43, no. 6 (November–December 2007): 955–80.

Guerrero, Alexander. "Living with Ignorance in a World of Experts." In *Perspectives on Ignorance from Moral and Social Philosophy*, edited by Rik Peels, 156–85. New York: Routledge, 2016.

Haack, Susan. "The Expert Witness: Lessons from the U.S. Experience." *Humana.Mente Journal of Philosophical Studies* 8, no. 28 (2015): 39–70.

Lane, Melissa. "When the Experts Are Uncertain: Scientific Knowledge and the Ethics of Democratic Judgment." *Episteme* 11, no. 1 (March 2014): 97–118.

Mueller, Christopher B., and Laird C. Kirkpatrick. *Federal Rules of Evidence with Advisory Committee Notes and Legislative History.* New York: Wolters Kluwer, 2017.

Ryan, Sean. "Backfire: Abandoning the Abuse of Discretion Standard of Review for Daubert Rulings Shoots Trial Courts in the Foot." *University of Toledo Law Review* 47, no. 2 (2016): 349–72.

Siegel, Max. Letter from Max Siegel on behalf of the American Psychological Association to Paul Cameron. December 2, 1983.

Watson, Jamie Carlin. *Expertise: A Philosophical Introduction*. London: Bloomsbury Academic, 2020.

Journal of Ethics and Social Philosophy
VOL. 22, NO. 2 · JULY 2022
https://doi.org/10.26556/jesp.v22i2.1633

PRIVACY RIGHTS FORFEITURE

Mark L. Hanin

Consider three scenarios: (i) a couple absentmindedly leaves its windows open during a loud fight, making it easy for neighbors or passersby to overhear the couple; (ii) a person repeatedly fires off sensitive emails despite mistyping the recipients' addresses and never taking time to double check them; (iii) an actress voluntarily places her DNA and name online while wanting to keep private a rare genetic disorder that she knows about.[1] Do the agents in these scenarios retain a moral right to privacy in their fight, emails, and disorder, respectively?

Suppose you think not. That is, presumably, because some or all of those agents have *forfeited* a right to privacy—no matter their intentions or protestations to the contrary. While *waiver* is a relatively straightforward notion that involves giving up a moral or legal entitlement in some voluntary way, *forfeiture* is a murkier concept that operates in spite of an agent's intentions. What are its normative foundations? And is it possible to specify conditions under which privacy rights can be forfeited?

I take up these questions here and propose a novel theory of privacy rights forfeiture. The theory takes its inspiration from Judith Thomson's canonical paper "The Right to Privacy." Thomson argues that privacy rights can be waived both intentionally and "unintentionally."[2] Regrettably, however, Thomson sows confusion by failing to distinguish clearly between waiver and forfeiture and by repeatedly speaking of "unintentional waiver" when it seems clear that *forfeiture* is really at issue. Still, taking a cue from Thomson's work, I will develop an account of privacy-rights forfeiture in the bulk of the paper.

The account, in brief, is as follows. Agents may forfeit a right to privacy in two main ways rooted in negligent or reckless conduct as it concerns their privacy interests. Yet forfeiture is not merely a normative consequence of acting in detriment solely to one's *own* interests. A doctrine along those lines would

1 For the first scenario, see Thomson, "The Right to Privacy," 306. For the third, see Rumbold and Wilson, "Privacy Rights and Public Information," 14–15.

2 Thomson, "The Right to Privacy," 302.

be too punitive. Rather, I will argue that considerations about unfairness to putative duty-bearers come into play, as well. Thus, I will defend a hybrid model of necessary and sufficient conditions for privacy forfeiture that includes both self-directed and other-directed considerations.

Toward the end of the paper, I will address some contrary views articulated in a recent article by Benedict Rumbold and James Wilson ("RW" for short). RW engage at length with Thomson's work and reject her idea that agents can be divested of privacy rights "unintentionally" (though RW, like Thomson, rarely speak of forfeiture itself).[3] I will respond to some of RW's criticisms and argue that RW's forfeiture-free model of privacy rights is unconvincing on moral grounds.

There is one issue that I should flag and set aside. RW's article is motivated by a narrower topic—privacy rights over *inferences*. For example, in the digital context, may someone who posts photos on Instagram legitimately assert a moral privacy right in certain inferences that can be drawn from those photos? Such inferences, after all, can yield highly sensitive information about one's health and personality, with one study showing that the choice of filter for Instagram photos can predict indicators of depression.[4] RW explore foundational issues about privacy in order to vindicate the reality and importance of what I will call *inferential privacy rights*.

I agree with RW that inferential privacy is an urgent and under-theorized subject. More important, I agree that such rights exist and can impose genuine moral constraints on what may be done with agents' information. In a nutshell, I believe that waiving or forfeiting privacy rights to certain facts—for example, one's public social media posts—does not *entail* waiving or forfeiting privacy rights over all possible inferences that can be drawn from that information, especially unpredictable and sensitive inferences.[5] The account of forfeiture I will develop in this paper offers new normative resources to assess when agents may assert inferential privacy rights and when those rights have been waived or forfeited. This paper, however, mostly focuses on privacy forfeiture in general rather than on the narrower subject of inferential privacy.

With respect to the normative foundations of privacy, I will aim to be ecu-

3 See Rumbold and Wilson, "Privacy Rights and Public Information," 6, 12–13. Rumbold and Wilson refer to forfeiture in a footnote ("Privacy Rights and Public Information," 15n32), and Thomson mentions it once in passing ("The Right to Privacy," 302).

4 See Reece and Danforth, "Instagram Photos Reveal Predictive Markers of Depression"; see also Wachter and Mittelstadt, "A Right to Reasonable Inferences," 505–14; cf. Barocas and Nissenbaum, "Big Data's End Run around Anonymity and Consent."

5 So, for example, I broadly agree with RW's criticism (though not their reasoning) of the Radar app that sought to infer facts about Twitter users' mental health from public posts without users' permission. See Rumbold and Wilson, "Privacy Rights and Public Information," 3–6, 24–25.

menical for purposes of this paper. I disagree with Thomson that privacy rights lack a unifying justification, and I am broadly sympathetic (as are RW) to Andrei Marmor's notion that privacy rights give us a reasonable degree of control over how we present ourselves to others.[6] But, with modifications, my account of forfeiture can be made consistent with theories of privacy that privilege control, intimacy, or contextual integrity.

I first develop a novel taxonomy of how moral and legal entitlements can be divested in general, focusing on waiver and forfeiture (section 1). I then define two species of forfeiture rooted in negligent and reckless conduct (section 2). Since those definitions are formulated at high levels of abstraction, I set out five application criteria for applying those definitions to specific cases (section 3). In particular, I will explore a puzzle about how different sensitivity levels of private information influence forfeiture (section 4). I will then engage with RW's account. I first consider an objection to my view having to do with distinct normative thresholds for forfeiting privacy rights and property rights in tangible goods (section 5). I will then argue that RW's forfeiture-free model of privacy *over*protects privacy rights by leaving out forfeiture and *under*protects privacy rights, particularly those of minorities and idiosyncratic agents (section 6). A brief conclusion follows (section 7).

1. DIVESTMENT OF ENTITLEMENTS: A TAXONOMY

Agents can be divested of entitlements in a variety of ways. This section offers a novel approach to mapping this conceptual terrain with a focus on waiver and forfeiture.

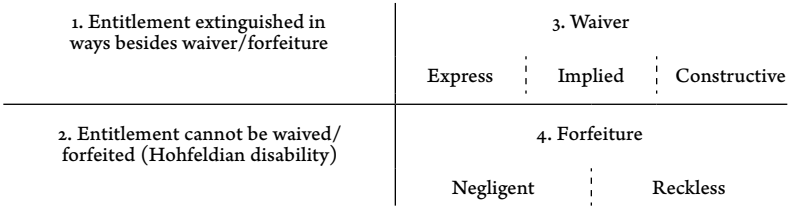

FIGURE 1 Divestment of Moral or Legal Entitlements: Four Modalities

Consider four modalities by which agents can be divested of moral or legal entitlements generally (figure 1), including first-order Hohfeldian entitlements (e.g., claim-rights and liberties) as well as higher-order entitlements (e.g., powers and

6 See Marmor, "What Is the Right to Privacy?" Marmor's theory, it seems to me, has trouble accounting for privacy violations in digital contexts that do not involve any human beings coming to know relevant private information, only machines. I will not develop this point here.

immunities). I will briefly address modalities 1 and 2 and set them aside. Agents can be divested of entitlements in ways other than via waiver and forfeiture (modality 1). That is, *P* may exercise a Hohfeldian power to modify *Q*'s entitlements in ways that need not depend in any direct way on *Q*'s acts or omissions. For example, the government may weaken or rescind a benefit-conferring law involving unemployment benefits or social insurance. Or parents can modify household rules by eliminating entitlements that their children had previously enjoyed. At the other extreme, so to speak, some entitlements cannot be waived or forfeited at all, so that agents have a Hohfeldian disability with respect to alienating them (modality 2). Under US law, for example, a litigant's right to contest subject-matter jurisdiction "can never be waived or forfeited."[7] Likewise, a criminal defendant's right to invoke certain constitutional defenses in post-trial proceedings cannot be waived or forfeited even following a guilty plea.[8] As to the moral realm, think of foundational human rights or rights that are correlative with self-directed duties of a Kantian sort, e.g., a duty of self-respect or a duty to protect one's own privacy.[9] Such rights, if they exist, cannot be alienated.

In the remainder of this section I will focus on waiver and forfeiture (modalities 3 and 4). In a case of *waiver*, an agent is divested of some Hohfeldian entitlement(s) in virtue of that agent's actual intentions, underlying attitudes or dispositions, or imputed intentions. I will distinguish below among three species of waiver—express, implied, and constructive—and consider how they manifest in both law and morality. While the terms I use appear in US law, I am defining them independently despite certain overlaps.[10]

Express waiver: When an agent indicates orally or in writing an intent to give up an entitlement, the agent expressly waives that entitlement. A defendant might tell a judge that she declines to assert her Fifth Amendment privilege against self-incrimination and agrees to testify at trial. Or a defendant may renounce a right to government counsel, to challenge extradition, to object to the introduction of certain evidence, and so forth. These and like cases are straightforward instances of express waiver.

Implied waiver: Waiver can also occur short of such express statements. Implied waiver may simply describe a situation in which an agent forms a clear intent to waive but does not articulate it openly. As a prosaic illustration, if I walk away from a self-checkout stand without taking my receipt, anticipating that it will shortly be thrown out by store staff, I may impliedly be waiving my right to

7 Gonzalez v. Thaler, 565 U.S. 134, 141 (2012).

8 See, e.g., Blackledge v. Perry, 417 U.S. 21 (1974); see also Westen, "Away from Waiver."

9 See Allen, "Protecting One's Own Privacy in a Big Data Economy."

10 See, e.g., Berg, "Understanding Waiver."

the receipt. Now consider two legal examples, keeping in mind that the law generally favors robust awareness and voluntariness conditions for waiver. If, after being given a lawful *Miranda* warning, a suspect voluntarily opts to talk with the police, the suspect may impliedly waive certain rights by initiating the conversation. Likewise, a party may impliedly waive the right to contest personal jurisdiction by omitting that objection from its initial answer to a complaint.

Implied waiver can also occur in a subtler way, which would be disfavored in law. I have in mind cases in which an agent (i) *has not* formed a clear intent at time *t* to waive an entitlement, but (ii) were she asked (at *t* or some later time) about her wishes at *t*, she would express an intent to waive. To illustrate, imagine that *P* is hanging up pictures in her new office or home without thinking much of it. When asked by *Q*, "So, you're giving up your privacy interest in these photos in relation to all your visitors?" *P* replies, "Yes, I suppose I am."

Constructive waiver: Finally, constructive waiver *imputes* to agents an intent to waive an entitlement irrespective of their mental states and modes of conduct. In law, agents are sometimes "deemed to have consented" to certain consequences if they ϕ.[11] The assumption, I take it, is that rational agents will ϕ only if they believe that doing so will advance their interests. In that case, they can be assumed to consent to relinquishing the specified entitlements by ϕ-ing. So whereas in implied waiver an agent actually intends (or would intend on reflection) to waive a right, but without expressly saying so, constructive waiver can occur even if an agent has no intention either way (for example, because she is unaware of the legal consequences of ϕ-ing) or intends *not* to relinquish her entitlement(s) (for example, if she wrongly believes that the legal provision is *ultra vires*).

The conditions outlined above are necessary but *insufficient* for each form of waiver to occur. That is because we must also account for potential defeating conditions.[12] An agent who tries to waive an entitlement could fail because, for example, (i) she does not actually have the entitlement, (ii) the entitlement is not alienable, or (iii) she has not evinced the right set of mental states and/or communicative acts. As for constructive waiver, which can occur without an agent's awareness, a defeating condition would arise if the provision at issue were unlawful.

How do the three types of waiver introduced above apply to the moral sphere? Express waiver is, again, straightforward. If *P* makes a promise to *Q* and *Q* later says that *P* does not need to abide by it, *Q* expressly waives the right to performance. The same will be true of any overt relinquishment of an entitlement,

11 See, e.g., 18 U.S.C. § 2334(e)(1); 25 C.F.R. § 162.458(b)(2); see also Parden v. Terminal R. Co., 377 U.S. 184, 192 (1964).

12 I am grateful to Laura K. Donohue for bringing this point to my attention.

assuming no defeating conditions. Agents can also impliedly waive moral rights. The pragmatics of the situation can convey waiver without requiring overt articulation. When *A* tells *B* a secret in ordinary circumstances, *A* impliedly waives her privacy right in that secret vis-à-vis *B* without having to say so directly. Or, as noted earlier, if *P* hangs photos in an office or home, *P* may impliedly waive privacy rights to those photos vis-à-vis *P*'s visitors. As to constructive waiver, I do not see this concept playing a major role in moral life. It is relevant, however, in social contract theories of political morality. On such views, actual agents may be said constructively to waive certain rights in virtue of choices made by hypothetical representative persons in something like an Original Position.

Finally, let us consider forfeiture. In a case of *forfeiture*, an agent is divested of some Hohfeldian entitlement(s) in a way that typically, though not invariably, fails to align with that agent's actual intentions or relevant attitudes or dispositions. Forfeiture typically damages, rather than advances, an agent's interests. And it usually manifests in some form of negligent or reckless conduct, a point I develop below in relation to privacy. (The preceding caveats—"not invariably," "typically," "usually"—are needed to leave room for cases of *intentional forfeiture*. For example, in a case of "suicide by cop," an agent may take deliberate steps to forfeit a legal right not to be intentionally killed.[13] Since that legal right is non-waivable, waiver cannot account for its divestment.)[14] As a final point, while the justificatory grounds of forfeiture will vary based on the circumstances, they will often (though not always) involve fairness to third parties.

In law, forfeiture comes in many stripes. Litigants can forfeit their right to rely on certain claims or defenses by failing to assert them in a timely manner—for example, by failing to plead an affirmative defense under Rule 8(c) of the Federal Rules of Civil Procedure or by leaving out an argument from an opening appellate brief.[15] And under Federal Rule of Evidence 804(b)(6), which used to be called "forfeiture by wrongdoing," a defendant can forfeit an immunity against the use of hearsay evidence if that defendant had played a role in the "declarant's

13 I am grateful to Matthew H. Kramer for flagging this sort of case.

14 So the difference between waiver and forfeiture does not lie in the fact that the former involves voluntary divestment of an entitlement, whereas the latter does not. That distinction is overly simplistic. There are cases of forfeiture in which an agent intentionally seeks to forfeit an entitlement (e.g., "suicide by cop"), and there are instances of waiver in which an agent does not intend to give up an entitlement (e.g., certain cases of constructive waiver).

