Aporia

Undergraduate Journal of the St Andrews Philosophy Society

Volume XIX — Issue No. 1

Aporia *is funded by the University of St Andrews Philosophy Society, which receives funds from the University of St Andrews Department of Philosophy, the University of St Andrews Students' Association, and independent benefactors.*

Contents

Acknowledgements

We would like to thank everyone who has contributed in any way towards the publishing and support of *Aporia*. In particular, a special thanks goes out to Andrew Fish, Andrew Ma, and Micheal Taylor for their significant efforts as a part of the Aporia Sub-Committee. Additionally, we'd like to thank Joe Waters, Ryan Miller, Quentin Pharr, Andrés Hernández Villarreal, and Stephano Lo Re for their help in the reviewing process. Furthermore, we'd like to thank Kimon-Sourlas-Kotzamanis for being responsible for the journal design and style, and Kyle Brady from the Online Journal Hosting Service at the University of St Andrews for helping set up our online interface. Finally, thank you to everyone who submitted a paper for our consideration.

Spencer Albert
Editor-in-chief

Arvo Muñoz Morán and Kyle Van Oosterum
Deputy editors

Where's the Harm in It?

Distinguishing epistemic and moral harm in cases of epistemic injustice

Kirsty Hardwick*
University of St Andrews

Abstract At the heart of the epistemic injustice debate is Fricker's claim that an agent can be harmed purely in their capacity as a knower. For Fricker, this harm occurs in cases of epistemic injustice, where an individual's testimony is undervalued due to the prejudice of their audience. In this paper, I consider Fricker's claim that these cases involve a 'distinctly epistemic kind of injustice.' I argue that Fricker relies too heavily on her virtue epistemological commitments which leads her to conflate moral and epistemic concerns in cases of epistemic injustice. Concluding that we therefore we need to be more precise about what it means for an agent to be harmed as a knower, in the second part of the paper I sketch a theory-neutral distinction between epistemic harms, where knowers are restricted from access to a knowledge exchange, and moral harms, where moral agents are negatively affected by morally impermissible actions. I suggest that this distinction can enable us to be clearer about the harm done in cases of epistemic injustice and help us identify who is responsible. The paper ends with suggestions for further research.

1 Introduction

Placed at the intersection of ethics and epistemology, the epistemic injustice debate examines where our roles as knowers and moral agents coincide. Crucial to this debate is the claim, proposed by Miranda Fricker in her seminal book *Epistemic Injustice,*[1] that

*Kirsty Hardwick is a recent philosophy graduate from the University of St Andrews. She focuses on philosophical topics which have a political bite, including social epistemology and the philosophy of sex. Putting off the real world for another year, she is about to begin an MPhil in Philosophy at University of Cambridge.

[1] Miranda Fricker, *Epistemic Injustice: Power and the Ethics of Knowing* (Oxford University Press, 2007), 1.

an individual can be harmed purely in their capacity as a knower. In Fricker's central cases this harm occurs when epistemic agents are not trusted to be competent testifiers due to the prejudice of their audiences.[2]

In this paper, I consider Fricker's claim and argue that we need to be more precise about what it would mean for a harm to be epistemic. I sketch a theory-neutral distinction between epistemic harms, where a knower has restricted access to a knowledge exchange, and moral harms, where moral agents are negatively affected by morally impermissible actions. I suggest that this distinction can enable us to be clearer about the harm done in cases of epistemic injustice and help us identify who is responsible.

In §2, I set out Fricker's account of epistemic injustice and discuss two central cases to illustrate the phenomenon. In §3 I consider Fricker's own characterisation of the harm done in cases of epistemic injustice and argue that her commitment to virtue epistemology causes her to conflate moral and epistemological concerns. In §4 I sketch a theory-neutral distinction between epistemic harms and moral harms and apply it to the cases introduced in §2 and in §5 I respond to possible objections to my distinction. Finally, I consider the challenge of extending the account to cases of structural injustice and discuss avenues for future research.

2 Defining epistemic injustice

2.1 Two concepts of epistemic injustice

David Coady distinguishes between two attempts to identify a type of injustice that is purely epistemic.[3] The first characterises epistemic injustice as an unjust distribution of epistemic goods (e.g. knowledge and education).[4] While certainly a prevalent social injustice, Fricker argues that this understanding of 'epistemic injustice' is not properly epistemic because the fact that the good in question is an epistemic good is, according to Fricker, 'incidental.'[5]

Coady disagrees, arguing that since the epistemic goods in question are intrinsically epistemically valuable, it is not incidental that the distribution of these goods is an epistemic issue.[6] However, this misses the point. It is the injustice—the harm caused—not the epistemic goods, which is not truly epistemic. Once we have identified what epistemic goods are, the question of their fair distribution, as Coady himself

[2] Discussion of Fricker's second type of epistemic injustice, hermeneutical injustice, on which an individual is disadvantaged because of a gap in her society's understanding of a concept needed to understand her experience, is beyond the scope of this essay.

[3] David Coady, 'Two Concepts of Epistemic Injustice', *Episteme* 7, no. 2 (2010): 101.

[4] Ibid.

[5] Fricker, *Epistemic Injustice*, 1.

[6] Coady, 'Two Concepts', 106.

notes,[7] collapses into the same discussion of distributive justice that occurs for other goods/property. Therefore, it is not a distinctly epistemic type of injustice.

The second conception of epistemic injustice, Fricker's own, aims to be more fundamentally epistemic.[8] The injustice lies in an individual not being acknowledged as a knower, resulting in their testimony being undervalued.[9] Crucially, note the harm done is 'not to be characterised as a non-receipt of one's fair share of a good (credibility).'[10] Instead, Fricker argues the harm is the distinctly epistemic harm of being undermined in one's capacity as a knower.

Before I examine whether epistemic injustice does involve distinctly epistemic harms, in §2.2, I introduce some terminology from Fricker to further elucidate her concept of epistemic injustice. In §2.3, I introduce two central cases which illustrate Fricker's characterisation of epistemic injustice.

2.2 Fricker on epistemic injustice

To understand Fricker's characterisation of epistemic injustice, we must briefly consider her views on testimony, the everyday epistemic practice of conveying knowledge to others.[11] When we listen to testimony, we face the decision of how much credibility to attribute to the speaker. Given the little information we have on which to judge the speaker's credibility, Fricker suggests that we often use stereotypes as heuristics to facilitate making credibility judgements.[12] These stereotypes may be useful, prejudicing us to trust the testimony of teachers over our peers for instance. Yet they may also introduce prejudice into our credibility judgements, causing testimonial injustice whereby a speaker's testimony is undervalued by a hearer on the basis of a facet of their identity such as their gender or race.[13]

Fricker identifies two ways in which testimony can be dysfunctional as a result of prejudice.[14] First, in cases of *credibility excess*, a speaker receives more credibility that she otherwise would have due to the prejudice of a speaker.[15] For example, we attribute an excess of credibility to doctors about medical matters because we are prejudiced to think that doctors know about medicine. Second, in cases of *credibility deficit*, a speaker receives less credibility than she otherwise would have due to the prejudice of

[7] Ibid., 103.

[8] Ibid., 101.

[9] Fricker, *Epistemic Injustice*, 1.

[10] Ibid., 20.

[11] Ibid., 16.

[12] Ibid., 30–33.

[13] Ibid., 1.

[14] Note that prejudice may be positive or negative. To be prejudiced to think *x* is, roughly, to be resistant to thinking not-*x* (ibid., 35).

[15] Ibid., 17.

the speaker.[16] For example, the boy-who-cried-wolf receives a credibility deficit from his audience since they are prejudiced to think he is lying.

Fricker further distinguishes between *systematic* and *incidental* cases of injustice. In incidental cases, the prejudice is highly localised to the situation. For instance, suppose a student encounters a teacher with a strong prejudice against people who write in fonts besides Times New Roman and undervalues the essay she submits in Calibri. In this case, the prejudice in question is not commonly-held so will not generalise, causing further injustices. In systematic cases, the prejudice which causes the credibility deficit is pervasive and connects the epistemic injustice to other types of injustices.[17] For example, testimonial injustices based on race or gender are examples of systematic injustice.

2.3 Central cases

Fricker's central cases of epistemic injustice are cases of systematic *identity-prejudicial credibility deficit.* By *identity-prejudicial* Fricker means that the prejudice that causes the credibility deficit is a *negative identity prejudice,* which we can understand simply as a commonly-held disparaging association between a social group and one or more attributes.[18] These cases are central for Fricker because they connect to other forms of social injustice that a subject might encounter, hence locating epistemic injustice within the 'broader pattern of social injustice.'[19]

> **Central case 1** For our first central case, let us follow Fricker in using a story from Harper Lee's *To Kill a Mockingbird.* Tom Robinson, a black man in 1930s Alabama, is on trial for beating and raping a white young woman. His lawyer has proved beyond reasonable doubt that Tom cannot be responsible for the beating yet despite this, the jury still refuse to believe a black man's testimony over the testimony of a white woman. Here, we have a clear-cut case of identity-prejudicial credibility deficit. The jury exhibits a negative identity prejudice against Tom because they associate lying with being black. This leads them to assume that Tom will not give testify the facts of the case. Therefore, Tom is attributed a strong credibility deficit due to the jury's racial negative identity prejudice and they cease to view him as a knower, that is someone who is capable of knowing the facts and transmitting them. Furthermore, since this prejudice is prevalent and the credibility deficit is one of many injustices the black community suffered, the injustice is systematic.[20]

[16] Fricker, *Epistemic Injustice.*

[17] Ibid., 27.

[18] For a more precise definition, see ibid., 35

[19] Ibid., 4.

[20] Ibid., 23–28.

A closely related term to testimonial injustice is silencing. I follow Kristie Dotson in understanding silencing as the conjunction of two distinct epistemic injustices. *Testimonial quieting* occurs when 'an audience fails to identify a speaker as a knower,'[21] e.g. in the Tom Robinson case as the jury does not identify Tom as a knower of the facts of the case. The second kind of epistemic injustice is *testimonial smothering,* sometimes referred to as *self-silencing,* which involves 'the truncating of one's own testimony.'[22] In testimonial smothering, the knowledge of the prevalence of identity-prejudicial credibility deficits causes an individual to self-censure their speech, resulting in, to use Fricker's term, *pre-emptive* testimonial injustice.[23] This leads us to our second central case:

> **Central case 2** GLAAD defines Bisexual Erasure as 'a pervasive problem in which the existence or legitimacy of bisexuality (either in general or in regard to an individual) is questioned or denied outright.'[24] This phenomenon can lead bisexuals to fear coming out since they think that they will not be believed by someone who denies the existence of bisexuality. The pervasiveness of this problem leads some to self-silence if they feel unable to come out to those around them because they do not think that they will be taken seriously as a knower of the facts about their own sexuality.[25] As our second central case, let us take the example of a bisexual man in a relationship with a woman who refrains from coming out to his friends and family out of fear of not being believed or being misidentified as homosexual.

Following Fricker I take both types of cases to be important examples of injustice and I agree that we should spend time examining them to highlight the prevalent social injustices they exemplify. However, in the following section I argue that we need a clearer account of the harm done in these injustices in order to call what is occurring a distinctly epistemic kind of injustice. Such an account should be able to make good on Fricker's claim that we can hurt in our capacity as a knower by distinguishing between epistemic and moral harms.

[21] Kristie Dotson, 'Tracking Epistemic Violence, Tracking Practices of Silencing', *Hypatia* 26, no. 2 (2011): 242.

[22] Ibid., 244.

[23] Fricker, *Epistemic Injustice,* 130.

[24] GLAAD, 'Erasure of Bisexuality', accessed 17 February 2019, `https://www.glaad.org/bisexual/bierasure`.

[25] Note that I do not claim that all individuals who choose not to come out are self-silencing as there are many legitimate reasons why someone may not come out which has nothing to do with fear of encountering prejudice. However, when an individual would like to come out but feels unable to on account of their knowledge of prejudicial attitudes it provides a case of self-silencing.

3 Fricker's account of the harm in epistemic injustice

We saw in §2.1 that Fricker's aim in *Epistemic Injustice* was to identify a 'distinctively epistemic kind of injustice.'[26] Yet, consider the following passage, in which Fricker explains the importance of understanding the wrong done in epistemic injustices:

> Any claim of injustice must rely on shared ethical intuition, but we achieve a clearer idea of why something constitutes an injustice if we can analyse the nature of the wrong inflicted.[27]

I think Fricker is right to note that there is a moral judgement involved in calling something an injustice, and that moreover injustice is commonly identified by means of ethical intuition or reasoning. However, if epistemic injustice is still to be properly *epistemic*, then there must be a sense in which victims of epistemic injustice are harmed epistemically, as well as morally.

While we might have expected Fricker to discuss the ways in which epistemic injustice leads to less overall transmission of knowledge, the harm Fricker actually identifies is less about knowledge and more about being recognised as a knower. In Chapter 6, Fricker clarifies her account of the harm in epistemic injustice, characterising it as a kind of *epistemic objectification*, on which individuals are treated as 'sources of information', not 'informants', or knowers.[28] Crucially, Fricker claims that to treat someone as a source of information or as an informant is to have a particular ethical attitude towards them,[29] not an epistemological one. She even adopts part of the Kantian framework to show that it is immoral to treat someone as a *mere* source of information, just as it is immoral to treat someone as a *mere* means to an end rather than an end in themselves.[30] The harm caused is moral, caused by a morally-impermissible ethical attitude, rather than epistemic as Fricker originally promised.

Asking why Fricker identifies the harm done in cases of epistemic injustice in this way reveals her virtue epistemological commitments. For a virtue epistemologist, the lines between epistemic value and moral value are already blurred because they view our role as knowers as an extension of our role as moral agents (that is, how we conduct ourselves as knowers affects should be guided by whether it would lead a flourishing life or not).

Fricker's own virtue epistemological leanings are clear from the fact that in addition to identifying the phenomenon of epistemic injustice, Fricker develops an account of the virtue of epistemic justice, which also informs her suggestions for reducing epi-

[26] Fricker, *Epistemic Injustice*, 1.

[27] Ibid., 5.

[28] Ibid., 134.

[29] Ibid., 131.

[30] Ibid., 133.

stemic injustices.[31] While a full analysis of this virtue is beyond the scope of the essay, it is worth noting that Fricker aims the virtue to be 'hybrid in kind: both intellectual and ethical'[32] just like epistemic injustice is meant to be both epistemic and ethical. Yet the epistemological concerns, at least in the case of epistemic injustice, appear to be considered less than the ethical concerns.

Indeed, the injustice Fricker describes could more accurately be called an identity injustice, since it rests on the moral wrong of undermining the dignity of someone's identity as a knower. Consider the case of a sexist employee who fails to consider their female boss a superior because they hold a negative identity prejudice that states that women do not have leadership skills. It seems to me that the harm identified by Fricker in cases of epistemic injustice also occurs in this case. This shows that the fact that the individual was undervalued as a *knower* is incidental, just as the fact that the goods were *epistemic goods* was incidental to the kind of injustice involved in Coady's first type of epistemic injustice. This is the case so long as we identify the wrong involved in epistemic injustice as holding the wrong kind of ethical attitude towards someone on the basis of a negative identity prejudice.

In sum: while there is no doubt that there is an important ethical dimension to epistemic injustice, Fricker has failed to identify the distinctly epistemic aspect of epistemic injustice which was meant to identify it as a sui generis kind of injustice. In the remainder of this paper, I aim to rescue Fricker from this criticism by providing a way of identifying epistemic injustice which retains the distinctly epistemic element of the injustice. I do this by sketching a distinction between being harmed as a knower (an epistemic harm) and being harmed as a moral agent (a moral harm). On this account, the epistemic harm involved in epistemic injustice is not that someone is undervalued as a knower but that a knowledge exchange has broken down, resulting in an obstacle to gaining knowledge. Further, epistemic injustice occurs when this is due to a morally culpable prejudicial credibility deficit against a knower.

4 Two kinds of harm

In distinguishing between moral and epistemic harms, the goal is to provide a theory-neutral account of the harms involved in cases of epistemic injustice in order to highlight that it is an epistemic phenomenon. After sketching my distinction in §4.1-4.2, in this section I show how the distinction accounts for the harms in the central cases of epistemic injustice and motivate the use of my distinction by arguing that it explains the intuitive harm that occurs in cases of credibility excess which Fricker dismisses.

[31] Ibid., chap. 4.

[32] Ibid., 6.

4.1 Moral harm

I define a moral harm as *a bad effect on a moral agent resulting from a morally impermissible action by a moral agent*.

There are two important features of this definition. First, it is not complete since we need to supply further definitions of 'bad effect' and 'morally impermissible'. How the definition is fleshed out will therefore depend on the moral theory that one espouses; a utilitarian might characterise 'bad effect' as 'non-optimum level of wellbeing' while the virtue ethicist could define it as 'diminished flourishing.' The benefit of using a loose definition is that it will show that moral harms are distinct from epistemic harms not just on one particular moral theory but in a broader sense.

Second, we should understand 'moral agent' as a member of the moral community, i.e. as someone who is morally responsible for their actions. Identifying the moral harms of a situation further identifies where the blame should be placed, i.e. on the moral agent who acted impermissibly. For example, acting on a morally culpable prejudice (e.g. negative identity prejudices) provides the moral harm in most epistemic injustices.[33]

4.2 Epistemic harm

I define an epistemic harm as *a restriction on access to a knowledge exchange*.

The first thing to note is that this definition is very broad. A young child who asks how babies are made is harmed epistemically when their parents do not give them the full answer, as is the student who cannot afford a particular textbook, since both are blocked from participation in an exchange of knowledge.

Secondly, note that not all epistemic harms are morally culpable. The former example of the parents fudging the truth a little provides an example of a morally innocuous epistemic harm. Since we said that epistemic injustices involve both moral and epistemic harms, note that not every case of epistemic harm will count as epistemic injustice.

Instead, I suggest Fricker is right that for epistemic injustice to be an injustice, there must be a credibility deficit caused by identity-prejudice, although this should be understood as a moral, not an epistemic harm. A case counts as epistemic injustice iff it includes:

(1) Moral and epistemic harms;

(2) The moral harm of being undermined as a knower due to a morally-culpable

[33] I consider exceptions to this rule in §4.3 and §7.

prejudice held by one's audience.

4.3 Central cases revisited

First, in the Tom Robinson case, which illustrates testimonial quietening, it is clear that on all plausible moral theories there are moral harms involved. Tom is harmed morally because he is discounted as a knower due to the morally impermissible prejudice of the jury, resulting in an unjust verdict of guilt. He is harmed epistemically because he cannot transfer his knowledge of the situation to the jury since they cannot believe his testimony. Hence, he cannot participate in a knowledge exchange and is impeded in his ability to be a knower.

An implication of my definition is that the jury are also harmed epistemically. Their inability to believe Tom's testimony blocks their receipt of the information they need to come to a true belief about what happened. This obstruction of this knowledge exchange counts as an epistemic harm to the jury as well as to Tom. Hence, victims and perpetrators are both harmed epistemically in cases of epistemic injustice.

This is an important result for two reasons. First, it captures the idea that prejudice is harmful to society since it silences whole social groups. We lose out on the knowledge that could be gained from their unique testimony and perspectives. Second, it suggests even privileged members of society should be motivated to combat epistemic injustice, since they too are epistemically harmed by injustice.

In cases of testimonial smothering, the moral harm can be harder to identify. In the case of the bisexual man, he might not come out to friends who would in fact support him since his awareness of biphobia broadly leads him to fear a negative response from his friends. We might not want to say that his friends have harmed him morally, nor that there is a pre-emptive or counterfactual harm, since his friends would have supported him. Instead, it looks like his society has harmed him morally, rather than an individual moral agent. When the prejudice involved in epistemic injustice is part of the structure of society,[34] the moral harms involved must be different to when injustice is the result of an individual's prejudice. I return to this worry later.

By contrast, the epistemic harm involved is clear. Just as a shy student is harmed epistemically by not engaging in discussion in class, in a similar way self-silencing is epistemically harmful to listeners and speakers since a perspective is lost to the discussion.[35]

[34] Fricker, *Epistemic Injustice,* 10–11.

[35] Note that epistemic blame must not relate to epistemic harm in an analogous way to moral blame since then the self-silencer would be epistemically responsible for the harm caused. However, I leave aside issues of developing an account of epistemic blame.

4.4 Credibility excess revisited

An important feature of Fricker's account of epistemic injustice is that it includes cases of credibility deficit, but excludes (most cases of) credibility excess.[36] Fricker argues that cases of credibility excess do not involve the withholding of 'a proper respect for the speaker *qua* subject of knowledge.'[37] However, I agree with Medina that 'Fricker's claim that a credibility excess does not handicap the speaker in the course of the exchange in the same way that a credibility deficit does is dubious.'[38]

By using the distinction between epistemic and moral harms, we can explain the intuitive wrong involved in certain cases of credibility excess without having to further characterise them as epistemic injustices. Consider Fricker's case of the professor who asks a junior colleague to give her comments on a paper she is presenting at a conference.[39] The junior colleague admires the professor and gives too much benefit of the doubt, resulting in his comments being less critical than usual. The professor is harmed epistemically since the junior colleague does not give their best comments on the paper and this restricts the professor's access to the colleague's knowledge. Yet, while this is as a result of prejudice—the junior colleague is prejudiced towards thinking the professor has good suggestions (i.e. resistant to thinking otherwise)—it is plausibly not a morally impermissible prejudice to hold. Hence, the professor is not harmed morally by the encounter and it does not count as an epistemic injustice.