15 See, e.g., Maalouf v. Islamic Republic of Iran, 923 F.3d 1095, 1107 (DC Cir. 2019); Al-Tamimi v. Adelson, 916 F.3d 1, 6 (DC Cir. 2019).

unavailability as a witness, and did so intending that result." [16] In each of these cases, forfeiture serves to ensure fair treatment of opposing parties.

In morality, forfeiture is also commonplace. A parent who abuses a child can forfeit moral liberty-rights and claim-rights to care for that child. A business partner who makes reckless and self-serving decisions can forfeit moral entitlements to run the business. An athlete who dopes may forfeit the moral privilege to compete (even if no one finds out). A craven politician can forfeit moral rights to govern. In these examples, forfeiture is, again, justified in part based on unfairness or disrespect toward others. But forfeiture can also come about just by violating rules imposed with legitimate authority. For example, a teenager who negligently or recklessly comes home late—despite her parents' warning that this may expose her to added chores next week—will forfeit her typical immunity against added housework over and above her weekly allotment.

Paying attention to mental states will often be crucial to classifying cases accurately. For, the same conduct can be consistent *either* with forfeiture *or* implied waiver. To illustrate, imagine that Javier invites friends over to watch a sports game. Walking toward the TV room in his house, they notice a large open wall safe with baseball memorabilia and an old watch. Assuming that Javier himself left the safe ajar, at least two interpretations are possible. Either Javier could not care less that his friends will see what is inside, suggesting implied waiver of his right to privacy in the safe's contents, or Javier forgot to shut the door, suggesting that he negligently forfeited his right to privacy. To choose between these readings, we need to know about Javier's mental states. And the same will be true in countless other cases where mental states will be decisive in distinguishing between forfeiture and implied waiver.

With the preceding taxonomy in mind, I will narrow my focus to forfeiture and, more particularly, to forfeiture of privacy rights.

2. PRIVACY RIGHTS FORFEITURE

Here and in the following two sections I introduce my account of privacy rights forfeiture. In doing so, I heed Massimo Renzo's insightful criticism of forfeiture theories of punishment. Renzo argues that merely adverting to forfeiture cannot itself explain *why* that upshot is justified. Other normative concepts must step in.[17] With that admonition in mind, I suggest two main ways in which a claim-right to privacy can be forfeited:

16 See US Federal Rule of Evidence 804(b)(6), advisory committee's notes (1997 amendment).

17 Renzo, "Rights Forfeiture and Liability to Harm," 326.

> *Negligent Forfeiture*: *P negligently* forfeits a privacy right if each of the following conditions obtains: (1) *P should have been aware* of a somewhat substantial risk* to *P*'s privacy interest(s) that will result from *P*'s act(s) or omission(s); (2) *P* fails to take reasonable precautions to safeguard *P*'s privacy interest(s) in circumstances that satisfy condition 1; and (3) by failing to take reasonable precautions, *P* would unduly impinge on *D*'s interests if *D* were required to act as if *P* had a right to privacy.
>
> *Reckless Forfeiture*: *P recklessly* forfeits a privacy right if each of the following conditions obtains: (1) *P is aware of, but disregards,* a somewhat substantial risk* to *P*'s privacy interest(s) that will result from *P*'s act(s) or omission(s); (2) *P* fails to take reasonable precautions to safeguard *P*'s privacy interest(s) in circumstances that satisfy condition 1; and (3) by failing to take reasonable precautions, *P* would unduly impinge on *D*'s interests if *D* were required to act as if *P* had a right to privacy.
>
> *Risk**: A heightened probability that a putative duty-bearer will become privy to what *P* wishes, or would have wished on reflection at the time, to keep private.

I will clarify these definitions below and outline more fine-grained criteria for applying them in sections 3 and 4.

Clause 1 in each definition echoes the US Model Penal Code's definitions of negligence and recklessness.[18] In keeping with the code's approach, I prefer a relatively clean distinction along the lines set out by Peter Cane: "The difference between recklessness (in its core sense) and negligence resides in the fact that the former has a mental element (deliberation and knowledge of risk) that the latter lacks."[19] Those who classify negligent and reckless mental states somewhat differently can, with suitable modifications, still accept my account of forfeiture. I have, however, replaced the code's language of "substantial risk" with "somewhat substantial risk*." The code sets quite a high bar for liability in part for evidentiary and other pragmatic reasons. As far as the moral realm is concerned, one can run less than a substantial risk of harm and still be negligent or reckless. At the same time, not every uptick in risk seems salient. Thus, I opt for the admittedly imprecise phrase "somewhat substantial risk*."

Note that I did not define risk* in terms of a pure probability that *D* will learn certain information about *P*. Doing so would be overbroad. The information

18 See US Model Penal Code §§ 2.02(2)(c) and (d).

19 Cane, *Responsibility in Law and Morality*, 80. I would, however, set the bar for what counts as "deliberation" quite low, particularly in the moral context.

must be such that "*P* wishes, or would have wished on reflection at the time, [for it] to be private." This proviso reflects the point made in section 1—namely, that mental states must be accounted for to distinguish accurately between cases of forfeiture and mere implied waiver.

Next, I should clarify the phrase "should have been aware of a somewhat substantial risk*" in the definition of negligent forfeiture. That standard can be construed in more or less stringent terms. Imagine that there is some norm *N* prevalent in a given community. Acting in accordance with *N* ordinarily signifies that one aims to relinquish a right to privacy. If *P* acts in accordance with *N* while being *reasonably unaware* of *N*'s existence and implications, will clause 1 be satisfied? I suggest not. If *P* merits no epistemic blame for being unaware of *N*, it would not be fair to hold it against *P* under a negligence standard because *P* has not been careless in any respect. Hence, the "should have been aware" standard would not be met and clause 1 would not be satisfied.[20]

I now want to address perhaps the thorniest dimension of forfeiture—how self-directed aspects and other-directed aspects interrelate. Whereas clauses 1 and 2 in each definition refer to negligent and reckless conduct vis-à-vis one's own interests, clause 3 focuses on "undu[e]" limitations on putative duty-bearers' interests. How do these facets interact? There is a tempting but ultimately spurious way of thinking about forfeiture as a kind of *comeuppance* in which an agent gets what she deserves.[21] But deserves for what? Perhaps for imprudently risking one's own privacy. The trouble with that account is that it makes forfeiture too punitive. Why should agents be stripped of normative protections afforded by a right to privacy for taking *self-directed* risks? Agents are generally free to risk their own interests without moral (as opposed to ethical/axiological) repercussions.

There is, however, a different and more compelling rationale for forfeiture rooted in *interpersonal* considerations. Roughly stated, when *P* is well placed to take reasonably available precautions to secure *P*'s own privacy interests, it is not fair to subject *D* to epistemic risks, moral risks, and liberty-constraining compliance burdens that accompany deontic duties. By "epistemic risk" I mean the risk of making an error about the existence or nonexistence of *P*'s right to privacy in a given case. By "moral risk" I mean the risk of committing a moral wrong and becoming a fitting object of blame, experiencing guilt, and owing remedial duties. And by "compliance burdens" I mean normative limitations on

20 In section 6, I will argue for a less forgiving stance toward duty-bearers, who may sometimes legitimately be held to a standard of strict liability in relation to faultless errors about right-holders' entitlements.

21 RW note, without endorsing, a view along these lines. See "Privacy Rights and Public Information," 15n32.

one's conduct, given that "protecting privacy for one person inevitably leads to restraints on the freedom of another or others."[22]

Suppose that, at t_1, private fact F about person P is not available to D. Then, in virtue of P's negligent or reckless conduct, F becomes readily accessible to D at t_2. At that point, the normative situation changes, introducing epistemic risk, moral risk, and potential compliance costs. D will need to consider whether or not P has a right to privacy in F. Someone who thinks that privacy rights cannot be forfeited will, of course, maintain that a privacy right persists. Even if so, D may make a reasonable mistake. If D then ascertains F, D may become a fitting object for reactive attitudes and owe P a remedial obligation.[23] And if D accurately concludes that P retains a privacy right (again, from the perspective of someone who denies that privacy forfeiture can occur), D may be encumbered with compliance duties to ensure that D refrains from ascertaining F. In my view, it is generally unfair to saddle D with such risks and compliance burdens when P was reasonably well positioned to secure P's own interests, yet failed to do so. The most challenging type of case for my account is one in which D (i) *knows for sure* that P wishes to keep fact F private (in spite of P's negligent or reckless conduct) and (ii) faces *de minimis* or nonexistent compliance burdens. If there is no meaningful interpersonal detriment to D, forfeiture will lose much of its normative appeal, since we would otherwise need to fall back on a "comeuppance" rationale.

Finally, note that even if forfeiture does occur, normative constraints can *still* govern D's conduct. There may be confidentiality-type limits on what D may do with acquired information. There can also be duties not to inflict gratuitous embarrassment, emotional distress, offense, and so on, in disseminating certain information. Not only that, but we may even criticize an agent for *acquiring* information over which P has forfeited a right to privacy. Such criticism will be ethical/axiological rather than moral. The moral realm, as I construe it here, covers what is deontically required, prohibited, and permissible. The ethical sphere is broader, encompassing virtues and excellences, including supererogatory norms.[24] Thus,

22 Nissenbaum, "Protecting Privacy in an Information Age," 571. In her prescient account of "privacy in public," Nissenbaum criticizes various extant theories of privacy for overvaluing duty-bearers' freedoms and undervaluing privacy interests in light of big-data aggregation and inferential analyses (570–75). I agree with those concerns and that "privacy in public" is a genuine phenomenon (e.g., in the form of inferential privacy rights). But Nissenbaum does not deny that duty-bearers' liberties deserve *some* normative weight, so the issue will be a matter of degree.

23 As we will see in section 6, RW adopt a view according to which, if D cannot reasonably conclude that P has a privacy right in a given context, D has no duty to respect it. I will disagree with RW's approach on moral grounds.

24 For this distinction, see Kramer, *Moral Realism as a Moral Doctrine*, 2–3.

even if *D* does not violate *P*'s right to privacy (or, put otherwise, does not wrong *P*) in virtue of forfeiture, *D*'s conduct may still be unseemly, "bad, Not Nice, not done by the best people."[25] Thomson adds: "From the point of view of conduct . . . bad behavior is bad behavior, whether it is a violation of a right or not."[26]

What, then, is the *point* of forfeiture if an agent's conduct remains open to normative (albeit ethical) critique? The answer lies in what is distinctive about violating moral rights as opposed to falling short of ethical ideals: only the former gives rise to remedial requirements. Consider the Remedy Principle set out by Matthew Kramer:

> *Remedy Principle*: If and only if *P* holds vis-à-vis *D* a moral right against *D*'s ϕ-ing, *D*'s ϕ-ing will place *D* under a moral obligation to *P* to remedy the resultant situation in some way.[27]

When a person fails to embody various virtues—e.g., temperance, modesty, bravery—that shortcoming does not itself trigger remedial duties. But failure to comply with deontic duties does. So the difference between a theory of privacy that makes room for forfeiture, and one that does not, has real-world consequences. On a forfeiture-friendly view, when *P* forfeits a privacy right, *D* will not wrong *P*—and thus will not incur a remedial duty—by acquiring relevant information about *P*, even though *D* will remain open to ethical criticism.

3. FIVE APPLICATION CRITERIA

I will now consider how the definitions introduced above apply in some specific circumstances. Thomson observes:

> It is not at all easy to say under what conditions [an agent] has waived [or forfeited] a right—by what acts of commission or omission and in what circumstances. The conditions vary, according as the right is more or less important; and while custom and convention, on the one hand, and the cost of securing the right, on the other hand, play very important roles, it is not clear precisely what roles.[28]

25 Thomson, "The Right to Privacy," 296. RW repeatedly invoke the infringing/violating distinction (which Thomson herself popularized). Because nothing in the paper turns on that distinction, so far as I can see, I will continue to speak of "violating" rights.

26 Thomson, "The Right to Privacy," 298.

27 Kramer, "Moral Rights and the Limits of the Ought-Implies-Can Principle," 313 (minor modifications added).

28 Thomson, "The Right to Privacy," 302.

To build on this terse but incisive sketch, I will suggest five factors to help determine an agent's degree of negligence or recklessness and the reasonableness of available safeguards. I will address factors 1–4 here and leave the fifth factor for the next section.

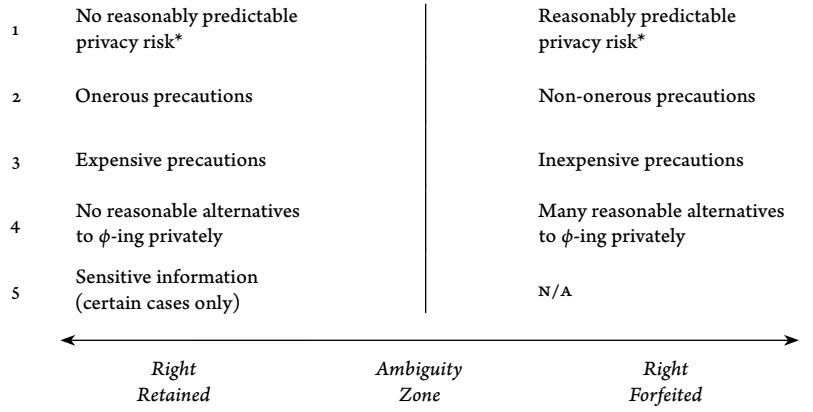

FIGURE 2 Factors Relevant to Forfeiture Analysis

Judgments will no doubt vary about how to interpret and apply these factors, how much weight they deserve in particular cases, and how they interconnect. Below, I will illustrate how the schema works with simple examples and flag issues for further discussion.

As to the first factor, some nuances are worth noting. To begin with, a privacy risk* can be spelled out at various levels of abstraction. For example, is it a risk* that one's personal data may be misused by (i) some entity in some way or (ii) misused by entity *E* in context *C*? The appropriate level of generality will depend on the circumstances. Next, recognizing *that* there is a privacy risk* to begin with can involve epistemic costs and effort that need to be accounted for. By the same token, if an agent is negligently unaware of salient privacy risks* in a given situation, this may count against that agent in a forfeiture analysis.[29] The various considerations adduced in this paragraph can be folded into the reasonableness qualifier within the first factor.

Consider, now, two of Thomson's scenarios in slightly modified form.

Open Windows: A couple has a loud fight in a low-floor apartment overlooking a street with pedestrian traffic. The couple has not "thought to

29 If an agent is negligently unaware of a certain privacy risk*, but it turns out that no reasonable precautions could be taken in any event in that situation, the agent's oversight would not increase the likelihood of forfeiture.