5 Objections

Having introduced my distinction between epistemic and moral harms and provide some reason for thinking that it is useful, in this section I consider two possible objections to my account.

5.1 The epistemic harm harmful?

In §4, I aimed to provide a theory-neutral account of moral and epistemic harms. While my definition of moral harm is neutral with regards to which moral theory is correct, it may be objected that my definition of epistemic harm commits me to a particular conception of epistemic value. The challenge my account faces is to answer how an epistemic harm can be a kind of harm even using loose definitions of knowledge and epistemic value.

[36] Fricker, *Epistemic Injustice*, 21.

[37] Ibid.

[38] José Medina, 'The Relevance of Credibility Excess in a Proportional View of Epistemic Injustice: Differential Epistemic Authority and the Social Imaginary', *Social Epistemology* 25, no. 1 (2011): 17.

[39] Fricker, *Epistemic Injustice*, 18.

The easiest way to see the harm in restricted access to a knowledge exchange is to say that we lose out on the knowledge that we would have gained from the knowledge exchange. By knowledge, I here mean a weak sense of knowledge used by Goldman, 'true belief'[40] and as with the moral harm definition, I leave the fleshing out of the concept to an individual's preferred theory of knowledge. Yet we are left with a regress of the question—we must now ask why it is harmful to miss out on knowledge, particularly when it is understood as mere true belief.

The obvious response, versions of which are endorsed by Coady[41] and Goldman is that knowledge has intrinsic value (if one that can be trumped by other values).[42] This view explains the wrong of cases of epistemic harm because being blocked from a knowledge exchange then restricts one's access to something which is intrinsically valuable. However, we can question whether all true beliefs are indeed intrinsically valuable: as Coady quips, 'the project of maximising true beliefs seems at best to be valuable for those who want to do well in the game of *Trivial Pursui*.'[43]

Consider as an example my true belief that 'Meghan Markle's baby is due in April 2019'. Prima facie, it does not appear to be intrinsically valuable to hold this belief, and hence not all true beliefs are intrinsically valuable. Thus, to preserve the sense in which we are harmed by missing out on knowledge we must amend the weak definition of knowledge beyond 'true belief'. I briefly consider two such attempts.

First, Greco argues that knowledge should be understood as true belief arrived at by a method that deserves credit.[44] The value of knowledge lies in arriving at the truth by a reliable method rather than accidentally. However, consider an investigative journalist who puts in the effort to reach this true belief, perhaps even checking with Kate Middleton's obstetrician. The burden of proof is still on Greco to explain why using reliable methods to arrive at trivial truths should be intrinsically valuable, particularly when such methods and effort could have been utilised to arrive at more important truths.

Second, Coady, following Goldman, restricts intrinsic value to true beliefs that answer:

> First, questions the agent happens to find interesting, second, questions the agent would find interesting if he or she had thought of them, and third, questions that the agent has an interest in having answered.[45]

Adopting this view implies knowers are only harmed epistemically when they lose

[40] Alvin I. Goldman, *Knowledge in a Social World* (Oxford University Press, 1999), 5.

[41] Coady, 'Two Concepts', 106.

[42] Goldman, *Knowledge in a Social World*, 6.

[43] Coady, 'Two Concepts', 103.

[44] John Greco, 'Knowledge as Credit for True Belief', in *Intellectual Virtue: Perspectives From Ethics and Epistemology*, ed. Michael DePaul and Linda Zagzebski (Clarendon Press, 2003), 116.

[45] Coady, 'Two Concepts', 103.

out on *interesting* true beliefs. Since such beliefs are taken to be intrinsically valuable, we can explain why cases of epistemic harm are harmful.

The problem is that the inclusion of an agent's interests makes the value more instrumental than intrinsic. It is plausible that knowing 'Meghan Markle's baby is due in April 2019' is instrumentally valuable to an avid royalist, aiding their goal to know trivia about royalty. However, I see no reason why being an avid royalist makes it an intrinsically valuable true belief to hold.

Yet, even if knowledge is only instrumentally valuable, I argue that we can still see the harm in epistemic harm. To the extent that our goals/interests are prudentially valuable to us, knowledge which furthers these goals/interests in valuable. Hence, a breakdown in an exchange of (interesting) knowledge does result in the loss of something valuable, and thereby we are harmed.

5.2 A distinct kind of harm?

Having established that knowledge need only be instrumentally valuable for epistemic harm to be harmful, a second possibly objection is that making this move causes the distinction to collapse between epistemic and moral harms. The worry here is that if our goals/interests are harmed, then we are harmed morally every time we are harmed epistemically.

The answer to this concern is to point out that our goals are not always moral. To illustrate, suppose you walk down the street and encounter an obviously shady character wearing a balaclava and holding an empty bag labelled '$ $ $'. If they ask you where the nearest bank is and you purposely deceive them, the would-be criminal is harmed epistemically since they lose out on interesting knowledge (i.e. knowledge that furthers their interest in robbing the bank). Yet, they are not harmed morally since (plausibly) you have not acted morally impermissibly by obstructing their own immoral act. The distinction survives since there are cases which involve epistemic, but not moral, harms.

6 Concluding remarks

In this paper, I aimed to do three things. First, I argued that Fricker's account of the harm done in cases of epistemic injustice misses the *epistemic* harm caused and fails to establish why epistemic injustice is 'distinctly epistemic.' Second, I sketched definitions of epistemic and moral harms and used these to analyse key cases of epistemic injustice and to identify the intuitive harm in some instances of credibility excess. Third, I responded to possible objections to my distinction and argued that we can separate the moral and epistemic elements of epistemic injustice.

One remaining worry with my argument is that it is difficult to attribute moral blame in epistemic injustice cases such as testimonial smothering. The crux of this issue is that the prejudice involved is not tied to an individual, but rather part of the fabric of the society we live in, i.e. it is *structural.* For example, in our second central case the mere awareness that pervasive biphobia and bisexual erasure exists can cause testimonial smothering, independent of any individual's biphobic belief. This threatens my suggestion that identifying the moral harm in a case of epistemic injustice further identifies who is to blame.

Even where the prejudice can be attributed to an individual, the prejudice is often an implicit bias the individual may be unaware they have. It is tempting to say that we cannot be blameworthy for such implicit biases. Yet, while this may be comforting, it risks demotivating our attempts to resist our implicit biases by assuming we have no control over our biases. Instead, there is empirical evidence that suggests we can improve the situation through reflective self-regulation.[46] Given that such self-regulation will be a long and effortful process, we need a motivation to even try. Taking responsibility for our implicit biases by seeing the moral harms that they cause provides an essential first step towards motivating the effort involved in such self-regulation.[47]

Of course, Fricker is right to argue that combating epistemic injustice requires both individual reflective self-regulation and enacting changes in structural mechanisms.[48] These are important topics and provide fruitful areas for further research. However, one promising direction for identifying the moral harms in structural epistemic injustice is to say that we are all to blame. Consequently, we are all responsible for working towards changes in ourselves, and in our society.

References

Coady, David. 'Two Concepts of Epistemic Injustice'. *Episteme* 7, no. 2 (2010): 101–13.

Dotson, Kristie. 'Tracking Epistemic Violence, Tracking Practices of Silencing'. *Hypatia* 26, no. 2 (2011): 236–57.

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

———. 'Replies to Alcoff, Goldberg, and Hookway on Epistemic Injustice'. *Episteme* 7, no. 2 (2010): 164–78.

[46] Margo Monteith, 'Self-regulation of Prejudiced Responses: Implications for Progress in Prejudice-reduction Efforts', *Journal of Personality and Social Psychology* 65, no. 3 (1993): 469–85.

[47] Many thanks to Dr. Elinor Mason for her presentation 'Taking Responsibility for Implicit Biases,' which influenced this discussion.

[48] Miranda Fricker, 'Replies to Alcoff, Goldberg, and Hookway on Epistemic Injustice', *Episteme* 7, no. 2 (2010): 165.

GLAAD. 'Erasure of Bisexuality'. Accessed 17 February 2019. https://www.glaad.org/bisexual/bierasure.

Goldman, Alvin I. *Knowledge in a Social World*. Oxford University Press, 1999.

Greco, John. 'Knowledge as Credit for True Belief'. In *Intellectual Virtue: Perspectives From Ethics and Epistemology*, edited by Michael DePaul and Linda Zagzebski, 111–34. Clarendon Press, 2003.

Mason, Elinor. *Taking Responsibility for Implicit Biases*. Presentation. Society of Women and Minorities in Philosophy, St Andrews, 2017.

Medina, José. 'The Relevance of Credibility Excess in a Proportional View of Epistemic Injustice: Differential Epistemic Authority and the Social Imaginary'. *Social Epistemology* 25, no. 1 (2011): 15–35.

Monteith, Margo. 'Self-regulation of Prejudiced Responses: Implications for Progress in Prejudice-reduction Efforts'. *Journal of Personality and Social Psychology* 65, no. 3 (1993): 469–85.

Ontological Dumpster Diving

A search for a four-dimensionalist account of a person

James Brown-Kinsella*
Princeton University

Abstract Throughout the literature on personal identity, the term 'four-dimensionalism' is poorly understood. Indeed, Mark Johnston deploys the concept of ontological trash to show that there is no feasible four-dimensionalist account of a person as an object entirely within spacetime, but he does not consider how any particular theory of spacetime or four-dimensionalism comes to bear on personhood. In this paper I will explain this line of reasoning, clarify four-dimensionalism, and show that there is a feasible account of personhood available on four-dimensionalism. In the introduction, I explain the concept of ontological trash and its threat to personhood. In the first section, I explain the concept of time dilation and use it, in conjunction with ontological trash, to prove that a person's life does not have an unqualified temporal duration. In the second section, I summarise Cody Gilmore's analysis of four-dimensionalism and explain how it comes to bear on persistence. In the third section, I sketch a new way to escape ontological trash in light of four-dimensionalism. In the fourth section, I apply this response to personhood, arguing that persons exist fully within spacetime and can withstand almost any psychological change. In the conclusion, I reflect on avenues for future research.

1 Persons and personites as ontological trash

Time is fleeting. Perhaps the most salient feature of a person's life is how little time they have to live. Indeed, the recognition of time's seemingly unjust imposition upon life helps motivate Mark Johnston's argument that it is rational to hope that this life,

*James Brown-Kinsella graduated from Princeton University in 2019, majoring in Philosophy and minoring in Humanistic Studies, French Language and Culture, and East Asian Studies. Next year, he will pursue an MPhil in China studies on a fellowship at the Yenching Academy of Peking University. James is interested in studying ancient Chinese thinkers in the way that ancient Greek thinkers have been studied.

confined to spacetime, is not all there is to one's existence.[1] His argument rests on the idea of ontological trash—i.e. a heap of nearly identical objects, all with equal claim to ontological priority. If we are ontological trash, so Johnston argues, then practical reason can be of no use to us. But practical reason *is* useful, so we ought to reject any account of personhood on which we are ontologically trashy, notably any four-dimensionalist account. In this introduction I will summarise Johnston's argument and motivate a closer look at four-dimensionalism. First, I will explain the concept of ontological trash, as well as the ontologically trashy version of a person: a personite. Finally, I will show how personites pose a threat to practical reason, and why this leads us to re-examine what it means for a theory of personhood to be four-dimensionalist.

A thing is ontological trash if in its nearest spatiotemporal vicinity there are many other things that are nearly identical to it, ontologically on par with it, and which all differ from it only in conditions of persistence.[2] Consider a fist as it exists through time. A fist comes into being when you clench your fingers all the way, and it ceases to exist when you unclench your fingers. On this way of looking at a fist, whenever you have a fist, you will also have a strew of other fist-like objects. There will be loads that are composed of the fist plus the fingers during the moments before they were clenched; there will be loads that are composed of the fist plus the fingers during the moments after they were unclenched; and there will be loads that are composed of the fist plus the fingers during the moments both before and after they were clenched. All of these objects are nearly identical to your fist; they are composed of the same stuff as your fist; and they differ from your fist only in their conditions of persistence. When you look at your fist it is therefore impossible to distinguish it from any of the other fist-like objects in its ontological trash heap.[3]

When it comes to persons, the ontological trash heap is piled high with personites. A personite coincides with a person and may share one but not two temporal end-points with a person.[4] It could come into existence and cease to be somewhere within its person's lifetime; it could come into existence at a time later than its person's origin and cease to be when its person ceases to be; or *vice versa*. Because a personite is made up of exactly the same stuff as its person and because it differs from its person only in the sort of changes it can survive, a personite is very person-like. If persons just are sums of instantaneous person-stages over time, then there is one personite for every interval of time within a person's lifetime (i.e. infinitely many). If persons are instead chains of physically or psychologically continuous person-stages, then any parameter of the continuity relation (e.g. the degree of connectedness necessary for continuity or whether the chain is maximal or not) could be tweaked to produce hoards of person-

[1] Mark Johnston, 'Is Hope For Another Life Rational?' (Unpublished, 2017), 4.

[2] Ibid., 7.

[3] Ibid.

[4] Mark Johnston, 'The Personite Problem: Should Practical Reason Be Tabled?', *Noûs* 50, no. 4 (2016): 199.

ites.[5] The worry that Johnston elaborates is that any four-dimensionalist account of a person will take one of these two forms and thus will be unable to separate the person from the legion of personites with equal ontological status.[6]

Ontological trash is mostly nontoxic. For objects like fists ontological trash only gets in the way of our analysis. When you throw a punch the host of fist-like objects that swing with your first do not help you hit any harder. Practically speaking, there will be only one impact and it does not matter that a multitude of objects were responsible for it. However, objects like persons get a special moral status in virtue of being the kind of object that they are. We ought, *ceteris paribus*, to avoid causing persons undue or uncompensated harm; to respect the intentions and future interests of ourselves and other persons; and to act according to many other uncontroversial moral imperatives. We ought to act this way toward persons just in virtue of a certain property or properties of personhood. Anything that possesses the same properties should be respected in the same way. An ontologically trashy theory of personhood thus corrupts our moral framework because personites are too person-like for their interests to be neglected. Taking personites into our moral calculus leads to at least ten destructive consequences.[7] I will highlight one of them. A trip to the gym would torture the personites who exist entirely during the pain of physical exertion and who cannot be compensated after they have ceased to exist. Similarly, learning a language or undertaking any sort of long-term investment that involves short-term pain, frustration, or other harm will oppress some personites without compensation.[8] So it seems impossible to promote our own interests without being morally flagrant. Personites, as ontological trash, inevitably pollute our moral thinking.

We are, in fact, practical and rational creatures. From this strong intuition Johnston invokes 'a kind of pragmatic *a priori*:'[9] that practical reason requires us to be able to make decisions about how to act, for otherwise we would be paralysed in deliberation. More specifically, in order to take any deliberate action, we are required to believe that we can avoid doing bad; that we can achieve some sort of good; and that we can use some form of ethics to guide our behaviour. But as we saw above, personites pollute our moral thinking, so practical reason demands that we reject any theory of personhood that yields personites. Because any four-dimensionalist account of personhood would yield personites, we are practically required not to believe it. Thus we can hold out hope that our existence is more than just our career in spacetime.

But what exactly does four-dimensionalism entail? Based on Johnston's use of the term, an object that exists four-dimensionally is one whose 'whole reality. . . is found within its spatiotemporal envelope.'[10] Cody Gilmore however has dedicated a paper to

[5] Johnston, 'Is Hope...', 7.

[6] Johnston, passim, especially 'The Personite Problem'.

[7] Johnston, 'The Personite Problem', 10.

[8] Ibid., 17–18.

[9] Johnston, 'Is Hope...', 7.

[10] Ibid., 6.

disambiguating all the theories that go by this name, and spelling out which are supported by the relativity theory of spacetime.[11] With such confusion, one can hold out hope that there is still an account of persistence that is four-dimensionalist in the sense that it explains the reality of an object while keeping it fully in spacetime and without throwing it in ontological trash heaps. In the following sections, I will first explore the concept of persistence through time by highlighting how the relativity theory of spacetime reveals more ontological trash, then I will disambiguate four-dimensionalist theories of persistence and sketch a four-dimensionalist strategy for avoiding ontological trash altogether. After investigating persistence, I will return to personal identity to propose a definition of a person that keeps us fully in spacetime.

2 Relativity and a problem for persistence

Before diving into ontological trash, it is important to remember that physics should always bear on metaphysics. Since the current trend in physics is that we are living in a relativistic spacetime, any account of persistence through time should be expressed in a relativistic account of time. However, the relativity theory of spacetime is a very complicated field of study in itself; I can only hope to scratch the surface here. Nevertheless, in this section I will first lay out the basics of the relativity theory of spacetime, and then offer a new, relativity-based variant of ontological trash that reveals hoards of objects lurking on any account of persistence wholly in spacetime.

On the relativistic view of spacetime, neither time nor space are held as absolute constants. Instead, it is the speed of light that is invariable. Light, when measured by any observer, will move through a certain medium with a certain speed in meters per second. Accordingly, space and time give way in order to accommodate this fact. Imagine you have a very odd clock that consists of a laser gun, a thin pole of a known length, a mirror at the end of that pole, and a receptor attached to the laser. You press a button and the laser fires a beam of light, which travels the known distance to the mirror, where it bounces back and travels to the receptor to be absorbed. Since the speed of light is constant, since you know the length of the pole, and since speed is just distance over time, you can infer the amount of time, in seconds, that passed during the laser beam's journey. Whenever you fire the laser, you can trust that it will accurately measure how time passes for you.

Now suppose you have a friend who measures your clock while you move very quickly. Perhaps you are standing on the caboose of a train moving at a known speed along a set of straight tracks and your friend watches from the train station with binoculars as you fire the laser at a right angle to the train's path. From your friend's perspective, the distance of the laser beam's path is slightly longer than it is from your

[11] Cody Gilmore, Damiano Costa and Claudio Calosi, 'Relativity and Three FourDimensionalisms', *Philosophy Compass* 11, no. 2 (2016): 102.

perspective. From your friend's perspective, it has to travel not only the length of the pole, but also the distance the train has traveled before its journey can come to an end. But light moves at the same speed for both you and your friend, so when your friend uses the speed of light to infer the time it took for the laser beam to complete its journey, because the distance he measured is *greater* than the length of the pole, he will infer that *more* time, in seconds, has passed during the beam's journey from his perspective than would have passed from your perspective.[12]

This example rings like a paradox, but it will make more sense given the notion of a reference frame. A reference frame is a collection of objects that are all at rest with respect to each other. In the context of a certain reference frame, the classical intuitions about time and distance enter back into the relativistic view of spacetime.[13] In your friend's reference frame, the train moves quickly away from the station, while in your reference frame it is the tracks that move quickly underneath the train. Neither perspective can describe the example better than the other, but because the speed of light must remain constant, we are forced by the relativity theory of spacetime to accept that time passes differently in different reference frames. Thus, the amount of time through which any object persists will be relative to the reference frame in which it is viewed.

If all there is to existence is spacetime, then this last fact—that temporal duration is relative to reference frame—should offer hope for even more ontological trash. The looming threat of ontological trash should hold generally for all objects that persist four-dimensionally in the sense that their whole reality is within spacetime, but I will use the example of a person to flesh it out. First, suppose that every life has a temporal length: the amount of time that passes from a person's birth to their death. Now suppose something even less controversial:

(Leibniz's Law) 'Objects x and y are numerically identical only if they have exactly the same properties.'[14]

entailing that two persons can be identical only if they have the same temporal length.

Suppose that a person, Fred, lives for 80 years. Any person that is identical to Fred must also live for 80 years. In Fred's own reference frame, his watch will tick at a steady rate, and when it reaches 80 years he will expire, having had no trouble discerning whether he was one and the same person over the course of his life. But if a friend of his ever escorted him to a train station and waved to him as he sped away, the person

[12] George F.R. Ellis and Ruth M. Williams, *Flat and Curved Space-times*, 2nd ed. (Oxford University Press, 2000), 28–29.

[13] Cody Gilmore, 'Persistence and Location in Relativistic Spacetime', *Philosophy Compass* 3, no. 6 (2008): 1226.

[14] Theodore Sider, 'Temporal Parts', in *Contemporary Debates in Metaphysics*, ed. Theodore Sider, John Hawthorne and Dean W. Zimmerman (Blackwell, 2007), 4.

the friend would have waved to, call him Fred*, couldn't be Fred. For as we saw above, Fred* will exist for slightly more time, as measured in the friend's reference frame, than Fred will when measured in his own reference frame. But then they wouldn't have the same temporal length, and by Leibniz's Law, Fred* couldn't be the same person as Fred.