> close the windows," so it can easily be heard. A passerby hears the argument and stops to listen.[30]
>
> *Closed Windows*: A couple is having a quiet fight in their apartment with windows closed. Unbeknownst to them, a neighbor across the street trains an amplifying device onto their windows that captures sound waves inaudible to the human ear that migrate beyond the closed window. The neighbor gleans the conversation.[31]

I agree with Thomson that the two cases showcase "not merely a difference in degree, but a difference in quality."[32] A privacy right is divested in the first case, but not the second. Thomson elaborates as follows (notably failing to distinguish forfeiture and implied waiver):

> If my husband and I are having a loud fight, behind open windows, so that we can easily be heard by the normal person who passes by, then if a passerby stops to listen, he violates no right of ours, and so in particular does not violate our right to privacy. Why doesn't he? I think it is because, though he listens to us, we have *let* him listen (whether intentionally or not), we have waived [or forfeited] our right to not be listened to—for we took none of the conventional and easily available steps (such as closing the windows and lowering our voices) to prevent listening.[33]

I will assume that the couple *did not* want to relinquish its privacy right, making the scenario a candidate for forfeiture. The privacy risks* here are obvious, per the first factor. Next, Thomson's phrase "conventional and easily available steps" gestures at the second and third factors. The epistemic costs of identifying precautions are in effect nil, since the steps are apparent as well as simple and costless. Finally, the couple has a reasonable alternative to fighting with its windows open: doing so with windows closed. The couple, in other words, negligently risked its privacy interests despite reasonably available precautions and, hence, forfeited its right to privacy. In Closed Windows, all the factors point the other way. The neighbor's use of an amplifying device is hardly predictable. Even if it were, there are no widely known, feasible steps to protect against it. If so, the couple's only alternative would be to argue some place other than its home, depriving it of reasonable alternatives. The couple, in other words, has

30 Thomson, "The Right to Privacy," 296; see also 306.

31 Thomson, "The Right to Privacy," 296.

32 Thomson, "The Right to Privacy," 296.

33 Thomson, "The Right to Privacy," 306.

not been negligent or reckless vis-à-vis its privacy interests and retains its claim-right to privacy.

As in Closed Windows, in another one of Thomson's examples—a park-bench scenario—there is no negligent or reckless conduct, and hence no forfeiture. Where a pair hoping to "talk over some personal matters" chooses a "bench far from the path," it retains its right to privacy against an eavesdropper who "creeps around in the bushes … and crouches [at the] back of the bench to listen."[34] To be clear, I think the explanation here does not turn on the eavesdropper's bad motives. Imagine that Sari enjoys sitting nestled in those bushes each afternoon and reading for pleasure. If, one day, Sari happens innocently to overhear the pair's exchange, the pair would *still* retain its right to privacy despite Sari's impeccable motives. The real explanation here, I believe, has to do with the *unpredictability* of risk* where agents have taken due precautions.

But even incurring predictable privacy risks* does not *entail* forfeiture. That is true, for example, when third parties constrain one's choice situation in morally untenable ways. Suppose that *A* and *B*'s house is far from their property line. *C* trespasses and positions himself by an open window. Assume, as well, that *A* and *B know* of *C*'s presence. Here, unlike in Open Windows, there would be no forfeiture. The difference lies is *C*'s independent moral wrong of trespass that illegitimately constrains *A* and *B*'s choice situation. Though it would be highly imprudent for *A* and *B* to keep the windows open if they know of *C*'s presence, forfeiture does not occur because the phrase "unduly impinged" in clause 3 of the reckless forfeiture definition is not met in virtue of *C*'s trespass.

The same point—that incurring predictable privacy risks* does not ineluctably result in forfeiture—applies when activities are integral to carrying out one's life plans and there are no reasonable alternatives to pursue them in privacy-preserving ways. Suppose that you live in an area with a sole utility provider. The provider informs you that it sells granular electricity usage data to marketers that could reveal various personal or intimate details, and there is no opt-out mechanism. Despite predictable risks*, you would not forfeit a moral right to privacy in that data by signing up, since electricity provision is an essential service and you are faced with a monopolistic provider.[35] In contrast, if you had a choice between two otherwise identical providers, only one of which sells granular data, you may forfeit (or waive) your moral right to privacy in relevant data by knowingly selecting the data-monetizing utility. (If the overall deal offered by

34 Thomson, "The Right to Privacy," 298.

35 A similar analysis would apply to Andrei Marmor's hypothetical in which the government openly announces its plans to record all telephone calls. See Marmor, "What Is the Right to Privacy?" 14–15.

the non-monetizing utility is materially worse than the monetizing utility, perhaps there still would not be a reasonable alternative.)

Finally, I would like to say a word about factors relevant to the *duty-bearer*'s situation, focusing on considerations introduced in section 2, albeit in reverse order: (1) compliance burdens and (2) epistemic and moral risks. As to 1, *D* may face new compliance burdens in virtue of *P*'s negligent or reckless conduct. For example, in Open Windows, passersby may have to modify their route or stopper their ears to avoid overhearing the couple. In RW's related hypothetical, a neighbor would need to "adopt a kind of wilful deafness"[36] to avoid overhearing neighbors having an altercation.[37] Or, as in some of Thomson's examples, *D* "would have to go to some trouble" to avoid acquiring relevant information or, more strongly, "cannot help but" acquire it due to *P*'s acts or omissions.[38]

In evaluating these compliance burdens, what might we say about assessing the importance of *D*'s liberty interests? Trying to assign them comparative weight is not at all straightforward. How would we decide whether *D*'s liberty interest in looking at *X* or listening to *Y* or walking near *Z* is more or less weighty in a particular case or in general? Matters are further complicated by asking such questions both about the interests of natural persons and corporations (commercial giants, startups, nonprofits, etc.) and trying to gauge whether the liberty interests of persons or corporate entities deserve more or less normative weight either as a general matter or in specific cases of putative privacy forfeiture. These issues deserve further reflection beyond this paper's scope. In any event, the bar for normative significance of liberty interests in relation to compliance burdens should be set relatively low, in my view, to satisfy clause 3 in both definitions of forfeiture. That is because we are already focused on the subset of cases in which clauses 1 and 2 have been met. That is—but for *P*'s reckless or negligent conduct in the face of somewhat substantial risks* and the presence of reasonable precautions—*D* would not be facing any added compliance burdens at all.

Now let us assume that compliance costs are *de minimis* or nonexistent. In that case, the existence of epistemic and moral risks can satisfy clause 3. In real-world scenarios such risks will typically exist. That is, *D* will not know *P*'s actual intentions and may thus make a (reasonable) mistake and violate a putative moral duty. To underscore this point, consider two contrived scenarios in which those risks are deliberately taken off the table. Suppose that *P* negligently misdirects a personal email to *D*'s inbox. Before *D* sees it, *P* calls *D* to say that the email was sent in error and asks *D* to refrain from opening it. What result? Given that

36 Rumbold and Wilson, "Privacy Rights and Public Information," 11.

37 See section 6 for further discussion of this case.

38 Thomson, "The Right to Privacy," 301, 303, 304.

D is apprised of *P*'s intent and compliance is not hard, no forfeiture would likely result.[39] Or consider a case based on RW's hypothetical to which I will return in section 5.[40] Imagine that *P* voluntarily uploads her name and DNA onto a public website but appends the following note: "I hereby assert a right to privacy in relation to any genetic disorder(s) that may be inferable from my DNA." Ignoring the folly of *P*'s conduct, with that note in place, putative duty-bearers would be fully apprised of *P*'s intent. Assuming that the compliance burden is *de minimis*, forfeiture again may be unjustified. After all, *D*s are left no worse off than if *P* had not uploaded the DNA in the first place. Unlike in these contrived cases, however, in the real world, *D*s typically will *not* be privy to *P*'s actual intentions. The epistemic and moral risks that will exist in those contexts can be sufficient to satisfy clause 3 in each definition of privacy forfeiture, even if compliance burdens with a putative right to privacy are not especially onerous or even nonexistent.

4. SENSITIVITY OF PRIVATE INFORMATION

I will now address the fifth and final factor introduced in the previous section. How should varying sensitivity levels of private information affect forfeiture?

To set up a puzzle, imagine two variants on Open Windows. In one, a couple argues about its dinner plans. In the other, it fights about whether the infidelity of one spouse should lead to divorce, airing salacious details. Factors 1 through 4—predictability, onerousness, costs, and reasonable alternatives—are identical in both cases. If that were the full story, both couples would be equally likely to forfeit their rights to privacy. But the second couple seems far more negligent, given its sensitive topic. Yet that fact may seem to count *against* forfeiture, precisely because passersby would then be morally free to listen in (while remaining open to ethical criticism). Conversely, the couple bickering about dinner plans appears less negligent, given its mundane topic. This fact may seem to make forfeiture *less* objectionable as compared to the previous case.

Which way does sensitivity cut, then? Does forfeiture become more likely as sensitivity rises, since it bespeaks greater negligence or recklessness? Or does forfeiture become less likely, since it would erode normative protections for sensitive facts?[41]

39 Should we take into account the fact that *D* may be *tempted* to open the email (or, more generally, access now-available information about *P*), and must exercise self-restraint to avoid doing so, as a factor that cuts in favor of forfeiture? I think not, since that would not be fair to *P*. But see note 42.

40 See Rumbold and Wilson, "Privacy Rights and Public Information," 14–15.

41 I ignore trickier cases relevant to inferential privacy where the underlying facts may be mun-

To begin with, various complications will arise in determining the true sensitivity level at issue. Does it depend solely on an agent's subjective judgments and attitudes? If not, what sort of reasonableness constraints are warranted? I set aside these issues and simply assume that we are dealing with sensitive facts. If so, one option is to contend that higher sensitivity always increases chances of forfeiture, at least to some extent. But that view strikes me as too punitive, much like the comeuppance rationale discussed in section 2. Granted, failing to take precautions with respect to very sensitive facts is more negligent or reckless than failing to do so vis-à-vis mundane facts. But that tells us nothing about the *duty-bearer*'s situation. For example, no matter what the couple fights about—dinner or marital trouble—passersby would need to take the identical steps to respect a right to privacy. If higher sensitivity could *in itself* be decisive for a forfeiture verdict, holding other factors fixed, the account would be overly harsh toward right-bearers.[42]

With these remarks in mind, my proposal to handle different sensitivity levels is reflected in figure 3 below. (In order not to beg any questions, sensitivity is excluded altogether from judgments about degrees of negligence and recklessness in the two columns.)

	Low/Moderate Negligence	*Gross Negligence and All Recklessness*
Low/Moderate Sensitivity	1. Neutral	2. Neutral
High Sensitivity	3. Tells somewhat against forfeiture	4. Neutral

FIGURE 3 Sensitivity Levels of Private Information and Forfeiture

Let us first consider box 3—cases where sensitivity is high but agents are no more than moderately negligent. Here, a somewhat forgiving attitude toward right-holders is justified because the right to privacy protects important interests whose normative weight increases with heightened sensitivity and forfei-

dane, but when coupled with other available data and processed by powerful analytics tools can yield sensitive inferences.

42 There is one set of circumstances in which quasi-punitive considerations could play a role when tethered to interpersonal harms: *serial* negligence or recklessness. Suppose that an agent negligently, and repeatedly, sends misdirected email messages. Even if she retains a privacy right at the *first* such transmittal, she could arguably forfeit her privacy right in an identical misdirected email by the *tenth* such erroneous transmittal.

ture is clearly detrimental to those interests. These considerations tell somewhat against forfeiture in box 3–type cases. But extending the same approach to cases of gross negligence or recklessness (boxes 2 and 4) is not justified given the higher level of risk that agents run, especially when they do so knowingly in cases of recklessness. But to avoid an excessively punitive verdict even in those cases, sensitivity can be treated as a neural factor that tilts the scales neither for nor against forfeiture. The same approach can be extended to cases where sensitivity is low to moderate (boxes 1 and 2), since it would make little sense to treat those circumstances either more or less favorably than cases with high sensitivity (boxes 3 and 4). With this rubric in mind, we can briefly reconsider the puzzle with which I started. The first variant of Open Windows (dinner dispute) can be slotted in box 2 and the second variant (infidelity dispute) in box 4. I have located both cases in the second column because, as in Thomson's original scenario, the privacy risks* are obvious and the safeguards are equally clear and costless. Sensitivity thus ends up playing a neutral role in both scenarios.

Assigning a role to the sensitivity of private information in forfeiture analysis raises a host of complications, some of which I have noted above. Most important, to steer clear of overly punitive results, high sensitivity need not work against right-bearers (boxes 1, 2, and 4) and, in one type of case, can favor them by counting somewhat against forfeiture (box 3).

5. FORFEITING PRIVACY RIGHTS VERSUS PROPERTY RIGHTS

Having outlined my account of privacy forfeiture, in the remainder of the paper I will engage with RW's criticisms of Thomson and their own take on privacy.

One way to test a philosophical thesis is to consider how it fits with one's other commitments. If the fit is incongruous, that is a defeasible strike against the thesis. I take up such an inquiry here by asking how the idea of privacy-rights forfeiture relates to forfeiture of a different kind of entitlement—property claim-rights in tangible goods. This comparison, which RW briefly invoke to impugn privacy rights forfeiture, raises broader issues of interest to philosophers and social scientists. The topic is especially apt because Thomson herself compares privacy rights and property rights, though she does not broach the issues that I will address here.[43]

RW imagine a famous actress, Annabel, who volunteers to support an initiative to promote genetic research. She agrees to donate her DNA to research and put it online with her name. Annabel has a rare, hard-to-diagnose genetic disor-

43 See Thomson, "The Right to Privacy," 304–6, 312–14.

der that she wants to keep private. After the DNA is put online, a geneticist downloads it, analyzes it, and finds the disorder. Would he violate Annabel's right to privacy by publicizing it?[44] (Note that this case involves inferential privacy, since the disorder is inferred from the DNA.)

RW think so, even though "Annabel intended to publicize the contents of her DNA."[45] They then raise the possibility that Annabel forfeited her right to privacy given the obvious risks of making her DNA public: "Although we might chastise Annabel for her naivety in this situation, it is far from clear that, simply by virtue of that naivety, we should also think she has somehow forfeited her right to privacy."[46] RW continue:

> After all, one does not forfeit one's right to private property simply by absent-mindedly leaving one's car keys in one's car. Failing to act in a way that ensures, as far as possible, that the car will not be stolen does not somehow mean that the car is no longer *ours*, or that we cannot make reasonable demands on others by virtue of our rights over it.[47]

RW's first sentence is undoubtedly right. But it is not enough to clinch RW's point as to Annabel's retention of her right to privacy. For the passage appears to depend on an unstated assumption: forfeiture thresholds for tangible property rights are not very different from those for privacy rights. But, as I will now suggest, there may be good reasons to doubt that assumption.

If Annabel forfeits a right to privacy in her genetic disorder—as my account suggests she does, assuming duty-bearers face relevant epistemic risks, moral risks, or compliance burdens—why does a person who absentmindedly leaves a wallet or laptop at a coffee shop *not* forfeit a right to those things, even if the level of negligence or recklessness exhibited is the same or greater than in Annabel's case?[48] Put another way, why do property rights seem stickier—harder to forfeit—than privacy rights? In taking up this question, I will contrast "pure" privacy cases with "pure" property cases. I bracket cases implicating both types of rights—for example, appropriation of a hard drive containing sensitive personal data—where I believe that the stricter property-rights standard should control.

As an initial matter, it *is* possible to forfeit property rights (paradigmatically, land rights) under the legal doctrine of "adverse possession." But that doctrine sets such a high bar—often requiring ten years of actual, exclusive, hostile, and

44 Rumbold and Wilson, "Privacy Rights and Public Information," 14.

45 Rumbold and Wilson, "Privacy Rights and Public Information," 14.

46 Rumbold and Wilson, "Privacy Rights and Public Information," 15n32.

47 Rumbold and Wilson, "Privacy Rights and Public Information," 15n32.

48 I grant that there may be ethical/axiological objections to publicizing the disorder.

"open and notorious" occupancy—that it is largely an exception that proves the rule that property rights are hard to forfeit.[49] Next, distinct forfeiture thresholds cannot simply be explained by distinct degrees of harm resulting from forfeiture of privacy rights and property rights, respectively. After all, losing ownership rights to a wallet or laptop can be much *less* harmful than forfeiting a privacy right to sensitive medical information, for example. Hence, I will now consider two other potential explanations, though I concede that these issues are not clear-cut and may resist any neat generalizations.