To be clear, Fred* is not Fred, but he is definitely very Fred-like; he is made of the same stuff as Fred; and he differs from Fred only in conditions of persistence. In order to show that there is ontological trash lurking, all that remains to be shown is that there are many more things like Fred*. To seek them out, consider that Fred*, too, has a temporal length associated with his life: $80 + x$ years, where x is a positive real number. Now consider the host of beings, Fred_n, that are exactly like Fred except that their lives *at most* $80 + nx$ years long, where n ranges through the natural numbers. These beings are not persons, since their lives do not have definite temporal lengths, but they are still very Fred-like (Fred_1 is identical to Fred*, and Fred_0 just is Fred!), they are made up of the same stuff as Fred, and they differ from Fred only in conditions of persistence. As n increases, each Fred_n will be able to survive the change to frames of reference that put him in motion for more and more time. In Fred's own reference frame, all of the Fred_n coincide on him.

What could there be to make Fred ontologically superior to any of the Fred_n? Without an answer to that question, it would seem like the relativity theory of spacetime turns Fred and all other objects whose existence is fully exhausted by spacetime into ontological trash. A simple change of reference frame is enough to change the properties of an object and introduce a hoard of ontological trash. But how could a change of reference frame, that is a change in the way one observes an object, actually change that object? Indeed, Johnston affirms this confusion when he sketches the concept of ontological trash in his paper 'On Being Ontological Trash.' He writes that it is not as if 'our more specific ways of looking at or conceiving of things *thereby bring other things into being*. Rather ... [they] select from among things that are already there. [italics in original]'[15] Change of reference frame should not change an object, so there should be no question as to whether the Fred_n are identical to Fred. These beings seemed different only because they each had a different unqualified temporal duration. So instead of highlighting a new layer of ontological trash that envelops ordinary objects, Fred and the Fred_n actually serve as a *reductio ad absurdum* to the conclusion that unqualified temporal length is not a property. Therefore, an account of persistence that keeps objects fully in spacetime is possible in a relativistic spacetime, so long as temporal duration is always relative to a reference frame. For the remainder of this paper, whenever I give an unqualified time or temporal duration associated with an object, *it should be interpreted as time relative to the reference frame where that object is at rest.*

[15] Mark Johnston, 'On Being Ontological Trash' (unpublished, 2017), 8.

3 The landscape of persistence

What other constraints might relativity place on an account of persistence that is four-dimensional in the sense that it keeps objects fully in spacetime? In 'Relativity and Three Four-Dimensionalisms,' Cody Gilmore explains how relativity comes to bear on two different four-dimesnionalist views:[16] mereological perdurantism and locational perdurantism.[17] In this section I will explain what these two perdurantisms entail, why a relativistic spacetime points strongly toward the truth of locational perdurantism, and argue that locational perdurance is four-dimensionalist in the sense relevant to a four-dimensionalist account of persons.

It will be easier to understand perdurantism by contrasting it with its negation, endurantism. First, there is the domain of mereology, which is the study of a how a whole is composed of its parts. Mereological perdurantism is the view that all complex objects are composed of temporal parts. An object mereologically perdures if and only if it is a series of achronal chunks, or object-stages, that succeed each other over time. Mereological endurantism, however, holds that objects do not have temporal parts. Instead, an object mereologically endures if and only if it is wholly present whenever it exists.[18] The heart of this side of the debate between perdurantism and endurantism is about which is more fundamental: an object's presence at each time it is present or its presence throughout its lifetime. A mereological perdurantist thinks an object's temporal parts explain its entirety, whereas a mereological endurantist thinks an object's entirety explains its presence at each time it is present.

Second, there is the domain of location, which concerns the precise region where an object is found. Locational perdurance is the view that material objects occupy only their whole spacetime path. So an object locationally perdures if and only if the single place it is located is the four-dimensional region that is its whole career, swept out through spacetime. Locational endurance, on the other hand, is the view that material objects occupy many different regions: the manifold achronal chunks of their path. So an object locationally endures if and only if it occupies many regions and each region it occupies is a three-dimensional slice of its path at a time. The crux of this side of the debate is over which sort of region is more fundamental for an object: its four-dimensional whole or its three-dimensional manifestations at times. The locational perdurantist believes that the four-dimensional region of an object's course through spacetime explains the smaller three-dimensional regions that that object has at different times, whereas the locational endurantist holds that an object's three-dimensional shape at the times when it is present explains the four-dimensional region it sweeps

[16] Gilmore, Costa and Calosi, 'Relativity...', 102.

[17] Gilmore, 'Persistence and Location in Relativistic Spacetime', 1227.

[18] Katherine Hawley, 'Temporal Parts', in *The Stanford Encyclopedia of Philosophy*, Spring 2018 edition, ed. Edward N. Zalta (2018), `https://plato.stanford.edu/archives/spr2018/entries/temporal-parts/`.

out.[19]

The issues of these two debates are quite similar, but they remain nevertheless independent. Likewise, relativity does not support both views in the same way. Gilmore presents detailed versions of the arguments from relativity theory to both forms of perdurantism in sections of 'Relativity and Three Four-Dimensionalisms'[20] and in 'Persistence and Location in Relativistic Spacetime,'[21] but they require too much knowledge of spacetime geometry to present here. Instead, I will take his conclusions that it is very likely that space and time are not fundamentally separate entities,[22] and that this implies locational perdurantism.[23] Therefore, a locationally perdurantist account of persistence will be consistent with a relativistic spacetime.

Such a view also implies that the reality of an object might be wholly exhausted by its spatiotemporal extent. For what is there to push an object outside of its spacetime envelope if the region it occupies just is its spacetime envelope? An object *could* exist partially outside of spacetime and could locationally perdure in the sense that the region in spacetime that it occupies is its four-dimensional career through spacetime even though this region is not all there is to the object. But this is just one flavor of the view. Locational perdurantism is also consistent with both mereological perdurantism and endurantism. Relativity thus leaves two options on the table for an account of perdurance that keeps objects fully in spacetime.

4 Taking out the trash

Now that it is clear that there is wiggle room within relativity for an account of persistence that is four-dimensionalist in the relevant sense, the next task is to see whether such an account can also avoid ontological trash. When Johnston sketches out the concept of ontological trash, he considers two possible accounts of persistence through time, both of which are consistent with locational perdurantism. The first is a type of mereological endurantism, in that 'at each time [there are] a plenitude of co-extensive objects, each with a different condition of survival, some of which get teased out by this or that change.' The second is mereological perdurantism, where 'sequences or parades or cross-time sums of short-lived objects, temporal stages of [things]' compose complex objects.[24] In this section, I will present Johnston's arguments to ontological trash from mereological perdurance and from a common form of mereological

[19] Cody Gilmore, 'Building Enduring Objects Out of Spacetime', in *Mereology and the Sciences: Parts and Wholes in the Contemporary Scientific Context*, ed. Claudio Calosi and Pierluigi Graziani (Springer, 2014), 9.

[20] Gilmore, Costa and Calosi, 'Relativity...', 11–14.

[21] Gilmore, 'Persistence and Location in Relativistic Spacetime', 1299–35.

[22] Gilmore, Costa and Calosi, 'Relativity...', 4.

[23] Gilmore, 'Persistence and Location in Relativistic Spacetime', 1235.

[24] Johnston, 'On Being...', 8.

endurance, and then show that amid the mereologically enduring ontological trash, there is always one object that can claim ontological supremacy.

Mereological perdurance straightforwardly entails that objects are ontological trash. First, consider a fist as a mereological perdurantist would see it. A fist only exists because a collection of a short-lived fist-stages succeed each other for a given interval of time. So whenever you have a fist, you will also have objects that are composed of all of the fist's temporal parts plus some temporal parts of the hand from right before it was clenched, from right after it was unclenched, or from both times. You will also have many objects that are composed of all of the fist's temporal parts except a couple from right after it was clenched, from right before it was unclenched, or from both. All of these objects are nearly identical with a fist; they are composed of the same stuff as a fist; and differ from a fist only in their conditions of persistence. Because they all overlap on your fully clenched fist, why are we to suppose we are looking at the maximal fist and not at any of its *doppelgängers*? Thus, persistence on mereological perdurantism is hopelessly ontologically trashy.[25]

The path to ontological trash from mereological endurantistism is a bit less obvious. Suppose that fists cannot be reduced to temporal parts. Instead, a fist is a hand that is clenched all the way and it survives until the hand is unclenched to a lesser a degree. However, if this account correctly describes an object, then there is also the half-fist: a hand clenched half of the way that survives until it is either clenched more or unclenched, as well as the quarter-fist and eighth-fist and so on for every fraction of a fist. And there is nothing to rule out the definition of the at-least-half fist, which is identical to the half-fist but which can survive further clenching, and the at-least-quarter fist and so on for every fractional fist. So whenever you have a fist, you have a host of at-least fists which are all very fist-like; they are all made up of the same stuff: a hand with fingers rolled to a degree or range of degrees; and they differ only in conditions of persistence. Thus, even on a mereologically enduring view of persistence, there is ontological trash.

We should not hold out hope for finding a non-ontologically-trashy definition of a fist, but we can acknowledge that buried at the bottom of the trash heap there is an object upon which all the others are ontologically derivative: the hand. Why must we demand that the fist exist in its own right? It's not as if the hand disappears when we look at the fist. Rather, the fist seems like a phase of the hand's existence. In general:

> If an object x is defined by possessing a property F continuously through time to degree d, where d could range through a plurality of values, the non-ontologically-trashy substitute for x will have F continuously through time *to any degree at all*.

This account will not render objects exactly as we expect them to be. We must wave

[25] Johnston, 'Is Hope...', 7.

goodbye to the idea of fists as ontologically basic. But only on this account can any object in the vicinity of the fist fully exist in spacetime and emerge from the ontological trash heap.

5 Salvaging the four-dimensional account of persons

With a four-dimensionalist account of persistence that avoids ontological trash, the path is clear to rescue the four-dimensionalist view of a person. In this section I will present such a view by adapting Derek Parfit's reductionist account of personhood[26] to the schema I introduced above, and then I will explain why this counterintuitive solution should make sense.

First, a word about Parfit's account. It is reductionist in the sense that it holds that all there is to personhood is the holding of other, more specific facts concerning psychological continuity and bodily continuity. Johnston adapts Parfit's view so that only psychological continuity is relevant to personhood and presents it as such:

> A person x, considered at t_1, is numerically one and the same person as a person y, considered at t_2, if and only if the mental profile (the congeries of mental states and events) exhibited by x at t_1 is D_o psychologically continuous with the mental profile exhibited by y at t_2; (where D_o is construed as the [relevant] degree of psychological connectedness. . .)[27]

Johnston demonstrates that, on this account, although persons mereologically endure,[28] there are still personites in the form of continuity variants that are psychologically continuous to more restrictive or less restrictive degrees. Thus, on the psychological variant of Parfit's account we are ontological trash. Johnston acknowledges, however, that a 'continuity variant that places the least demands on connectedness, if there is such a one' would be the only way out of this case of personites. That way, 'all the other continuity variants. . . might be able to be construed as phases on such least demanding wholes.'[29]

My proposal is that a person just is that least demanding continuity variant. Put more precisely:

> A person x, considered at t_1, is numerically one and the same person as a person y, considered at t_2, if and only if the mental profile (the congeries of

[26] Derek Parfit, *Reasons and Persons* (Oxford University Press, 1984), 207, quoted in Mark Johnston, 'Personites, Maximality And Ontological Trash', *Philosophical Perspectives* 30, no. 1 (2017): 225

[27] Ibid.

[28] Ibid., 224.

[29] Ibid., 227.

> mental states and events) exhibited by x at t_1 is psychologically continuous *to any degree at all* with the mental profile exhibited by y at t_2.

Questions of survival in the classic cases of amnesia, teletransportation, fusion, and fission, as well as the possibility of a resurrection, should all be treated similarly under this view as they were under the psychological variant of Parfit's original view.[30] Parfit would not endorse my view, since he held that '[i]f there was only a single [direct psychological] connection, x [today] and y [yesterday] would not be on the revised Lockean view the same person,'[31] and this minimal psychological connection just is the criterion of identity on my account. However, if spacetime is all there is, then my account is a non-trashy, ontologically superior alternative to Parfit's account and Johnston's continuity variants.

Psychological connectedness, and therefore its ancestral relation continuity, does admit of degrees, but that is no reason to think that a stronger degree of connectedness enables some psychologically persisting entities to survive where other, more weakly continuous entities would cease to be. To get a sense for why this is so, imagine two persons, Joan and Joni. Suppose that there is *a trace amount* of psychological continuity between Joan considered at t_1 and Joni considered at t_2, but not enough for Joan and Joni to be numerically one and the same. Perhaps this is a very severe case of partial amnesia. Joan's friends and family would certainly think that Joni is a ghost of the person they knew, and they would likely mourn the absence of Joan. But would Joni, a newly minted person, have to apply for citizenship? Should Joan's next of kin execute the final will and testament of their late beloved? And if Joni attempted to learn Joan's tendencies and to embrace Joan's personality, would she be at best an imposter for the real Joan who disappeared long ago? This borderline case should show that although our feelings about personhood respond to a minimum threshold of psychological continuity, falling below that threshold should not *actually constitute death*. Some very person-like thing does survive such a drastic change. What could it be other than that very person?

6 Conclusion

In this paper I have shown through a *reductio ad absurdum* that, in a relativistic spacetime, unqualified temporal duration is not a property; identified a version of persistence that is permissible in a relativistic spacetime and is properly four-dimensionalist in the sense it keeps objects fully in spacetime; given a four-dimensionalist account of persistence that avoids ontological trash; and finally, defended a definition of a person that persists in such a manner.

[30] This view could even be supplemented with an additional clause about some sort of bodily continuity if evidence is found to suggest that bodily continuity should also matter in survival.

[31] Parfit, *Reasons and Persons*, 207, quoted in Johnston, 'Is Hope...', 8

With such a resilient interpretation of what it is to be a person, we can coherently think of ourselves as objects whose existence is fully captured by our spacetime envelopes. We are surrounded by ontological trash (e.g. our fists) but we are not, ourselves, ontological trash. Thus, we can rescue practical reason from the personite problem without believing that part of us must be outside of spacetime. But in order to lift ourselves above the personites, we must admit that the single person seems to survive *too much*. All it takes to survive is a chain of continuity made of the weakest possible links of psychological connectedness. Assuming that Shoemaker's theory of psychological connectedness as causal dependence is the most tenable account, a robust theory of personal identity will explore the weakest sort of a causal dependence that still counts as psychological connected.[32] For instance, do mental states only transitively causally linked still count as psychologically connected? Consider a person's mental state at t_1 when writing something down and their mental state at t_2 when reading what they wrote. These mental states are ordinarily connected, e.g. in the case of a to-do list. Are they still connected if the person suffers amnesia between t_1 and t_2? How does this case differ from the relationship between the mental states of the writing author and the reading reader? Answering these and similar questions will lead us to a clearer picture of the new sort of personhood we should welcome on four-dimensionalism.[33]

References

Ellis, George F.R., and Ruth M. Williams. *Flat and Curved Space-times*. 2nd ed. Oxford University Press, 2000.

Gilmore, Cody. 'Building Enduring Objects Out of Spacetime'. In *Mereology and the Sciences: Parts and Wholes in the Contemporary Scientific Context*, edited by Claudio Calosi and Pierluigi Graziani, 5–34. Springer, 2014.

———. 'Persistence and Location in Relativistic Spacetime'. *Philosophy Compass* 3, no. 6 (2008): 1224–54.

———. 'Where in the Relativistic World Are We?' *Philosophical Perspectives* 20, no. 1 (2006): 199–236.

Gilmore, Cody, Damiano Costa and Claudio Calosi. 'Relativity and Three FourDimensionalisms'. *Philosophy Compass* 11, no. 2 (2016): 102–20.

[32] Eric T. Olson, 'Personal Identity', in *The Stanford Encyclopedia of Philosophy*, Summer 2017 edition, ed. Edward N. Zalta (2017), `https://plato.stanford.edu/archives/sum2017/entries/identity-personal/`.

[33] I am very grateful to the Princeton University Office of Undergraduate Research and the President's Fund for making it possible for me to present this paper at the 23rd Annual Pacific University Philosophy Conference on April 6, 2019 in Forest Grove, Oregon. The conference was a wonderful opportunity to present my ideas, field questions, hear what my peers on the west coast thought about this topic, and to be exposed to their interesting ideas.

Hawley, Katherine. 'Temporal Parts'. In *The Stanford Encyclopedia of Philosophy*, Spring 2018 edition, edited by Edward N. Zalta. 2018. `https://plato.stanford.edu/archives/spr2018/entries/temporal-parts/`.

Huggett, Nick, and Carl Hoefer. 'Absolute and Relational Theories of Space and Motion'. In *The Stanford Encyclopedia of Philosophy*, Spring 2018 edition, edited by Edward N. Zalta. 2018. `https://plato.stanford.edu/archives/spr2018/entries/spacetime-theories/`.

Johnston, Mark. 'Is Hope For Another Life Rational?' Unpublished, 2017.

———. 'On Being Ontological Trash'. Unpublished, 2017.

———. 'Personites, Maximality And Ontological Trash'. *Philosophical Perspectives* 30, no. 1 (2017): 198–228.

———. 'The Personite Problem: Should Practical Reason Be Tabled?' *Noûs* 50, no. 4 (2016): 617–44.

Olson, Eric T. 'Personal Identity'. In *The Stanford Encyclopedia of Philosophy*, Summer 2017 edition, edited by Edward N. Zalta. 2017. `https://plato.stanford.edu/archives/sum2017/entries/identity-personal/`.

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

Sider, Theodore. 'Temporal Parts'. In *Contemporary Debates in Metaphysics*, edited by Theodore Sider, John Hawthorne and Dean W. Zimmerman, 241–62. Blackwell, 2007.

Has Horowitz Split Level-Splitting?

Savannah Leon[*]
University of California, Berkeley

Abstract What is to be done when first- and higher-order evidence point in opposite directions concerning the truth about *p*? The traditional response goes that ideally rational agents ought to privilege one evidential order over the other, such that an agent's belief that *p* co-varies with her total evidence. But the level-splitter zigs where others zag. Since each evidential order appears perfectly good in isolation, she supposes her credences should be partitioned accordingly. On penalty of believing against her total evidence, she responds to the pull of both evidential orders. In other words, she is epistemically akratic. Sophie Horowitz has recently argued that level-splitting views are almost universally irrational. To show as much, she points to some cases of peer disagreement where a pro-akrasia verdict requires (irrationally) concluding that S's evidence is misleading. The purpose of this paper is to deny that an on-off conception of agent-specific defeaters is called for: that is, I argue that peer disagreement need not necessarily banish first-order evidence to the realm of the misleading, and that a different approach is available to the pro-akrasia crowd.

1 Introduction

In 'Epistemic Akrasia,' Sophie Horowitz argues that while rational epistemic akrasia can be warranted in special cases, a pro-akrasia solution in standard cases is too intuitively costly to be right (Horowitz 2014). The problem of akratic belief states has inspired a triad of potential solutions in the literature of epistemology of disagreement: two traditional, one contemporary. Traditional responses reject the notion of rational epistemic akrasia. That is, they reject the view that one ought to believe *p* while simultaneously believing that *p* is unsupported by evidence. A contemporary response

[*] Savannah Leon is now a fourth year philosophy student at the University of California, Berkeley. Her philosophical interests include epistemology, philosophy of language, and some formal topics. Besides doing philosophy, she also enjoys travelling, reading poetry, and cooking. She plans on applying to philosophy programs in the UK and US as a prospective graduate student after finishing at Berkeley.

is that such belief states are only *prima facie* problematic and may even be rationally required.

The paper is structured as follows. In §1, I introduce Horowitz's Sleepy Detective Problem and three responses to it: the traditional (anti-akrasia) *conciliatory* and *steadfast* verdicts and the newer (pro-akrasia) *level-splitting* verdict. In §2, I motivate rational epistemic akrasia. I foreshadow Horowitz's reasons for thinking that level-splitting is irrational in most cases, yet rationally required in others. In §3, I discuss two further cases and give an overview of Horowitz's account of evidential uncertainty and how evidence can be considered either truth- or falsity-guiding. In §4, I give her argument in support of level-splitting's nearly universal irrationality, which features cases of peer disagreement where pro-akrasia requires (irrationally) concluding that one's evidence is misleading. In §5, I argue that these cases hinge on an assumption that a certain species of defeater commits level-splitters to forming the belief that their evidence is misleading, and that since this assumption is false, it doesn't follow that level-splitters must conclude that their evidence is falsity-guiding. In doing this, I indirectly offer an alternative explanation for why akratic belief states may sometimes be rational despite the worry posed by Horowitz.

2 What is epistemic akrasia?

'Akrasia' has classically meant *weakness of will*, such as in cases where S acts against her better judgment.[1] Correspondingly, 'epistemic akrasia' refers to cases wherein S arguably believes p against her better judgment. (For example, when her belief that p appears inadequately supported by her available evidence.) While not all-encompassing (and purposely somewhat imprecise, given current debates as to just what epistemic akrasia *is*), this definition shall serve as a reasonable point of departure in understanding what it is to be epistemically akratic.