First, I suspect that—at least in some cases—distinct forfeiture thresholds may have to do with varying levels of intrusiveness of the correlative duties. While obligations to respect tangible property rights center on outward conduct, duties of privacy can involve intimate psychological processes, requiring agents not to look, scrutinize, read, listen, and so forth. RW, for their part, even contemplate a duty to refrain from certain private *thought processes* that may yield sensitive inferences about other people.[50] While I believe that such a duty goes too far, it underscores the point that duties of privacy can be psychologically intrusive in ways that property-related duties are not and, for that reason, may justify a somewhat lower forfeiture bar. Granted, complying with some privacy obligations may not be very invasive psychologically; meanwhile, duties to respect property rights can be onerous in their own ways and impinge considerably on our liberties.[51] Thus, the contrast drawn in this paragraph is hardly decisive.

Second, forfeiture of property rights and privacy rights has different upshots for social order and coordination, which typically require clear and predictable rules about who owns what and how scarce resources are appropriated, apportioned, and transferred. Tangible goods such as land, houses, cars, wallets, and computers are rival and excludable. If the forfeiture threshold for such goods were set too low, property rights may become less stable and predictable because many more disputes over property will arise; perverse incentives may materialize to cause others to forfeit their property rights; and more resources may be needed to secure property interests, likely to the disadvantage of those who are worst off in society. Granted, these are empirical conjectures.[52] Still, it seems plausible that a relatively high threshold for property-rights forfeiture can help avert those destabilizing social outcomes.

49 See, e.g., Nome 2000 v. Fagerstrom, 799 P.2d 304 (Alaska 1990).

50 See Rumbold and Wilson, "Privacy Rights and Public Information," 15.

51 I am grateful to an anonymous referee for pressing these points.

52 There does not appear to be a meaningful empirical literature on comparative property forfeiture rules.

A comparatively lower threshold for privacy-rights forfeiture does not pose analogous risks for social stability. For one thing, privacy rights usually do not involve rivalrous goods that can be fought over like scarce resources. If D_1 learns *P*'s secret or sees *P*'s picture, that does not leave any less of anything for D_2, D_3, etc. While privacy rights can surely involve market commodities—for example, information bought and sold by data brokers—they are also often untethered from market prices, so that agents do not suffer direct economic losses from forfeiture. Crudely stated, if I forfeit title to my laptop, I am out $1,500; if someone overhears me in Open Windows, it does not leave me any poorer to pay the bills (even if the harm that I suffer in losing title to my laptop is *less serious* than the harm that would be caused by a major privacy violation). In short, the threat to core prerequisites for civic order that arises if agents can readily forfeit tangible property rights does not apply with the same force to privacy rights.

Relatedly, low forfeiture thresholds for property rights may intensify worries about intrusive involvement of legal-governmental officials in people's lives. Low thresholds of that sort would almost certainly multiply disputes about whether forfeiture has occurred and who is the new rightful owner of forfeited goods. Those conflicts, in turn, may intensify involvement of law enforcement and the legal system in people's lives (at least in regimes with robust rule of law where property rights are well enforced). Privacy, again, is different. The state should not be in the business of resolving many privacy disputes in the first place, as in Open Windows or who overheard what at the office watercooler. And while common-law privacy torts vindicate important interests, the law is also ill equipped to protect privacy in various respects.[53] So a comparatively lower threshold for privacy-rights forfeiture likely will not supercharge state intervention in people's affairs as much as low forfeiture thresholds for property rights would.

Clearly, more must be said on this score. But explanations along the lines canvassed above can, I suspect, make sense of disparate forfeiture thresholds for privacy rights and tangible property rights. While RW helpfully allude to this contrast, its sheer existence, without further argumentation, does not cast doubt on privacy-rights forfeiture.

6. RW Overprotect and Underprotect Privacy Rights

In this final substantive section, I turn to RW's positive model of privacy. Their account has many dimensions, so I will focus on just two. First, RW make no

53 See Gavison, "Privacy and the Limits of Law," 370–72.

explicit provision for forfeiture in their model, resulting in a theory that appears too solicitous toward right-holders. Second, to balance out that normative picture, RW enter two key caveats. Those caveats, I will argue, veer too far in the other direction, *under*protecting privacy interests of faultless right-holders and biasing outcomes against minority preferences in morally problematic ways.

I first want to rule out a purely terminological dispute. Objecting to Thomson's idea of "unintentional waiver," RW say that "if one is to waive a right, one would seem to need actually to waive it." In one sense, I agree.[54] Thomson invites needless confusion by speaking of "unintentional wavier" instead of "forfeiture" even when it is clear that agents *do not* wish to relinquish an entitlement.[55] If that were the whole dispute, RW could accept the bottom line in Thomson's examples but redescribe relevant cases in terms of forfeiture, as I have done. But RW appear to press a deeper objection. They seem to contest the very idea that agents can be divested of privacy rights involuntarily. They say that agents can expressly waive a privacy right and that rights can become "defunct" (a topic I address below).[56] But what agents "may not do is *unintentionally waive* their rights, which is to say, accidentally absolve duty-bearers of their rights [*sic*]."[57] To the extent that this remark is meant to go beyond terminological quibbles, I interpret it—together with the absence of any acknowledgement of forfeiture in their article—as ruling out privacy forfeiture.

On that view, the right to privacy becomes an outlier, a fundamentally different sort of right from other moral and legal entitlements. If one can forfeit a right to a friend's trust by deceiving her, or to raise a child by mistreating him, or to rely on certain legal claims or defenses by failing to assert them in a timely way, why can one not forfeit a right to privacy by failing to take readily available safeguards when privacy risks* are predictable? On RW's picture, the lodestar is a right-holder's *intention* as to which facts to keep private (bracketing RW's caveats that I discuss below).[58] It appears that, no matter how negligent or reckless an agent may be, if she *unintentionally* discloses private information that she does

54 If by "actually . . . waive," RW mean "expressly waive," they may be overstating matters, given my account of implied waiver in section 1.

55 See, e.g., Thomson, "The Right to Privacy," 301–2 (an agent who "positively want[s] that nobody shall look at the[ir] picture" "unintentionally waives" a right to privacy by leaving the picture in a public place).

56 Rumbold and Wilson, "Privacy Rights and Public Information," 12–13.

57 Rumbold and Wilson, "Privacy Rights and Public Information," 13. RW presumably intended to write "obligations" or "duties" rather than "rights."

58 Rumbold and Wilson, "Privacy Rights and Public Information," 13–14.

not wish to make public, the costs of securing her privacy could, in principle, be imposed on others.

As one illustration, consider RW's variant on Thomson's Open Windows. RW imagine neighbors having a "highly personal, but also very loud, argument that you cannot help but overhear," while knowing that they "have absolutely no intention of broadcasting their discussion."[59] According to RW, one has a moral duty "to adopt a kind of wilful deafness" in order not to "pay too close attention to precisely *what* they are saying."[60] But that strikes me as a lopsided, unfair result. The idea that others must stopper their ears to accommodate *me* if *I* am speaking loudly when I have neighbors—and perhaps apologize to me if they hear me too distinctly—is too solicitous toward me, since I have been negligent or reckless by failing to take basic safeguards. While adopting "willful deafness" may be *virtuous* in that context, doing so is fully consistent with endorsing a forfeiture verdict as to the privacy right(s) at issue.

The same misallocation of moral benefits and burdens will characterize countless other cases on a theory of privacy that makes no room for forfeiture. Whether it is the negligent couple in Open Windows or a person who serially misdirects sensitive emails, or Annabel intentionally posting her DNA online (without wishing to make public her genetic disorder), the onus to protect privacy—and the attendant moral risks, including the possibility of being subject to reactive attitudes and remedial obligations—can, in principle, fall onto others on RW's model.

Recognizing that their account may seem too onerous for duty-bearers, RW adjust it in two main ways. One centers on physical/psychological constraints faced by putative duty-bearers and the other focuses on epistemic constraints. These caveats, I will now suggest, are morally problematic in their own right.

First, RW say that privacy rights become "defunct" if a duty-bearer cannot, in some physical and/or psychological sense, help but learn a private fact (or a slightly weaker condition is met).[61] Say *P* leaves a picture in a public place or has a loud fight with the windows open. *D* comes along and simply *sees* the picture or *hears* the fight. In such cases, privacy rights become "defunct" and duty-bearers do nothing wrong in acquiring private information.

How does the notion of "defunct" rights intersect with my account of privacy forfeiture? Figure 4 below captures areas of functional convergence and divergence.[62]

59 Rumbold and Wilson, "Privacy Rights and Public Information," 11.

60 Rumbold and Wilson, "Privacy Rights and Public Information," 11.

61 Rumbold and Wilson, "Privacy Rights and Public Information," 10–11, 20–21.

62 I am assuming in this chart that RW's second caveat (discussed below) is not relevant.

	Right "Defunct"	*Right Not "Defunct"*
Right Forfeited	1. Functional convergence	2. RW overprotect privacy
Right Not Forfeited	3. RW underprotect privacy	4. Functional convergence

FIGURE 4 "Defunct" Rights and Forfeiture

Our verdicts coincide in two scenarios: when a right is "defunct" *and* forfeited (box 1) and when a right is *neither* "defunct" *nor* forfeited (box 4). I will focus instead on the points of disagreement—boxes 2 and 3. In box 2, where a right that is not "defunct" is arguably forfeited on my account—e.g., Open Windows, RW's loud neighbors case, Annabel's case—RW are, again, being overly solicitous toward negligent or reckless right-holders. Since I addressed this problem earlier, I will focus here on box 3, where rights are "defunct" but *not* forfeited. Here, RW err in the other direction. They give unduly short shrift to full-fledged privacy interests by stripping them of normative force through no fault of a right-holder's own. In such cases, the better view is to say that, while duty-bearers may be acting non-culpably, they could *still* violate a right to privacy and owe innocent right-holders a remedial duty, even if a modest one.[63] I will return to this idea below, since it tallies with my response to RW's second strategy for cabining the burdens that fall on duty-bearers.

RW's second approach centers on epistemic constraints that arise in deliberating about one's obligations: "Whether or not *Q*'s actions infringe or even violate *P*'s privacy rights depends in part on what *Q* could *reasonably have expected P*'s concerns were with regard to once-private information that now finds itself in the public domain."[64] So RW accept:

> *Conscientious Non-Violation*: If "a conscientious agent, taking the utmost care to respect an individual's right to privacy, will still get it wrong about what that individual intended or did not intend when they made a certain

63 See Kramer, "Moral Rights and the Limits of the Ought-Implies-Can Principle," 317–22, 328–31. RW may hold other philosophical commitments that would lead them to contest the possibility of non-culpable rights violations, a point of disagreement that would go beyond issues of privacy.

64 Rumbold and Wilson, "Privacy Rights and Public Information," 19, emphasis added.

> piece of information public," that is "a cause for regret in the harm they do to *P*," without constituting a violation of *P*'s rights.[65]

Note that, earlier on, RW say that the "right to privacy … track[s] the extent to which an individual *intended* to make a piece of information public."[66] Conscientious Non-Violation waters down this key commitment by making the normative bite of privacy rights turn in part on duty-bearers' myriad epistemic circumstances. Even if *P has not* waived a right, that right becomes inert whenever duty-bearers deliberate blamelessly but mistakenly about *P*'s interests. *P* is then out of luck, merely a "cause for regret."[67] So Conscientious Non-Violation succeeds at cabining the burdens on duty-bearers. But it does so at the unfair expense of innocent right-bearers.

Conscientious Non-Violation has a more serious moral strike against it—one that RW acknowledge and strive to deflect, albeit unsuccessfully in my view. The disadvantages that Conscientious Non-Violation places on right-holders are not distributed equally or randomly. Rather, they disproportionately disadvantage preferences, dispositions, and norms that are in the minority in a given group or community. When a putative duty-bearer does not know *P*'s actual intentions, she will need to rely on certain heuristics, simplifying assumptions, generalizations, and so forth. Those, in turn, will usually track majoritarian views to maximize the chances of getting it right, which is an epistemically sensible strategy. So Conscientious Non-Violation will systematically discriminate against the privacy interests of minorities and idiosyncratic parties. That itself is a good reason to reject the thesis, RW's efforts to reply notwithstanding.

Take RW's own example of an Orthodox Jewish woman's right to privacy in a photograph depicting her hair, which religious custom forbids her from doing in public if she is married. RW imagine that the photo falls into the hands of "a British citizen living in a close-knit rural village in England."[68] The villager decides whether to disseminate it based on "social norms prevalent in her society."[69] Since the villager "could not be reasonably expected to think that [the woman] would have wanted such a photograph to be kept private," the villager would not

65 Rumbold and Wilson, "Privacy Rights and Public Information," 19–20.

66 Rumbold and Wilson, "Privacy Rights and Public Information," 14.

67 Rumbold and Wilson, "Privacy Rights and Public Information," 20.

68 Rumbold and Wilson, "Privacy Rights and Public Information," 22.

69 Rumbold and Wilson, "Privacy Rights and Public Information," 22.

violate the woman's privacy right.[70] Hence, on RW's view, the woman has "no justifiable complaint" against the villager.[71]

We can substitute the Orthodox woman's preferences for the practices of any minority group or community or any agent's eccentric predilections. RW's verdict in such cases will render non-waived, non-forfeited privacy rights a dead letter, unlike in the case of agents who happen to harbor mainstream preferences that duty-bearers are better positioned, epistemically speaking, to discern. No matter how faultless the villager's deliberation may be, in my view she could *still* violate the Orthodox woman's right to privacy (again, assuming it has not been forfeited), and owe her a remedial duty.

RW concede that their account "appears to discriminate against minorities."[72] But their strategy to neutralize the criticism is not convincing. They clarify that "what constitutes a reasonable expectation . . . cannot be defined simply by the most prevalent norms within a given society," and they advise duty-bearers to "consider the likelihood of P being a member of [a] minority group and the norms held by such groups."[73] That is fair enough. But it hardly guarantees that *D* will reach the right conclusion. *D* may lack certain information or may reasonably misinterpret available facts. The normative force of privacy rights, particularly those of agents whose preferences are atypical in a given group or community, will remain hostage to duty-bearers' various epistemic blind spots and limitations.

My alternative view allows us to avoid the lopsided upshots of Conscientious Non-Violation. Granted, one implication of my account is that duty-bearers can incur remedial obligations for violating rights in non-culpable ways. To preempt potential objections to this upshot, I will make four points. These points extend, *mutatis mutandis*, to the way that I propose to handle cases of non-forfeited but "defunct" rights in RW's sense.

First, there will be fewer privacy rights on my model to begin with, since rights can be both waived and forfeited. In cases of genuine forfeiture, deliberative errors on the part of putative duty-bearers will be immaterial, since the entitlement in question will not exist.

Second, remedial duties resulting from erroneous but faultless deliberation will be attenuated, often significantly, compared to the remedial duties resulting from culpable violations.[74] In the villager's case, it may amount to an apology, an

70 Rumbold and Wilson, "Privacy Rights and Public Information," 22.

71 Rumbold and Wilson, "Privacy Rights and Public Information," 22.

72 Rumbold and Wilson, "Privacy Rights and Public Information," 22.

73 Rumbold and Wilson, "Privacy Rights and Public Information," 22.

74 For a related point, see Kramer, "Moral Rights and the Limits of the Ought-Implies-Can Principle," 328–31.

explanation, and perhaps a sincere expression of intent to make amends in some way. Nothing more would be needed, especially because the reputational and emotional harms that could result from disseminating the picture in the woman's community would not arise in the villager's social context.