The paper will proceed, as Horowitz does, with the evidentialist approach to understanding epistemic akrasia. Provisionally, I will consider the puzzle of whether epistemically akratic belief states are to be thought rational as one which can be solved or dissolved by settling on how an agent should apportion her belief that p with respect to her total evidence concerning p. If we suppose that two or more crucial parts of an agent's available evidence is both in favor of and against believing that p (and that she is unable to suspend judgment), and the agent is ultimately uncertain as to whether p, it otherwise remains unclear how and in which circumstances her belief that p could

[1] An attempt to account for the 'paradoxical irrationality' of akratic agents is of course discussed in the Protagoras, and also in Davidson (1982). Another account, that of Levy (2018), argues that what we take to be (epistemically) akratic states are an agent's mistaken belief that they believe that p. Levy differentiates between belief that p and a(n) agent's indistinct first-order 'beliefy' representation(s) that p.

be rationally required.

Part of this puzzle's traction is owed to emerging controversies about evidence. The controversy currently at hand turns on a tension between two sorts of evidence: first- and higher-order evidence. The first term is maybe the most familiar: 'first-order evidence' is that which bears directly on *p*'s truth value. Standard sources of first-order evidence include (non-exhaustively) perception and memory. Contrastingly, 'higher-order evidence' is typically thought to be evidence *about* one's evidence, or evidence 'bearing on the functioning of one's rational faculties, or on the significance of other evidence that one has' (see Horowitz, forthcoming). Higher-order evidence, then, can speak to how S ought to interpret her first-order evidence, and often crops up in the form of testimony (as in cases of peer disagreement) or even reflection upon one's own epistemic state (such the recognition of impairment or lack of expertise). For instance, higher-order evidence's impact on total evidence (especially concerning defeat, i.e., whether higher-order evidence can undercut or rebut first-order evidence) remains at large in recent literature (Christensen 2010; Feldman 2009; Lasonen-Aarnio 2014), and what differentiates higher-order evidence from first-order evidence is also an open question.

There are many cases that serve as excellent candidates for framing the kind of evidential tension which might rationalize epistemic akrasia.[2] For simplicity, I'll use Horowitz's: its key details bear centrally and specifically on her criticisms of level-splitting and the subject of this paper. The case:

> **Horowitz's Sleepy Detective** Sam is a police detective, working to identify a jewel thief. He knows he has good evidence—out of the many suspects, it will strongly support one of them. Late one night, after hours of cracking codes and scrutinizing photographs and letters, he finally comes to the conclusion that the thief was Lucy. Sam is quite confident that his evidence points to Lucy's guilt, and he is quite confident that Lucy committed the crime. In fact, he has accommodated his evidence correctly, and his beliefs are justified. He calls his partner, Alex. 'I've gone through all the evidence,' Sam says, 'and it all points to one person! I've found the thief!' But Alex is unimpressed. She replies: 'I can tell you've been up all night working on this. Nine times out of the last ten, your late-night reasoning has been quite sloppy. You're always very confident that you've found the culprit, but you're almost always wrong about what the evidence supports. So your evidence probably doesn't support Lucy in this case.' Though Sam hadn't attended to his track

[2] Consider, for example, Alvin Plantinga's 'letter filching case' (1986), where a man accused of stealing a letter has (arguably excellent) first-order evidence that he didn't take it—in this case, his memory of walking in the woods. He is in fact correct about his woods memory. As with the Sleepy Detective case, there's nevertheless a large body of higher-order evidence against him: he's done similar things before, and an extremely reliable person testifies that she witnessed the theft.

> record before, he rationally trusts Alex and believes that she is right—that he is usually wrong about what the evidence supports on occasions similar to this one (Horowitz 2014, 2).

To specify: Sam's first-order evidence (hereafter 'FOE') is the aforementioned codes, letters, and photographs. Sam's higher-order evidence (hereafter 'HOE') consists in Alex's testimony that Sam is probably unable to properly interpret FOE. Taking together Sam's first- and higher-order evidence provides a (rough) picture of his 'total evidence.'[3] Next I will spell out what each possible verdict has to say about how these evidential orders should interact with one another, if at all.

One might think that the two evidential orders *don't* (or shouldn't) interact, and furthermore that Sam might be rational in believing both that Lucy is the culprit while also accepting Alex's testimony that he is often wrong in situations such as these. This option permits rational epistemic akrasia.[4] The view that akratic attitudes can be rationally required in cases like Sleepy Detective is what Horowitz dubs the 'level-splitting' position. For our purposes, the shortened 'pro-akrasia' will often be used. If the detective is epistemically akratic, then he'll continue to believe that Lucy is the jewel thief while believing that his total evidence doesn't support this. On pro-akrasia, then, he should split his levels of confidence and hold onto the first-order belief that p given FOE and the higher-order belief about his unreliability given HOE.

The traditional 'anti-akrasia' response is to deny that epistemically akratic states are rational. On this view, it is often thought that epistemic levels should *never* operate separately. An anti-akrasia proponent would apply this notion in the Sleepy Detective case by requiring that Sam must base his belief in Lucy's guilt on his total evidence. For Sam, this would mean that he must either steadfastly remain confident that Lucy is guilty on the basis of FOE, or else be persuaded to reduce his confidence in her guilt given Alex's testimony that he ought to doubt his initial conclusion to the contrary.

3 Level-splitting

Horowitz calls the pro-akrasia view that epistemic levels should operate separately 'level-splitting' (Horowitz 2014), which is to be properly differentiated from epistemic akrasia. Rather than merely labelling cases of 'divergence between first- and higher-

[3] It's not uncontroversial that Sam's total evidence might contain more or less than what is said here; this can vary depending on one's views concerning which evidence an agent ought to consider.

[4] I'm grateful to the aforementioned anonymous referee for pointing out that although level-splitters hold the view that evidential orders should operate separately and that this goes some way in rationalizing epistemic akrasia, one's commitment to the interaction of evidential orders isn't simultaneously a commitment to level-splitting in epistemic akrasia cases. One might think, for example, that epistemic akrasia is irrational without this entailing that evidential orders shouldn't influence one another, and vice-versa.

order attitudes' (Horowitz 2014), level-splitting views rationally require such divergences.[5]

In other words: level-splitting is a normative position, whereas epistemic akrasia is used descriptively. I'd also like to note that Horowitz perceives that first- and higher-order belief mismatches as coming in degrees:

- Moderate level-splitting prescribes being highly confident that p despite having high confidence that your evidence that doesn't support your degree of credence in p.

- Extreme level-splitting recommends high confidence in p while also being rationally highly confident that (a) your evidence doesn't support p, (b) your evidence supports low confidence in p, or (c) your evidence supports $\neg p$.

These distinctions are a sticking point in section 4, where I'll zero in on the belief state depicted in 'C'. Unless otherwise specified, it can be assumed for now that when I refer to level-splitting, I have in mind the gamut of akratic states given above.

A defense of pro-akrasia is founded in the thought that if either evidential order seems perfectly good to us in isolation, our belief state ought to reflect this somehow. On the basis of both FOE + HOE, the level-splitter believes that p and believes that there's something fishy about her total evidence for p. Again, many grounds for fishiness exist: misleading evidence, poorly-interpreted evidence, insignificant evidence, etc.

Horowitz argues that, barring a complex and much-discussed case, epistemic akrasia is universally irrational.[6] This caveat leads her to distinguish between two kinds of cases, standard and nonstandard. We shall feature an example or two from each in sections 2.3 and 2.4. The cases are classed according to a couple of contrasting features: (1) different types of uncertainty, and (2) opposite background expectations about how our evidence should point to the truth about p.

Uncertainty In standard cases, a would-be akratic agent is uncertain about what her total evidence supports. The difficulty here lies in discerning which order of evidence is really getting it right about p. In nonstandard cases, even if she can be sure of what her evidence should support, the problem is that she can't be sure of what her evidence *is*.

Truth- and falsity-guiding evidence Standardly, and nearly unanimously, we expect our evidence to be truth-guiding, so that 'when it justifies high confidence in a

[5] The two might easily be seen as interchangeable. I wish to avoid this confusion.

[6] The exceptional case is that of Williamson's 'irritatingly austere' clock (from Williamson 2011, 2014 discussed at length in Elga 2013); see p. 6 for Horowitz's adaptation.

proposition, that proposition is usually true, and when it justifies low confidence in a proposition, that proposition is usually false' (Horowitz 2014). But in unusual cases, we can have the background expectation that our evidence—whatever it is—will vary *falsely* with the proposition it is meant to support. That is, we expect that it will be FALSITY-GUIDING: it'll support high confidence in a false proposition, and low confidence in a proposition we think is probably true. Horowitz concludes that special cases like these are rare, but plausibly justify pro-akrasia.

Horowitz considers a number of epistemologists who argue that epistemic akrasia is rationally required.[7] We shall consider a pair of cases from two authors in particular: Weatherson (n.d.), who uses a familiar case of justified moral akrasia, and Williamson (2011, 2014), whose case is purely epistemic. The abridged version of each:

> **Weatherson's Kantian Professor** By way of sophisticated and persuasive argumentation, suppose your Kantian professor has given you good evidence to believe that lying is categorically wrong. Nevertheless, when a murderer inquires as to your roommate's whereabouts, you lie, since lying is what you ought to do. (Horowitz 2014)

> **Williamson's Long Deduction** Suppose a rational agent comes to know a long series of claims and deduces their conjunction, *C*. She's done so competently, but she realizes that since oftentimes memory and logical ability are limited, people in her situation often make inferential errors while completing long deductions. It's then highly probable on her evidence that she herself has made such an error, and thus that she doesn't know the conjunction. Still, given that she's competently deduced *C*, she knows *C*: its evidential probability is 1. It's nevertheless highly probable on her evidence about fallibility during long deductions that she doesn't know *C*. So, she should be highly confident in *C* despite her high confidence that she doesn't know *C*. (Horowitz 2014)

Borrowing a line from Lewis (1996), it seems a level-splitting agent can rationally 'properly ignore' evidence across epistemic levels: that is, form a kind of provisional belief that p. Acknowledging the evidential force of FOE + HOE rids us of the drawback of ignoring good evidence—in these cases, a properly performed proof and a well-formulated normative claim.

[7] Whatever the authors themselves may make of level-splitting, I'll follow Horowitz in proceeding as though the views presented here commit them to it.

4 Problems for level-splitting

Horowitz gives more than a few examples where making the choice to split epistemic levels goes terribly wrong. Central to this paper, however, is one case which is supposed to demonstrate that Sam—were he epistemically akratic—would have strange beliefs indeed about where his evidence points. In particular, Horowitz says: 'If [the detective] takes both "Lucy is the jewel thief" and "my evidence doesn't support Lucy" as premises, it seems he can engage in some patently bad reasoning' (Horowitz 2014).