Third, a choice between two alternatives seems inevitable. Either a duty-bearer's faultless but mistaken deliberations will leave innocent right-holders without normative recourse (RW's position) or epistemically faultless but mistaken duty-bearers can sometimes incur remedial duties, even if minimal ones (my position). The latter approach is preferable on moral grounds, including non-biased treatment of minorities and idiosyncratic agents. There is a further point, as well. While Hohfeldian rights and duties are correlative, a *justificatory* asymmetry is nevertheless at work. Privacy rights and duties exist to protect *right-holders'* interests. My position respects this asymmetry by extending slightly greater protections to innocent right-holders and placing slightly greater burdens on non-culpable violators.

Fourth, it may be objected that I am poorly placed to criticize RW for overemphasizing majoritarian preferences because my definition of negligent forfeiture may tacitly privilege such preferences. But that is not so, given my remarks in section 2 about the phrase "should have been aware of a somewhat substantial risk.*" I said that if *P* acts in line with social norm *N* that ordinarily signals an intent to relinquish a privacy right, yet *P* is *reasonably unaware* of *N* (perhaps because *P* is part of a minority group or community), the "should have been aware" standard will not be satisfied. So I am treating epistemically innocent right-holders in a more forgiving way than epistemically innocent duty-bearers. This asymmetric treatment is justified. As I noted in the third response, above, *some* tradeoff must be made between the interests of right-holders and duty-bearers. And given the important values secured by the right to privacy, privileging non-culpable right-holders who would otherwise forfeit a right to privacy under a negligence standard is the better tack, particularly because remedial obligations for non-culpable duty-bearers will typically be substantially attenuated.

In sum, RW's account overprotects and underprotects privacy. By seemingly leaving out forfeiture, RW extend normative protections to reckless or negligent right-holders at the unfair expense of duty-bearers. Then, overcorrecting the other way, RW let duty-bearers off the hook too readily in cases of "defunct" rights and cases of faultless but mistaken deliberation by duty-bearers in which *right-holders* have done nothing culpable. On my view, agents can waive a right to privacy or forfeit it in the ways that I have specified. Short of this, privacy rights will usually retain their normative bite. Nonculpable right-holders should not be at the mercy of duty-bearers' epistemic blind spots, and the rights of minori-

ties and idiosyncratic individuals should not be systematically shortchanged. One implication of my account is that duty-bearers can incur remedial duties for non-culpable rights violations. This upshot, for the reasons given above, is preferable on moral grounds to RW's approach.

7. CONCLUSION

Though much work remains to make sense of how privacy rights (and other rights) can be forfeited, I have made three contributions in this paper toward these efforts. First, I proposed a novel taxonomy of how Hohfeldian entitlements can be divested generally and distinguished among varying species of waiver and forfeiture. Second, I developed a theory of privacy rights forfeiture according to which privacy rights can be forfeited in one of two main ways rooted in negligent and reckless conduct. In making that case, I reconstructed and amplified key ideas in Thomson's canonical work on privacy, clarified how distinct sensitivity levels of private information affect forfeiture, and suggested that there may be legitimate reasons why forfeiture thresholds for property rights in tangible goods are more stringent than forfeiture thresholds for privacy rights. Finally, I advanced a moral critique of RW's forfeiture-free model of privacy by arguing that it at once underprotects and overprotects privacy rights.

My account also offers new normative resources to make progress on the subject of inferential privacy. The definitions of privacy forfeiture that I proposed, along with the five application criteria, can help draw principled distinctions between cases where inferential privacy rights may be asserted and where they have been forfeited. The key point to flag here is that one need not choose between a Thomson-inspired theory of privacy-rights forfeiture and inferential privacy rights. One can, and should, endorse both.[75]

Wilmer Cutler Pickering Hale and Dorr LLP
mark.hanin@gmail.com

75 For very helpful comments and/or discussion I would like to thank Marcello Antosh, Ben Bronner, Laura K. Donohue, William English, Boris Hanin, Matthew H. Kramer, Maggie Little, and anonymous journal referees. For funding support in 2020–21, I am grateful to the Fritz Family Postdoctoral Fellowship at Georgetown University.

REFERENCES

Allen, Anita L. "Protecting One's Own Privacy in a Big Data Economy." *Harvard Law Review Forum* 130 (December 2016): 71–78.

Barocas, Solon, and Helen Nissenbaum. "Big Data's End Run around Anonymity and Consent." In *Privacy, Big Data, and the Public Good: Frameworks for Engagement*, edited by Julia Lane, Victoria Stodden, Stefan Bender, and Helen Nissenbaum, 45–75. Cambridge: Cambridge University Press, 2013.

Berg, Jessica Wilen. "Understanding Waiver." *Houston Law Review* 40, no. 2 (Summer 2003): 281–344.

Cane, Peter. *Responsibility in Law and Morality*. Oxford: Hart Publishing, 2002.

Gavison, Ruth. "Privacy and the Limits of Law." In *Philosophical Dimensions of Privacy: An Anthology*, edited by Ferdinand D. Schoeman, 346–402. New York: Cambridge University Press, 1984.

Kramer, Matthew H. *Moral Realism as a Moral Doctrine*. Chichester, UK: Wiley-Blackwell, 2009.

———. "Moral Rights and the Limits of the Ought-Implies-Can Principle: Why Impeccable Precautions Are No Excuse." *Inquiry* 48, no. 4 (2005): 307–55.

Marmor, Andrei. "What Is the Right to Privacy?" *Philosophy and Public Affairs* 43, no. 1 (Winter 2015): 3–26.

Nissenbaum, Helen. "Protecting Privacy in an Information Age: The Problem of Privacy in Public." *Law and Philosophy* 17, nos. 5–6 (November 1998): 559–96.

Reece, Andrew G., and Christopher M. Danforth. "Instagram Photos Reveal Predictive Markers of Depression." *EPJ Data Science* 6 (August 2017): 1–12.

Renzo, Massimo. "Rights Forfeiture and Liability to Harm." *Journal of Political Philosophy* 25, no. 3 (September 2017): 324–42.

Rumbold, Benedict, and James Wilson. "Privacy Rights and Public Information." *Journal of Political Philosophy* 27, no. 1 (March 2019): 3–25.

Thomson, Judith Jarvis. "The Right to Privacy." *Philosophy and Public Affairs* 4, no. 4 (Summer 1975): 295–314.

Wachter, Sandra, and Brent Mittelstadt. "A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI." *Columbia Business Law Review* 2019, no. 2 (May 2019): 494–620.

Westen, Peter. "Away from Waiver: A Rationale for the Forfeiture of Constitutional Rights in Criminal Procedure." *Michigan Law Review* 75, nos. 5–6 (April–May 1977): 1214–61.

Journal of Ethics and Social Philosophy
VOL. 22, NO. 2 · JULY 2022
https://doi.org/10.26556/jesp.v22i2.2488

INTRODUCTION TO "ACTION AND PRODUCTION"

Pamela Hieronymi

"ACTION AND PRODUCTION" was written by Stephen J. White, whose life was cut tragically short in the early spring of 2021, at the age of 38. He earned his BA at Pomona College, completed his PhD at UCLA, and joined the Department of Philosophy at Northwestern University in 2012, earning tenure in 2019.

White's research spans ethics, philosophy of action, political philosophy, and epistemology. By focusing on the distribution of responsibility—on who is responsible for what and, in particular, how we are distinctively responsible for our own lives—he unearths striking insights in each field. Topics include how to reason responsibly to conclusions about what to do, how predictions about your own future actions might figure into such conclusions, and how ethical considerations might be brought to bear when thinking, responsibly and with others, about what is true. Responsibility—in particular, *taking* responsibility—becomes, in his hands, not only a way to avoid alienation from oneself and others, but a way to find harmony in living as oneself with others.

I will briefly sketch some themes from his published work. In "Responsibility and the Demands of Morality" and "The Centrality of One's Own Life," White approaches the well-worn issues of how to integrate the impartial demands of morality with our particular personal interests, projects, and attachments in a novel way: by thinking about responsibility. He notes that responsibilities bring with them something like rights: if I have a duty to ensure that your life goes well, then I should also enjoy some say over how you choose to live it. Thus, rights limit responsibilities: your right to decide, for yourself, how to live your life brings with it limitations on my responsibility for whether it goes well. As he put it in an unpublished research summary, "If one's own judgment about what is worth doing is to have the right kind of authority in relation to one's decisions about what to pursue, one must assume the primary responsibility for one's own life."

Thinking about how responsibility is apportioned between people contin-

ued to guide White's thinking in other, even less expected topics. In "On the Moral Objection to Coercion," White argues that the wrong of coercion is best understood by noting that the coercer imposes an illegitimate responsibility on the coerced (e.g., a responsibility to ensure that the coercer does not physically assault the coerced). Two papers co-authored with Berislav Marušić—"How Can Beliefs Wrong? A Strawsonian Epistemology" and "Disagreement and Alienation"—present novel accounts both of doxastic wrongdoing and the significance of peer disagreement by attending to the ways in which reasoning to belief is often a shared activity, requiring shareable reasons and accountability to one another.

A remaining constellation of published papers concerns how we can take responsibility for our actions, given that they extend across time. "The Problem of Self-Torture" considers the person who smokes themselves to death with the thought that "one more can't hurt." Thinking well about what to do requires conceiving of one's actions as part of a larger whole. The nature of that part-whole relation is the subject of "Intention and Prediction in Means-End Reasoning," where White argues that means to our ends are to be chosen *qua* means. The question of how our reasoning about one action relates to our choice of others animates both "Transmission Failures" and "Self-Prediction in Practical Reasoning." The present publication, "Action and Production," extends this line of thought to the tricky question of how the recognition of the aptness of one's own practical reasoning might become part of that reasoning. He aims to bring into view "how an agent might recognize and be moved by the ethical value of performing a certain action for certain reasons, without thereby treating this as a further end to be promoted."

To close, and in honor of Steve, I want to echo a thought that was voiced repeatedly both at the workshop held in his honor at Northwestern and at his memorial in September 2021: Steve was resolutely focused on what is good. That is not to say that he was Pollyannaish; he saw clearly what is awful, and its importance, and even its urgency. However, if you spent time with Steve, he would inspire joy and laughter, wonder and interest. He would focus your joint attention on what was good. As Louis-Philippe Hodgson put it at Steve's memorial:

> Steve's career and Steve's life were cut short, and I don't think it would be at all in the spirit of Steve's outlook to pretend that this is anything less than tragic. But it is in the spirit of Steve's outlook to rejoice in the beauty of the life that he led, and in how much joy and light he brought to all who knew him, whether personally or professionally.

Philosophy was the third love of Steve's life. He is survived by the first two: his wife, Jessica, and his daughter, Lucy.

* * * * *

White's collected papers, both previously published and unpublished work, are planned for a volume titled *Responsibility and Alienation*, edited by Kyla Ebels-Duggan and Berislav Marušić.

University of California, Los Angeles
hieronymi@ucla.edu

REFERENCES

Marušić, Berislav, and Stephen J. White. "Disagreement and Alienation" *Philosophical Perspectives* (forthcoming).

———. "How Can Beliefs Wrong? A Strawsonian Epistemology." *Philosophical Topics* 46, no. 1 (Spring 2018): 97–114.

White, Stephen J. "The Centrality of One's Own Life." In *Oxford Studies in Normative Ethics*, vol. 7, edited by Mark C. Timmons, 229–50. Oxford: Oxford University Press, 2018

———. "Intention and Prediction in Means-End Reasoning." *American Philosophical Quarterly* 55, no. 3 (July 2018): 251–66.

———. "On the Moral Objection to Coercion." *Philosophy and Public Affairs* 45, no. 3 (Summer 2017): 199–231.

———. "The Problem of Self-Torture: What's Being Done?" *Philosophy and Phenomenological Research* 94, no. 3 (May 2017): 584–605.

———. "Responsibility and the Demands of Morality." *Journal of Moral Philosophy* 14, no. 3 (2017): 315–38.

———. "Self-Prediction in Practical Reasoning: Its Role and Limits." *Noûs* 55, no. 4 (December 2021): 825–41.

———. "Transmission Failures." *Ethics* 127, no. 3 (April 2017): 719–32.

Journal of Ethics and Social Philosophy
VOL. 22, NO. 2 · JULY 2022
https://doi.org/10.26556/jesp.v22i2.2487

ACTION AND PRODUCTION

Stephen J. White

Production-oriented theories of practical reason hold, roughly, that our reasons for performing an action ultimately depend on (1) how things are likely to turn out if we perform that action—what would happen, which states of affairs would result—and (2) what reasons we have to care about whether things turn out that way. And when it comes to choosing between two courses of action, the reasons to favor one over the other will depend on a comparison of the way things would go were one to perform one as opposed to the other.[1] This way of thinking about reasons for action can seem quite natural. Actions do make a difference in the world. They effect and prevent changes to the way things are, and it can hardly be denied that our reasons to care about the changes we effect or prevent by and through our actions are relevant when it comes to the issue of whether we should *do* various things.

Douglas Portmore, in a recent paper, puts the idea like this:

> If our actions are the means by which we affect the way the world goes, and if our intentional actions necessarily aim at making the world go a certain way, then it is only natural to suppose that what we have most reason to do is determined by which way we have most reason to want the world to go.[2]

The general picture looks even more promising once we notice that, even where there appear to be considerations that bear on what one should do independently of whether a given consequence is thereby brought about, it often seems to be

1 Many philosophers who explicitly discuss this conception use the label "teleological." See, for example, Scanlon, *What We Owe to Each Other*; Hurka, "Value and Friendship"; Wallace, "The Publicity of Reasons"; and Portmore, "The Teleological Conception of Practical Reasons" and *Commonsense Consequentialism*. But it is in some ways an unfortunate label. For one thing, the ethical theories of Aristotle and Aquinas deserve to be called "teleological" if any do. But not in the present sense. A better name would probably be the "production" view, after Aristotle's conception of production as that which is done for the sake of some further end. See Aristotle, *Nicomachean Ethics,* bk. VI, chs. 2, 4.

2 Portmore, "The Teleological Conception of Practical Reasons," 118.

a fairly trivial matter to redescribe our concern in terms of a wider event or state that *is* brought about (or hindered) via one's action. For instance, we may think that one has a moral reason to keep a promise even in cases where the advantages of breaking it outweigh the benefits to the promisee (and others) of keeping it. But we need not understand this as a case in which, even though the end being pursued is legitimate, a certain means (breaking one's promise) is ruled out—and ruled out for reasons that have nothing to do with the effectiveness of that means in realizing the end. We need not understand things this way because it is obvious that, if one breaks a promise, one thereby brings it about that the promise is broken. And the moral value of the promise may provide decisive reason to prefer the state of affairs in which one's promise is kept.

Once we acknowledge that this "consequentializing" move is available, it can be hard to see what sort of consideration or principle could possibly bear on how we should act that cannot ultimately be understood as a matter of our actions' efficacy in producing (or preventing) changes in the way the world is. It seems, then, that if we are asking after the point of performing a particular action, we should simply look to the way the world would be altered by its performance and ask what could be said for wanting the world to be that way.

In this paper I hope to undermine this production-oriented conception of practical reason. My argument has two main parts. First I will offer a counterexample and then argue that the most promising reply—one that employs the above "consequentializing" move—fails. Though I do think the counterexample has intuitive force, my primary aim in this part of the discussion will be to establish that the possibilities for making this consequentializing move are limited in a crucial respect.[3] In particular, while it may be possible for the production model to accommodate reasons to perform actions for their own sake, I will argue that one cannot employ the same strategy with respect to an agent's motives—the reasons for which she acts. It seems sometimes to matter how an agent is motivated. But the relevance of this for the agent cannot coherently be understood in terms of reasons for wanting it to be the case that she acts for certain reasons. This move is ruled out, in this case, by a constraint on what it is to be a reason—what I call the Deliberative Constraint on reasons.

3 At least this is so in the context of a theory of practical reason. "Consequentializing" is a term most frequently employed in discussions of moral theories conceived of as theories of right action. See, for instance, Dreier, "Structures of Normative Theories"; Portmore, "Consequentializing Moral Theories"; Brown, "Consequentialize This"; and Schroeder, "Teleology, Agent-Relative Value, and 'Good.'" Because my arguments below depend on the nature of reasons and reasoning, they do not bear directly on the consequentializing project as applied to theories of the right. The implications of this paper for that project will depend upon one's view of the relation between moral rightness and practical reason.

The second main task of the paper is to identify a regulative principle of action that is not reducible to any concern having to do with one's efficacy in producing or preventing various outcomes (including those that consist in the performance of some action). The principle I identify is one that enjoins us, as rational agents, to act *for* the reasons we judge to decisively support our actions. I argue that this principle best explains the rational failure involved in certain cases of self-manipulation—a type of rational failure that we cannot explain if we assume a production-oriented account of practical reason.

1. Production Theories of Practical Reason

Let us begin by being more precise about what constitutes a production theory of practical reason. We can usefully distinguish two components of the view.[4] The first is a thesis stating necessary and sufficient conditions for reasons for action. (As I will use the term, an "outcome" associated with an action may cover any event or state of affairs resulting from that action's performance.)

> *Equivalence Thesis*: (1) One has a reason to perform some action, ϕ, in circumstances C if and only if (a) one has a reason to want some type of outcome to obtain, and (b) one's ϕ-ing in C would result in an outcome of that type; (2) one has more reason to ϕ than to ψ in C if and only if one has more reason to want the outcome associated with one's ϕ-ing than to want the outcome associated with one's ψ-ing.

The second component is a thesis about the dependence of reasons for action on reasons for preferring that certain outcomes obtain:

> *Dependence Thesis*: (1) If one has reason to ϕ in C, this is in virtue of the fact that one's ϕ-ing in C would result in a type of outcome one has reason to want; (2) if one has more reason to ϕ than to ψ in C, this is in virtue of the fact that one has more reason to want the outcome associated with one's ϕ-ing than the outcome associated with one's ψ-ing.

It is important to note that, at this level of generality, nothing is implied about what kinds of reasons an agent might have for wanting some outcome to obtain. In particular, this picture is neutral on the questions of whether the rational ranking of outcomes needs to be in any sense impartial, or whether the event or state of affairs that one has most reason to promote is the one that is imperson-

4 These theses are based on Portmore's characterization of the position in "The Teleological Conception of Practical Reasons," 120.

ally best. Such views will count as production theories, but they import further assumptions that the general conception is not committed to as such.[5]

It is also important to be explicit that the notion of *outcome* as it is deployed in the two theses is broader than the ordinary notion. In particular, as mentioned in the introduction, some accounts will construe the outcome associated with an action so as to include the performance of the act itself.

The production conception is perhaps most naturally associated with traditional consequentialist views in ethics and with standard decision theory.[6] But because of its capaciousness about the types of outcomes that may be counted as relevant, the view is much more widely held and cuts across some traditional ways of drawing the distinction between consequentialist and deontological moral theories.[7] Rawls, for instance, categorizes moral theories as deontological if they do not define the right in terms of some prior conception of the good.[8] This does not tell us anything about how the right is related to one's reasons for action. It is easy to imagine a view according to which (a) the right is prior to the good in Rawls's sense, (b) one has reason to prefer outcomes in which one acts rightly to outcomes in which one fails to act rightly, and (c) this is what explains why one has reason to do the right thing.[9]

5 Cf. Scanlon, *What We Owe to Each Other*, 80–81.

6 I do not mean to suggest that all consequentialists accept the production view of practical reason. Some are not interested in practical reason at all. Moreover, to move from a theory like utilitarianism or decision theory to a particular production-oriented account of reasons for action, one would need to make some further assumptions. For example, since classical decision theory does not rationally evaluate preferences one by one, we would have to understand claims about what one has *reason* to prefer as determined by its consistency (in the sense defined by decision theory) with one's other preferences.

7 Thomas Nagel, for example, is explicit that reasons for action always take the form of reasons to promote events (including sometimes the event consisting of a certain action being performed), though he goes on to reject consequentialism as the correct account of morality. See Nagel, *The Possibility of Altruism*. And Portmore's project is to show how the moral permissions and constraints recognized by common sense are compatible with the production conception of reason, precisely in order to defend the view that we always have most reason to do what morality requires. See Portmore, *Commonsense Consequentialism*. For earlier discussion relevant to this issue, see Broome, *Weighing Goods*; and Dreier, "Structures of Normative Theories."

8 Rawls, *A Theory of Justice*, 24–26.

9 There are also theories that appear to endorse a production-oriented view of reasons *for* action, while incorporating negative side constraints as among the reasons that count *against* this or that action. Even if, on such views, the moral reason not to kill an innocent person has nothing to do with the reasons to want certain outcomes (not) to obtain, the positive reasons in favor of performing particular actions will all depend on the desirability of the re-

2. THE IDEA OF EXCLUDED REASONS AND A COUNTEREXAMPLE

Assuming the basic production conception, the reason to perform any particular action is that it will be a (perhaps constitutive) means of producing an event or state of affairs that one has some reason to want realized. And if one has a reason not to perform some action, this reason is given by the fact that performing the action would produce an event or state of affairs that one has some reason to want not to be realized.

Given this structure, the production conception yields the following principle:

> *Excluded Reasons*: If there is any constraint on the kinds of considerations on which we can legitimately act, this constraint must be explicable in terms of a constraint on the considerations that can serve as bases for wanting a given outcome to obtain.[10]

In this section I will suggest that there are sometimes constraints on what can count as a reason for action that do not apply to reasons for desiring outcomes. If this is right, and the Excluded Reasons principle does not hold for these cases, it would undermine the Equivalence Thesis.

I will begin by considering an argument of Scanlon's against the production view, an argument that relies on distinguishing different ways in which considerations can bear on what a person should do.

Scanlon thinks that the structure of practical reasoning is more complex than would be allowed by the thesis that all reasons for action are derived from the desirability or undesirability of various states of affairs. This complexity is due to the possibility of reasons that, in Scanlon's words, bear "not on the desirability of outcomes, but rather on the eligibility or ineligibility of various other reasons."[11] Examples of this phenomenon include what Joseph Raz calls "exclusionary reasons."[12] Consider Raz's view of promising, for instance. A promise, according to Raz, serves not only as a reason that counts in favor of performing the promised action, but also as a second-order reason excluding consideration of other first-order reasons, reasons that would otherwise be relevant in the situation. For instance, if I promise to read your manuscript and get you comments by the end

---

sulting outcomes. The argument from self-manipulation, which I present below, will apply also to such "constrained" production views.

10 In the next section, I will consider and reject an argument to the effect that the production view does not in fact entail the Excluded Reasons principle.

11 Scanlon, *What We Owe to Each Other*, 84.

12 See Raz, *Practical Reason and Norms*.

of the week, then not only do I have reason to do this, but I have reason to treat a range of other considerations as simply irrelevant to the question of whether or not I should do this. Although reading your manuscript will mean I will have less time to get my own writing done, and although my interest in having more time to write would normally be a reason not to spend so much time commenting on your work, my promise to you does not merely compete with the reason I have to spend a significant portion of my time writing, perhaps outweighing it. The fact that I have made the promise means that I should not regard the time it will take away from my own work as a factor to be considered at all in deciding ultimately whether to fulfill my promise. The promise simply renders this consideration irrelevant. So says Raz.

Let us grant for now that practical reasons do sometimes exhibit this complex structure. Sometimes factors that would normally be relevant are excluded from consideration. Is this incompatible with the production account of reasons for action? Scanlon seems to suggest as much in the passage quoted above. One thing he may have in mind is that, if we accept a view according to which the reasons for and against, say, keeping one's promise derive from the desirability of the respective outcomes of doing so versus doing something else, we would be committed to a simple weighing model of promissory obligations. Thus, the production conception would falsely imply that certain considerations (like my interest in having time to write) are relevant to a decision (even if they were ultimately outweighed) in situations where, in fact, they are not relevant at all.

But the production conception of reasons does not entail a simple weighing model of (e.g.) promissory obligations. For it is at least formally open to the production theorist to say that, where one has made a promise, then, because of this, certain considerations will be irrelevant to the appropriate preference ranking of possible outcomes. Thus, if Raz's model of promises is correct, it is possible that this is in virtue of the correctness of an analogous model according to which promises provide exclusionary reasons for preferring certain outcomes to others. My interest in spending more time writing is neither here nor there when it comes to the question of whether I should keep my promise to read your manuscript. But, the production theorist will say, that is because it is neither here nor there when it comes to the question of whether I should want it to be the case that I keep my promise rather than break it.[13]

13 Though this move is formally open on the production conception, I think it should strike us as a fairly counterintuitive view of promise keeping. To think that what explains the force of one's reason to do what one has promised to do is that one has special reason to want the world to go a certain way—namely, in such a way that one keeps one's promise—seems to get things backward. Rather, it seems to me that if I want it to be the case that I have kept

Nevertheless, there seem to be cases in which, although the above response is formally open to the production theorist, in fact there are constraints on our reasons for action that do not apply to our reasons for wanting the associated outcomes to obtain. That is, there are cases in which the Excluded Reasons principle does not appear to hold.

Consider that the sort of practical reasoning that is appropriate to social or institutional roles often involves bracketing certain considerations that would otherwise be relevant.[14] For example, imagine you are a state official charged with awarding a contract for a large public project. Given your position, you ought to treat certain considerations as irrelevant to your decision about whom to award the contract to. The fact that you are old friends with the owner of one of the firms vying for the job, for instance, should play no role in your deliberation. Now, the question is whether, in this sort of case, your reasons for wanting or hoping for a certain outcome should be constrained in the same way.

It seems clear that the requirement of impartiality here does not limit the kinds of considerations that may properly ground one's desires or preferences for outcomes in the same way that it limits the appropriate grounds for action. In contexts where this sort of impartiality is called for, it may be very important that a given consequence of your action be at most a foreseeable side effect. But this does not imply that it contributes nothing to the desirability for you of the outcome. The responsible exercise of public authority does not require that you refrain from giving the job to someone who happens to be your friend. It merely requires that you not allow the fact that she is your friend to influence your decision. But it seems perfectly consistent with the requisite degree of impartiality to hope that in the end it is your friend who wins the contract—and to hope for this result because she is your friend and therefore someone you should want to see succeed. It does not amount to cronyism to want your friend to do well or

---

my promise, this is because I recognize that that is something I have reason to do. And indeed, once we reverse the order of explanation, it is easy to become suspicious of the claim that doing what one has promised to is, in itself, and independently of any more "concrete" effects, a desirable way for the world to go. Suspicions only grow once we notice that one apparently has no comparable reason to care about whether others keep their promises. Thus one may aim at outcomes wherein one keeps one's own promises, though this results in several other promises being broken by other people. And does this not reveal our commonsense views about promissory obligations to be paradoxical? It seems to me that, on the contrary, what we should be suspicious of is the initial move—the explanatory reversal—that set us on this path to paradox. However, my main line of argument does not rest on the intuitions expressed in this note. On the supposed paradox of deontological constraints, see Scheffler, *The Rejection of Consequentialism*.

14 Scanlon, *What We Owe to Each Other*, 52.

to be glad that, as it has turned out, you were able to award the contract to her—just so long as you did not award the contract to her *because* she was your friend.

3. MOTIVES AS "OUTCOMES"

There is a natural response to this sort of case. The production theorist may complain that we have not fully specified the outcomes to be compared and that once we do we will be able to capture the relevant complexity involved in such practical reasoning in a way that is consistent with the production conception. The relevant outcome associated with your act, the suggestion goes, should include not only the performance of the act itself (e.g., awarding the contract to your friend), but also the motives that led to its performance.

The idea, applied to our example, is that there would be some unfairness, not in awarding the contract to your friend per se, but in awarding her the contract in part because she is your friend. But this, it will be said, can be framed as, in the first instance, a point about the outcomes you have reason to prefer. You should not take your friendship with a certain person to be relevant in considering whom to give the job to because you should prefer the state of affairs in which your decision is based solely on the qualifications relevant to the project over states of affairs in which you take into account other sorts of reasons, such as that one of the applicants is a friend of yours.

This expanded view of the types of events and states of affairs whose desirability is relevant in practical reasoning appears to allow for a version of the production theory that is compatible with both of the following claims:

1. The fact that someone (call her Green) is a friend of yours is a reason to prefer an outcome in which you hire her for the job.
2. The fact that Green is a friend of yours is *not* relevant to your decision about which applicant to hire for the job.

How is this possible on the production conception? Notice that the sense in which 2 is true cannot be that your friendship with Green provides literally no reason to hire her. If that were the proper interpretation of 2 then it, together with 1, would imply that the Equivalence Thesis was false. Rather, the thought is that 2 can be interpreted so as to follow from the following further claims about the relative desirability of various outcomes:

3. You have reason to prefer an outcome in which your decision is based only on the professional qualifications of the candidates, and Green is

offered the contract, to an outcome in which Green is offered the contract in part because she is your friend.

4. You have reason to prefer an outcome in which your decision is based only on the professional qualifications of the candidates and Green is *not* offered the contract to an outcome in which Green is offered the contract in part because she is your friend.

Claims 3 and 4 are consistent with 1 and seem to provide a plausible interpretation of the claim (in 2) that your friendship with Green is not relevant to your decision. Your friendship with Green is irrelevant in the sense that, whether or not in the end you should choose Green's firm for the job, the outcome of doing so on the basis of deliberation that gives weight to your friendship will always be worse than other outcomes it is in your power to realize.[15] All of this is perfectly in keeping with the production conception of reasons for action.

The response being considered, in effect, denies that the production conception commits us to the Excluded Reasons principle. We can, consistently with the two theses that constitute the production theory (Equivalence and Dependence), interpret constraints on the reasons for which one may legitimately *act* without denying that those reasons are genuine grounds for wanting certain outcomes to obtain. The claim is that this is possible because we can account for the former constraint having to do with action, not by denying that the "excluded" reasons are genuine reasons to act, but by appeal to entirely different reasons—reasons that bear on the desirability of the outcomes associated with the agent's motivations.

I do not think, however, that this is a tenable response. It might secure the Equivalence Thesis, according to which your reasons to want certain outcomes to obtain perfectly correlate with your reasons to act in certain ways. But it does so at the expense of the Dependence Thesis. We cannot explain why some consideration provides you with a reason to act in terms of the desirability of an event that consists in your acting for that reason.

The proposal under consideration is, again, that although you have reason to want Green to get the contract because she is your friend, all things considered, the outcome you have most reason to prefer is that you award the contract to the best qualified candidate (whomever that turns out to be) and do so solely because they are the best qualified. The proposal depends on a further turn of the consequentializing screw—a further expansion of the types of "outcomes" that can occupy the position of a reasonable goal to be intentionally pursued. That

15 Compare Portmore's response to Scanlon's tennis player example in Portmore, "The Teleological Conception of Practical Reasons," 137.

is, we have reached an iteration of the production theory according to which the agent, in considering what to do, will need to consider her own motives as potentially relevant components of the outcome she has reason to promote or prevent.

I do not want to deny that one might have good reason to want it to be the case that one acts on the basis of some considerations as opposed to others. The question is whether it makes sense to say that this gives one reason to act on the basis of those considerations—for instance, to hire a candidate on the basis of their qualifications alone, leaving aside personal connections. The fundamental objection I wish to press against this account is that it gets the order of dependence backward. In general, we have reason to want it to be the case that we act on the basis of a certain consideration because that consideration really does give us good reason to act, and not the other way around.

Let us begin with a general point. Even if it is true that it would be good or desirable or useful in some respect for a person to ϕ, under the circumstances, this does not necessarily constitute a *reason* for that person to ϕ. For one thing, whether she ϕ's or not may not be in any sense up to her. So we need to ask: What more is needed, then, to establish that a given consideration provides a reason for a person to ϕ, beyond its indicating that her ϕ-ing under the circumstances would be in some way good or desirable? I propose we look to the potential role that consideration could intelligibly play in the agent's reasoning or deliberation about whether to ϕ. As John Searle puts it, "You have to be able to *reason* with reasons."[16] Along these lines, I offer the following, fairly weak condition on something's being a reason:

> *Deliberative Constraint*: *R* is a consideration that provides a reason for *S* to ϕ only if it is possible that *S* could, without irrationality, take a consideration of type *R* to support ϕ-ing in the course of deliberating about whether to ϕ.

Now, we have been considering a proposal on behalf of the production theorist about how to explain why it might be that an agent should ignore certain consid-

16 Searle, *Rationality in Action*, 104. See also, Williams, "Internal and External Reasons"; Korsgaard, "Skepticism about Practical Reason"; Kolodny, "Why Be Rational?"; Shah, "A New Argument for Evidentialism"; Setiya, *Reasons without Rationalism*; and Raz, *From Normativity to Responsibility*. Some of their formulations of the connection between reasons and reasoning are stronger than others. Williams's and Shah's versions, for instance, seem to require that, to be a reason, a consideration must be able to explain what an agent does given the agent's actual attitudes and dispositions. The version I formulate in the text is much weaker and does not imply that whether some consideration provides a reason for an agent depends on the contents of that agent's subjective motivational set.

erations—like your friendship with one of the bidders for a contract—considerations that, by the production theorist's lights, should, strictly speaking, provide the agent with reasons to act. The proposal was that we can appeal to the fact that the agent has good reason to want, as an outcome of her actions, that she acts on the basis of certain considerations as opposed to others. I will now argue that such a fact (or the agent's recognition of that fact) could not satisfy the Deliberative Constraint on reasons, and so cannot constitute a reason for action.

An initial problem with the proposal is that it looks like the very attempt to promote the relevant outcomes, in response to one's recognition of their desirable features, would undermine any possibility of success. To see this, let us return to our example. You are to decide, let us suppose, between two architects bidding on a state contract: Green and Brown. The project-related considerations—proposed designs, projected costs of the proposals, etc.—appear to favor Brown (if only slightly). On the other hand, Green is an old friend of yours. Now, we are supposing that the question relevant to your deliberation about what to do is *not* whether your hiring Brown would be a more desirable state of affairs than your hiring Green. Rather, the question is whether your hiring Brown solely on the basis of his qualifications (call this outcome *B*) would be a more desirable state of affairs than your hiring Green partly on the basis of your friendship with her (call this *G*).

Let us grant that you have more reason to want *B* to obtain than *G*. How are you supposed to bring it about that *B* obtains? It might seem easy. Once you realize you have reason to prefer *B* to *G*, you just award the contract to Brown. But doing this does not in fact bring about outcome *B*. Rather, it brings about a distinct outcome, viz., the event consisting in your hiring Brown on the basis of your recognition that hiring Brown on the basis of his qualifications is preferable to hiring Green because she is your friend. To hire Brown for the reason that outcome *B* is preferable to outcome *G* is to hire him for a different reason than that specified in outcome *B*, and so fails to realize the latter outcome. Thus, the attempt to act so as to produce the outcome you have most reason to want—if it is undertaken on the grounds that it is the outcome you have most reason to want—is self-undermining.

The general problem is that one's own motivations are not themselves directly subject to one's choice or will. One cannot come to be motivated by a consideration through one's judgment that *being so motivated* would be a good thing.[17] Rather, for the consideration to motivate, insofar as you are rational, you must take it to count in favor of the action you are deciding on. Of course, you

17 Cf. Velleman, "Deciding How to Decide"; Hieronymi, "Controlling Attitudes."

might hire Brown on the basis of his qualifications not because you take this complex outcome to be preferable to the alternatives, but simply because you did not think about your friendship with Green, say, or because you overestimated Brown's qualifications relative to Green's. In that case you would have brought about the outcome you had reason to want. But you would have done so by sheer luck—not in response to the reasons you had for preferring that outcome.

One might reply that, though it may not be possible to immediately determine one's motives at will, one can act indirectly, arranging things so as to ensure that one has the motivations one wants to have. One could arrange to be deprived of certain information, for instance, or even, with some ingenuity, be provided with misleading information.[18]

This will not be enough, however, to capture the intuition behind the cronyism counterexample. For we can imagine that you, in your role as a public official, had no feasible indirect means of bringing it about that your motives were strictly impartial. If you had had such an opportunity, then the production theorist might be right to insist that you take it. But if we stipulate that no such opportunity was available, then, given the Deliberative Constraint on reasons (along with our earlier assumptions about the case), the production view implies that you have most reason to award the contract to Green because she is your friend, your role as a public official notwithstanding. And this does not seem right. Your mistake consists not only in your failure to take opportunities (supposing you had them) to manipulate your motives; it consists in your giving favor to your friend.

Even if we set aside the question of *how* to go about producing the motives in oneself that one supposes would be desirable to have, there is a more fundamental problem with this way of understanding practical reasons. The problem, as I will now argue, is that it implies that the question of what one has good reason to do is never directly relevant to one's decision about what to do. To explain this implication, I need to make two intermediate points about the sort of production-oriented account that includes the agent's motives among the outcomes whose desirability furnishes reasons for action.

The first point is that this view generates a type of regress. Suppose that considerations A and B provide reasons for desiring the outcome associated with performing an action, ϕ. As we have seen, on the current proposal, this does not

18 This would be to exercise what Hieronymi calls "managerial control" over one's mental states. Managerial control is analogous to the kind of control we exercise over objects in our environment and contrasts with the kind of activity involved in committing ourselves to some course of action (intention formation) or to the truth of some proposition (belief formation). See Hieronymi, "Controlling Attitudes."

show that I should ϕ for these reasons. It may be that some further consideration, *C*, gives me reason to prefer the outcome associated with ϕ-ing on the basis of *A* alone. Let us grant for the sake of argument that there is some way for me to bring it about, in response to *C*, that I ϕ solely on the basis of *A*. But now another question immediately arises: Whatever I propose to do to bring it about that I ϕ on the basis of *A*, should I do *that* on the basis of *C*? *C* shows that ϕ-ing on the basis of *A* is a desirable outcome, but it is an open question whether I should treat *C* as a reason to promote that outcome. This will itself depend on whether there are good reasons to desire the outcome associated with this complexly motivated action. But then, of course, the same sort of question will arise again. Call this the regress point.

The second point I want to raise is that there is simply no necessary connection between the desirability of the outcome associated with ϕ-ing on the basis of *R* and *R*'s being a consideration that counts in favor of ϕ-ing. There is nothing to rule out, for example, that the outcome associated with *drinking a cup of coffee on the grounds that one loves Sophocles* might be preferable in certain respects to other outcomes one might promote.[19] But this is obviously not enough by itself to show that the fact that one loves Sophocles is a good reason to drink coffee. There is no rational connection between these two things. (Indeed, we could imagine situations where the lack of a rational connection is part of what explains why drinking the coffee on this basis is a desirable thing to have occur.)

Now, the production theorist will presumably agree that one's love of Sophocles is no reason to drink coffee. But she will insist that although the fact that I love Sophocles is not a good reason for me to drink coffee, it *could* nevertheless be the case that I have good reason to drink coffee on the grounds that I love Sophocles. This means, however, that the fact that some consideration is *not* a good reason to perform a given action does not warrant me in concluding that I should not treat it as one, even a decisive one. The question of what I should take into account in deliberating is not settled by consideration of which factors provide rational support for the various courses of action I am considering. Call this the irrational-motivation point.

Let us put these two points together. The irrational-motivation point shows that, on the current view, there is no constraint on the agent to give weight to, or choose on the basis of, only good reasons—reasons that genuinely bear on the object of her choice. For there may be reason, one level up, to think that it would be preferable for her to act on the basis of irrelevant reasons. The regress point, however, shows that the same could be true at that second level as well.

19 The example comes from Raz, *Engaging Reason*, 8.

And indeed, at each point at which the agent might consider the practical question of whether some outcome is worth promoting, it could in principle be that it would be best if the agent were actually to *settle* that question on completely irrelevant grounds. It follows that at no stage of the deliberative process will the agent reach a point where what is directly relevant to how she should conclude her deliberation is what she has good reason to do. For a conclusion about what one has good or most reason to do—in the sense that the balance of relevant considerations favors one of the available courses of action—settles nothing. Such a conclusion is, according to this view, compatible with its being the case that one has good (higher-order) reason to act on the basis of some irrelevant (or rationally overridden) consideration, and thereby be led to do something other than what one has good (first-order) reason to do. And yet, this conclusion—that one has good reason to do such and such on the basis of such and such considerations—does not settle the matter either. It is possible that one has good reasons to ignore the considerations that led to that latter conclusion (the regress point again). At no point will an answer to any of the questions in this series license one to go ahead and form the intention to promote the relevant outcome.

This consequence seems sufficient to rule out any attempt to derive reasons for action from the desirability of states of affairs consisting in the agent's having acted on the basis of certain reasons or motives. A theory of practical reason cannot be coherent if it implies that the question of what I have good reason to do is never directly relevant to first-person deliberation.

Let me summarize our progress so far. I raised a counterexample to the Equivalence Thesis based on the nature of our objection to cronyism by public officials. It seems very plausible that personal ties may give a public official good reason to desire a certain outcome, without thereby giving her any reason to do what is in her power to effect that outcome. The wanted outcome should be, at most, a side effect of the official's actions. I then considered the suggestion that we can save the Equivalence Thesis, and so the production conception, by analyzing the case as one where the official has reason to treat her personal connection as irrelevant, not because, strictly speaking, it is not a reason for her, but because she has reason to want it to be the case that she is motivated differently. If we include the agent's motives as possible components of the "outcomes" mentioned in the Equivalence Thesis, the thought went, we can see that our anti-cronyism intuitions are consistent with the production conception after all.

But we should reject this strategy for responding to the cronyism counterexample. There are good objections to any view that attempts to derive reasons for an agent to be motivated by certain considerations from the fact that the agent

has reason to want to be so motivated. Such a view both violates the Deliberative Constraint on reasons and yields an account of practical reason and deliberation on which the question of what one has good reason to do is never treated as directly relevant to one's decision about what to do. The production theorist must therefore abandon at least one of her theses. Either there are considerations (such as one's role as a public official) that bear on how we should conduct ourselves, though they do not bear on the desirability of associated outcomes—and thus the Equivalence Thesis is false—or, if they do bear on the desirability of the outcome (given that we are conceiving of the outcome as consisting in one's acting for certain reasons), then it cannot be *in virtue of this* that they provide reasons to act—and thus the Dependence Thesis is false.

Of course, it remains open to the production theorist to reject the substantive normative intuitions motivating the example—for instance, that one ought not to take into account personal affections in the administration of public duties. Nevertheless, we have achieved an important result, which is independent of substantive ethical intuitions. What the above arguments show is that there is a limit to what can be built into the "outcome" associated with an action for the purposes of deriving reasons for action. If there are cases where one has reason to bring it about (indirectly) that one acts from a certain motivating reason, irrespective of whether that motivating reason rationally supports the action, such cases will necessarily be exceptional. It cannot in general be an open question whether acting *for* the reasons one rightly takes to count in favor of one's decision is itself something one has reason to do. If there is a sense in which one ought to act for certain reasons on a given occasion, this will not (except perhaps in abnormal cases) be explained by the fact that the state of affairs constituted by one's having acted for these reasons is, on balance, to be preferred to the alternatives. This is a result that I will rely on in what follows.

4. SELF-MANIPULATIVE ACTS

I turn now to a different kind of case that raises problems for the production conception. Often, an agent will have reason to want herself to act in a certain way—either because her acting in that way will have further desirable consequences or because it will itself be a desirable outcome (or both). On the production conception of reasons, this will give that agent some reason to perform the action in question. But it may also provide some reason for her to perform a different action, by which she can get herself to perform the first. One can sometimes bring it about that one does something, not simply by doing it, but less directly, through various forms of self-manipulation, self-inducement, or

self-entrapment. It seems to me evident that, in certain cases, employing such indirect methods of bringing oneself to act is irrational. What I hope to show is that the best explanation of the irrationality in these cases appeals to the fact that, although the agent's self-manipulative act produces a result she has reason to want, this does not give her any reason to act in that way. Indeed, I will go even further and argue that the agent's failure to recognize this fact betrays an incomplete grasp of the concept of a reason for action.

Consider an example. Marge sees that she has hardly any food in her refrigerator and that she therefore has reason to go to the grocery store. She realizes that, if she smokes her last two cigarettes, she will be out and will need to go to the store to buy another pack. Once she is there, she knows that she will also purchase the groceries she needs (in the past, nearly every time she has stopped at the grocery store to buy cigarettes, it has occurred to her to take advantage of the opportunity to stock up on some groceries as well). And so she decides to smoke her last two cigarettes in order to bring it about that she buys groceries.

We can imagine circumstances in which this would be a sensible way for Marge to proceed. For example, she might be worried that if she does not smoke her remaining cigarettes she will forget to go to the store, or succumb to laziness or agoraphobia. But the point I want to make now is that, the less concerned she is with such things, the less intelligible her choice seems. Marge's smoking to get herself to go grocery shopping appears permissible only if she expects that she would not otherwise go.

To bring this out, let us stipulate that Marge is not worried that if she does not smoke she will not shop. She does not anticipate any rational or agential defect that might otherwise prevent her from getting herself to the store. She merely sees two options she could take, both of which she is reasonably sure will result in her buying groceries. One option is to simply get in her car, drive to the store, and buy groceries. The second option is to smoke her last two cigarettes, which she is confident will lead to her buying groceries. Given this stipulation, it seems clear that there is something irrational about Marge's behavior. Her choice makes no sense.

Notice, however, how nicely Marge's reasoning fits with the production view. She recognizes that she has reason to want a certain state of affairs to obtain—namely, that she goes to the store and buys groceries. And she recognizes that by acting in a certain way (smoking her last cigarettes) she can promote that state of affairs. What she does, she does in order to bring about an outcome she has reason to want. So what is wrong with Marge's reasoning here? Why should it count as less than fully rational?

There are three possible explanations we should consider. One is that the de-

sirability of Marge's buying groceries simply does not give Marge any reason to smoke her cigarettes, despite the fact that it would lead her to buy the groceries. A second possible explanation would be that, while the desirability of her grocery shopping does give her a reason to smoke, there are much stronger reasons for her to go directly to the store instead. On this account, the overall balance of reasons is not sufficient to support her smoking the cigarettes. Finally, we might try to explain Marge's irrationality not by considering the reasons she has (or lacks) for the various courses of action she might take, but rather by citing her failure to conform to some rational requirement of coherence or consistency among her attitudes.

If we accept the production conception of reasons, we will not have recourse to the first of these explanations. This is because, on that conception, action is judged ultimately in terms of its (probable) efficacy in bringing about this or that outcome. And, by hypothesis, the case is one in which (a) the intended outcome of Marge's action is one she has reason to want, and (b) she has correct causal beliefs about what she can do to promote that outcome.

It might be suggested that this is wrong, that Marge's buying groceries is a voluntary action and as such cannot be understood as a causal product of her earlier actions. Her act of smoking cigarettes, on this view, would not contribute to bringing about the state of affairs in which she buys groceries, and therefore the desirability of the latter would not provide any reason for the former. This libertarian conception of action, according to which freely willed acts are not caused or explained by prior events, is probably not one that will be embraced by many production theorists. But even if we were to accept such a metaphysics of free will, we should still take Marge's smoking to play a non-superfluous role in bringing about the outcome in which she buys groceries. By smoking, she provides herself with a motive or incentive to go to the store she would not otherwise have—an incentive we can assume she freely responds to. Whatever the precise nature of the influence this incentive has on her will, the earlier act of smoking was clearly not a redundant factor in bringing her to act as she does. Without it, she would not have the reason to go that she subsequently acts on.

By the lights of the production view, then, Marge is right to think that the desirability of her buying groceries gives her at least *a* reason to smoke her cigarettes. Therefore, if the production conception is correct, there must be some other explanation of Marge's irrationality. I will now argue that neither of the potential alternative explanations is adequate.

Let us turn to the second reasons-based explanation. Is there some special reason Marge has not to get herself to shop by way of smoking her remaining cigarettes? Is there, for instance, some further interest of hers (beyond her need

for groceries) that would be better served by going directly to the store, rather than smoking?

Well, one obvious factor is that smoking is bad for Marge's health. Isn't this a reason not to take this route to the store? This suggestion, however, pretty clearly does not get to the heart of the matter. For one thing, it would be easy to alter the example in order to eliminate such health-related concerns. Would it really be any less crazy for Marge to chew her last stick of nicotine gum (or brew her last cup of green tea) in order to promote her purchase of groceries?

But the more important point is this: if the issue is that Marge has health-related reasons not to smoke, then these would equally count against a decision to smoke a couple of cigarettes that is entirely independent of her decision to go grocery shopping. She might decide to go shopping and then decide that, before she does, she will have a smoke. The health-related concerns would kick in here as well. But this choice, and the deliberation leading to it, do not seem irrational in the same way as in the original case. Her decision might be unwise, but it is intelligible. What this brings out is that the health-related reasons do not speak to the apparent irrationality of smoking *as a way* of bringing it about that she buys groceries. But this is what needs to be explained. The problem is with the rational connection between the proposed means and the end. The fact that there is some additional objection to the act she adopts as her means will not help us understand that problem.

The appeal to the health-related costs of smoking to explain where Marge has gone wrong appears to miss the point. And the problem with this explanation will generalize to any appeal the production theorist might make to reasons Marge has not to get herself to go buy groceries in the way she does. Any appeal to the additional disvalue or cost, whatever form it takes, that Marge will incur by smoking rather than simply going, will constitute an objection to her overt behavior—smoking, then going to the store, then buying more cigarettes along with groceries—regardless of whether she smokes *in order* to get herself to buy groceries. But, intuitively, the problem with Marge does not surface until we look beyond her outward behavior to the structure of her motivations.

Perhaps it will be suggested that this is not so. The reason Marge has not to smoke is precisely that it is inefficient *as a way* of achieving the desired result that she buys groceries. But this does not really help. To say that it is an inefficient way of achieving the result is to say that there is an alternative way of achieving the same result (buying groceries) that carries fewer costs. It is the avoidance of these costs that matters. In other words, the reason to take the more efficient rather than less efficient means to an end is that it helps minimize the negative

impact on one's other ends and interests.[20] But again, this is a matter of what one does, not why one does it.

At this point, the production theorist might try to offer a more direct explanation of Marge's mistake. Perhaps the event of Marge's *smoking in order to bring it about that she buys groceries* can itself be construed as an "outcome"—a way the world might go—that is on balance undesirable. The objection to Marge would then be that she chooses to act in such a way as to bring about this undesirable outcome.

This, however, is just a version of the move I argued against in the previous section. It appeals to an "outcome" consisting in an agent's acting on the basis of certain reasons or motives. And though there is nothing wrong in this context with allowing that there is a broad sense of "outcome" that covers this sort of thing, the move violates a constraint on the types of outcomes that can sensibly figure in the Dependence Thesis.

Consider, then, the other sort of explanation a production theorist might offer to account for Marge's irrationality. Perhaps she has violated some standard of rationality that applies to her reasoning, or to the way her attitudes cohere with one another, but that does not impugn the claim that she had a reason to smoke. It is possible, after all, to be irrational in forming an intention to do what one in fact has reason to do.[21]

The problem for this type of explanation will be to identify the relevant rational requirement that Marge is violating in this case. There does not appear to be any inconsistency among her beliefs, or between her beliefs and her intentions. She is not intending means she knows to be insufficient for her end.[22]

It is tempting to think that Marge is somehow *akratic* or weak willed. To the extent that this is plausible, I think it is because it seems so obvious that she should not smoke as a way of bringing herself to go to the store unless for some reason she cannot bring herself to go to the store immediately. But this is the very intuition we are trying to explain. To appeal to it in order to support a charge of *akrasia* would beg the question.

There is also a principled objection to the production theorist's appeal to this third way of explaining the case. It is extremely plausible that if an agent arrives

20 Cf. Korsgaard, "The Myth of Egoism."

21 There is a large literature on the distinction between what rationality requires of one and what one has reason to do or think. See, for example, Broome, *Rationality through Reasoning*; Brunero, "The Scope of Rational Requirements"; Kolodny, "Why Be Rational?"; Parfit, "Rationality and Reasons"; and Scanlon, "Structural Irrationality."

22 For discussion of these and related requirements of practical rationality, see Bratman, *Intentions, Plans, and Practical Reason*.

at justified conclusions about what she has good reason to do, and decides what to do on that basis, then she has reasoned well. But if we assume the production view is correct, then this is just what Marge has done. This would mean that the production conception is incompatible with even this minimal connection between good practical reasoning and the normative reasons that apply to an agent. To offer an account of reasons for action that had so little to do with practical thought and deliberation would be tantamount to changing the subject.

I conclude that the best explanation of what has gone wrong in Marge's case is that the desirability of her buying groceries gives her no reason to smoke her cigarettes. And this is incompatible with what a production-oriented account would imply about the reasons Marge has in this case.

5. ACTING FOR THE RIGHT REASONS

If the argument of the last section is sound, the reason Marge has to buy groceries—to perform that action—is not equivalent or reducible to a reason for her to do something that will have the result that she buys groceries. But this leaves us with a further question. Why does Marge's need for groceries not give her a reason to smoke? Why should the fact that her smoking will promote a desirable outcome not count at all in favor of doing it?

Some further explanation is needed, especially since this fact would seem to give someone in her situation good reason to smoke if we were to suppose that she would otherwise be too lazy or forgetful to go to the store. Such self-manipulative activity is not necessarily irrational.[23] Most of us engage in some form of it from time to time. One might deliberately avoid a certain topic of conversation, knowing that it tends to get one so worked up that one behaves badly. Or one might make a bet with a friend that one will stick to one's diet, hoping this will counteract the temptation to overeat. There are often good reasons to take such measures (even if those reasons are sometimes outweighed by other considerations). Why should things look so different in Marge's case?

In response, we can begin by noting that, when Marge smokes her remaining cigarettes, her intention is to introduce a new motive or reason for her to go to the store, beyond her need for groceries. She gets herself to go to the store; but when she does go, she goes in order to buy cigarettes, not groceries. She thus brings herself to go, not for the reason she takes to show that she should go, but for some other reason.

It is plausible, however, that acting *for* the reasons that normatively support

23 See, however, the discussion of this question in Julius, "The Possibility of Exchange" and "Reconstruction."

one's actions is an ideal of rational agency.[24] More precisely, it seems that if some consideration counts decisively in favor of one's performing a particular action then one should perform the action on the basis of that consideration. In choosing to act as she does, Marge flouts this ideal. She brings herself to do what she has decisive reason to do, but she arranges it so that she will not do it on the grounds that rationally support her act in the first place.

Now, someone who acts as Marge does will fail with respect to this standard whether we suppose that she smokes because she believes she will otherwise fail to get herself to the store, or whether we stipulate that she has no such concern. Still, there is a difference between two ways of imagining the case. If we assume, on the one hand, that someone (let us call her Midge) is attempting to ensure that she does not fail to do what she thinks she should do, then we are not prevented from attributing to her some awareness of the fact that, ideally, she would go to the store for the reason that establishes that this is something she should do. We can attribute this recognition to her since, we may suppose, were Midge to regard herself as fully capable of going for this reason—were she not expecting to fail in exactly this respect—she would see no reason to smoke as a way of bringing herself to go. It is not that she fails to grasp the ideal, it is just that she regards it as temporarily out of her reach.

By contrast, Marge, as we were imagining her, appears simply to be indifferent as to whether or not she goes to the store for the "right" reasons. She sees no obstacle to acting directly on the reason that counts (decisively) in favor of going. That is (we have stipulated), she regards herself as perfectly capable of going to the store in order to buy groceries. And yet she chooses the less direct route. This implies that, although the feature of her act that she sees as making it worth performing is that it will enable her to purchase groceries, she fails to see this as a consideration *on the basis of which* she should go. In this way, it seems she fails to grasp fully the import of the relevant consideration as one that provides her with a reason to act.

I want to suggest that this helps explain the intuitive difference between Marge's unnecessary self-manipulation and Midge's attempt to deal with her own faulty agency. When it comes to Midge, it makes sense to assume that, in taking some consideration to provide her with a decisive reason to act, she understands it as a reason that, ideally, she would act on. My conjecture is that this understanding is partly constitutive of the judgment or recognition that some-

24 Compare the discussion in Markovits, "Acting for the Right Reasons," concerning the connection between morally worthy action and acting on the basis of the specific reasons that count in favor of one's action. For a very interesting argument linking something like this ideal with a certain form of freedom, see Julius, "Reconstruction."

thing counts as a decisive reason for an action. This would explain why Marge's thought and decision make so little sense to us. She appears to take a certain reason to show that she should act in a particular way, while at the same time she does not take it to be a reason that, if possible, she should act on. But this is an incoherent stance to take toward the relevant consideration. It is as though she both does and does not regard her need for groceries as a decisive reason for her to go to the store.

Because Marge knows that she can go directly to the store in order to buy her groceries, it would be irrational for her to regard her need for groceries as equally providing her with a reason to smoke, and thereby bring herself to go to the store, since that would have the effect of her going, not in order to buy groceries, but in order to get cigarettes. My claim is that the irrationality here is best explained by the fact that the latter is ruled out by the concept of a (decisive) reason for action. If that is right then it would seem to follow that the productive link between smoking and the purchase of groceries provides no reason at all for Marge to smoke.

6. CONCLUSION

I argued in section 3 that the production conception cannot make sense of the idea that, in some circumstances, what practical reason requires or recommends is not just an act's performance, but an act's being motivated in certain ways. The problem was, roughly, that it makes no sense to regard such a requirement or recommendation in terms of some outcome one has reason to bring about.

I have just argued that, quite generally, judging that some consideration provides one with a reason for action commits one to the proposition that what it requires or recommends is action that is motivated in a certain way. There thus appears to be a deep difference between rational action, on the one hand, and, on the other, the mere production of events or states of affairs one has reason to want. We cannot fully understand the former in terms of the latter. As we have seen, an agent who regards her reasons for action *merely* as reasons to produce a state of affairs in which she does what the reasons indicate she should do has failed to fully grasp the sense in which she has decisive reason to act. We cannot really understand our reasons for action so long as we think of our actions merely as means of production.

REFERENCES

Aristotle. *Nicomachean Ethics*. Translated by Terence Irwin, Indianapolis, IN: Hackett Publishing, 1999.

Bratman, Michael. *Intentions, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press, 1987.

Broome, John. *Rationality through Reasoning*, Oxford: Wiley-Blackwell, 2013.

———. *Weighing Goods*. Cambridge, MA: Blackwell, 1991.

Brown, Campbell. "Consequentialize This." *Ethics* 121, no. 4 (July 2011): 749–71.

Brunero, John. "The Scope of Rational Requirements." *Philosophical Quarterly* 60, no. 238 (January 2010): 28–49.

Dreier, James. "Structures of Normative Theories." *Monist* 76, no. 1 (January 1993): 22–40.

Hieronymi, Pamela. "Controlling Attitudes." *Pacific Philosophical Quarterly* 87, no. 1 (March 2006): 45–74.

Hurka, Thomas. "Value and Friendship: A More Subtle View." *Utilitas* 18, no. 3 (September 2006): 232–42.

Julius, A.J. "The Possibility of Exchange." *Politics, Philosophy and Economics* 12, no. 4 (November 2013): 361–74.

———. *Reconstruction*. Princeton, NJ: Princeton University Press, forthcoming.

Kolodny, Niko. "Why Be Rational?" *Mind* 114, no. 455 (July 2005): 509–63

Korsgaard, Christine M. "The Myth of Egoism." In *The Constitution of Agency*, 69–99. Oxford: Oxford University Press, 2008.

———. "Skepticism about Practical Reason." In *Creating the Kingdom of Ends*, 311–34. Cambridge, MA: Harvard University Press, 1996.

Markovits, Julia. "Acting for the Right Reasons." *Philosophical Review* 119, no. 2 (April 2010): 201–42.

Nagel, Thomas. *The Possibility of Altruism*. Princeton, NJ: Princeton University Press, 1970.

Parfit, Derek. "Rationality and Reasons." In *Exploring Practical Philosophy: From Action to Values*, edited by Dan Egonsson, Jonas Josefsson, Björn Petersson, and Toni Rønnow-Rasmussen, 17–40. Farnham, UK: Ashgate Publishing, 2001.

Portmore, Douglas W. *Commonsense Consequentialism: Wherein Morality Meets Rationality*. New York: Oxford University Press, 2011.

———. "Consequentializing Moral Theories." *Pacific Philosophical Quarterly* 88, no. 1 (March 2007): 39–73.

———. "The Teleological Conception of Practical Reasons." *Mind* 120, no. 477 (July 2011): 117–53.

Rawls, John. *A Theory of Justice*. Cambridge MA: Harvard University Press, 1971.
Raz, Joseph. *Engaging Reason: On the Theory of Value and Action*. Oxford: Oxford University Press, 2000.
———. *From Normativity to Responsibility*. Oxford: Oxford University Press, 2011.
———. *Practical Reason and Norms*. Oxford: Oxford University Press, 1999.
Scanlon, Thomas M. "Structural Irrationality." In *Common Minds: Themes from the Philosophy of Philip Pettit*, edited by Geoffrey Brennan, Robert Goodin, Frank Jackson, and Michael Smith, 84–103. Oxford: Oxford University Press, 2007.
———. *What We Owe to Each Other*. Cambridge, MA: Harvard University Press, 1998.
Scheffler, Samuel. *The Rejection of Consequentialism*. Oxford: Oxford University Press, 1994.
Schroeder, Mark. "Teleology, Agent-Relative Value, and 'Good.'" *Ethics* 117, no. 2 (January 2007): 265–95.
Searle, John. *Rationality in Action*. Cambridge, MA: MIT Press, 2001.
Setiya, Kieran. *Reasons without Rationalism*. Princeton: Princeton University Press, 2007.
Shah, Nishi. "A New Argument for Evidentialism." *Philosophical Quarterly* 56, no. 225 (October 2006): 481–98.
Velleman, J. David. "Deciding How to Decide." In *The Possibility of Practical Reason*. Oxford: Oxford University Press, 2000.
Wallace, R. Jay. "The Publicity of Reasons." *Philosophical Perspectives* 23 (2009): 471–97.
Williams, Bernard. "Internal and External Reasons." In *Moral Luck*, 101–13. Cambridge: Cambridge University Press, 1981.