Horowitz's argument is as follows. Suppose that Sam trusts Alex's testimony and forms the belief that the odds are 1:9 that Lucy is guilty. Suppose further that despite this, he remains confident that she is the culprit. Horowitz reasons that the detective must then think that, given such low odds, he 'just got lucky' about his true belief. Given that he rationally trusts Alex, Sam *should* be confident FOE doesn't support Lucy's guilt. (That is, he should have low confidence that p.) His HOE therefore pushes him towards high confidence that $\neg p$. A plausible (extreme) pro-akrasia reading of his *total* evidence, then, is that it's falsity-guiding: that is, it supports high confidence in a false proposition, $\neg p$.

Horowitz gives an analogous case, spiked coffee, where Sam has received HOE (Alex's testimony) that his coffee's been spiked with a reason-distorting serum (Horowitz 2014). His coffee has not in fact been spiked. Although he has the reasonable background expectation that his FOE would be truth-guiding under normal conditions, his partner's testimony causes him to form the belief that he can't properly evaluate his evidence. Nevertheless, he has properly evaluated his evidence, and p is a true belief. As in the above example, however, he also believes that his evidence is misleading on the basis of HOE. This, argues Horowitz, *shouldn't* be enough to cause Sam to form the belief that his FOE is misleading, so it's irrational to believe as much given HOE.

In what follows, I'll examine the claim that extreme pro-akrasia verdicts of the form 'my evidence doesn't support p', or 'my evidence supports low confidence in p' must land level-splitters in the predicament given above. The Spiked Coffee case, when paired with its 'just got lucky' predecessor, is supposed to point out the following absurdity. It would be absurd for an agent to form the belief that her evidence supports $\neg p$—that is, that her evidence is misleading—merely on the basis of a defeater bearing on her agent's diminished capacities (a 'self-doubting defeater').

5 Spiked Coffee, revisited

Horowitz writes:

> If Level-Splitting is right, and extreme cases of epistemic akrasia can be rational in Sleepy Detective, there is nothing wrong with Sam's concluding that his evidence is misleading in this way. But there *is* something wrong with Sam's concluding that his evidence is misleading in this case. This suggests that there is something wrong with Level-Splitting. (Horowitz 2014).

For the record, I'm confident Horowitz is right on the order of it being silly for the detective, spiked coffee or no, to draw the conclusion that his evidence is falsity-guiding. It doesn't seem at all plausible that a claim bearing on an agent's capacities should tarnish first-order evidence to the effect that it supports $\neg p$.

I'll argue against Horowitz that a level-splitter needn't conclude from a self-doubting defeater ('your coffee's been spiked, so you should doubt your ability to interpret your evidence!') that her first-order evidence is misleading. This would, as Horowitz has claimed, be irrational. I leave open the possibility that that some level-splitter or other might draw such a conclusion, but by my lights, there is an available alternative.

Might we say instead that the detective's confidence that Lucy is guilty should be *reduced* by his knowledge of his track record? Alternatively, we could just as well leave open the possibility that the detective might refrain from believing anything about whether his first-order evidence points to falsehood or truth. It's not clear why self-doubting defeaters must have an on-off effect on belief in such a way as to pressure extreme level-splitters to immediately interpret their evidence as only truth-guiding or falsity-guiding, but not something in between. The defeater might be classified as undercutting, or alternatively as neither raising nor lowering the probability that p. These interpretations sidestep the need for self-doubting defeaters to outright rebut the proposition that one's first-order evidence supports p.

If, for example, I were to espouse an extreme level-splitting view and then interpret my evidence under the influence of Irish coffee, I wouldn't form the belief that my FOE is—or must be—falsity-guiding. Depending on the amount of Irish coffee involved, I might reduce my confidence as to where my evidence points, or alternatively take the 'wait-and-see' approach until morning. Still, I needn't believe that my spiked coffee points away from the truth about p, and neither does the detective. He could refrain from believing anything about whether his first-order evidence points towards truth or falsehood, or merely reduce his confidence that Lucy is guilty upon becoming aware of his track-record.

In short, level-splitters aren't committed to the view that self-doubting defeaters are necessarily falsity-guiding, or that such defeaters should always have the effect of transforming one's evidence that p into evidence that $\neg p$. When paired with other compelling strands of evidence, a creeping suspicion of one's own unreliability *might* lead an agent to wonder about what her first-order evidence really supports, but then again it may not.

The same point evinces, I think, in another stock example of epistemic akrasia. If we consider Williamson's Long Deduction case, we might also think that a tendency to make inferential errors of this sort shouldn't serve the same evidential role as an added line in a proof that points away from your conclusion that *C*. Still, the extremely akratic logician doesn't appear committed to concluding that her evidence is misleading. This is because she can just as easily maintain high confidence that *C* while believing that her evidence could be something unnervingly short of truth-guiding.

If my analysis is correct, the detective needn't worry about his partner's testimony showing his FOE to be falsity-guiding; i.e., that it perplexingly supports Lucy's innocence ($\neg p$)[8], or somehow shrouds the truth about p in a thinner sense than rebutting it.[9] In much the same way, it would also helpfully tie up the loose end that the extreme level-splitter's answer in long deduction (as initially stated) isn't doomed to the same fate.

6 Conclusion

On Horowitz's view, an immediate problem with level-splitting is that it permits irrationally concluding that one's evidence is misleading in cases like Sleepy Detective. This, she argues, is problematic: clearly the Sleepy Detective Problem's would-be akratic agent can avoid being misled and can even point to a belief of his that should be revised given his total evidence. I've concluded that if level-splitting is correct, then it's not the case that this evidence *must* be permissibly interpreted as misleading. There may be internecine disagreements among level-splitters as to why and whether this kind of move can be rational. In this case, level-splitters need not *all* think it permissible to interpret evidence in this fashion, so Horowitz's criticism doesn't seem to count against the entire position.

The Sleepy Detective Problem's very setup permits believing (rationally) that one's first-order evidence might not be misleading. Why? I suspect this is because the self-doubting defeater contained in Alex's testimony is a long way off from serving as positive evidence for the devastating conclusion that Sam, were he to continue believing p, has been knowingly misled and yet still believes p. By my lights, this also tells a plausible story about why a level-splitter's belief state can be stable in cases like Williamson's Long Deduction. Indeed, it's hard to see how just about any close reading of the case would lead one to believe that the akratic logician's higher-order evidence supports $\neg p$. For again, if I competently completed a long proof, the likelihood of my

[8] While I find this perplexing given how I've read the evidential support relation as being 'HOE rationalizes low confidence in p' (and so on, and so on), I realize that the detective could very well have it that HOE makes Lucy's guilt less likely than her innocence. Thanks to an anonymous commentator at UCSD for pressing me on this.

[9] Thanks to Jennifer Carr for a rich discussion of this point and for an elegant formulation of it.

getting it wrong about the conclusion wouldn't be enough to induce credence x that something *other* than C holds: I'd just have good higher-order evidence to suspend belief that C, or reduce my credence in the proposition 'I know C.' I'd hardly be forced to revise my belief state such that I have low confidence in C being the answer.

Even if Horowitz has split level-splitting, I've disagreed with the portion of her strategy wherein an on-off conception of defeaters is called for: that is, if peer disagreement *must* banish first-order evidence to the realm of the misleading. I've offered an alternative that can better explain how akratic belief states might be rational. This point has perhaps a small yield in terms of the broader debate about responding to counterevidence, but if correct, it extricates split-friendly epistemologists from the view that their first-order evidence must be misleading due to higher-order evidence to the contrary. It allows the level-splitter to retain confidence that her evidence is truth-guiding and vindicates the evidence they've evaluated from being doomed to falsity-guiding status in paradigm cases.[10]

References

Alston, William P. 1980. 'Level-Confusions in Epistemology'. *Midwest Studies in Philosophy* 5 (1): 135–50.

Carr, Jennifer. n.d. 'Imprecise Evidence Without Imprecise Credences'. Unpublished.

Christensen, David. 2010. 'Higher-Order Evidence'. *Philosophy and Phenomenological Research* 81 (1): 185–215.

Davidson, Donald. 1982. 'Paradoxes of Irrationality'. In *Problems of Rationality,* edited by Donald Davidson. Oxford University Press.

Egan, Andy, and Adam Elga. 2005. 'I Can't Believe I'm Stupid'. *Philosophical Perspectives* 19 (1): 77–93.

Elga, Adam. 2013. 'The Puzzle of the Unmarked Clock and the New Rational Reflection Principle'. *Philosophical Studies* 164 (1): 127–39.

Feldman, Richard. 2009. 'Evidentialism, Higher-Order Evidence, and Disagreement'. *Episteme* 6 (3): 294–312.

Horowitz, Sophie. 2014. 'Epistemic Akrasia'. *Noûs* 48 (4): 718–44.

[10] I'm indebted to Jennifer Carr and the audience at Columbia University. Milo Phillips-Brown, Kevin Dorst, and Dee Payton have provided extensive comments on this paper for which I am also grateful. Enormous thanks are in order to an anonymous referee at St Andrews whose constructive feedback led to many positive changes in the paper.

Horowitz, Sophie. Forthcoming. ‘Predictably Misleading Evidence’. In *Higher-Order Evidence: New Essays,* edited by Mattias Skipper and Asbjørn Steglich-Petersen. Oxford University Press.

Lasonen-Aarnio, Maria. 2014. ‘Higher-Order Evidence and the Limits of Defeat’. *Philosophy and Phenomenological Research* 88 (2): 314–45.

Levy, Neil. 2018. ‘You Meta Believe It’. *European Journal of Philosophy* 26 (2): 814–26.

Lewis, David. 1996. ‘Elusive Knowledge’. *Australasian Journal of Philosophy* 74 (4): 549–67.

Plantinga, Alvin. 1986. ‘The Foundations of Theism: A Reply’. *Faith and Philosophy* 3 (3): 313–96.

Weatherson, Brian. n.d. ‘Do Judgments Screen Evidence?’ Unpublished.

Williamson, Timothy. 2011. ‘Improbable Knowing’. In *Evidentialism and its Discontents,* edited by T. Dougherty, 147–64. Oxford University Press.

———. 2014. ‘Very Improbable Knowing’. *Erkenntnis* 79 (5): 971–99.

Why Virtue Ethics?

Action and motivation in virtue ethics

Norah Woodcock[*]
McGill University

Abstract Contemporary virtue ethics, an agent-centred ethical theory, has been presented as a response to inadequacies in more traditional act-centred theories. In this paper, I argue that such a response is insufficient: contemporary virtue ethics fails to avoid the inadequacies that it purports to avoid, and brings with it problems of its own. This paper is divided into 5 sections, in the first of which I introduce contemporary virtue ethics as an agent-centred and pluralistic ethical theory. In section 2, I present inadequacies that virtue ethics claims to avoid: being too reductive, too algorithmic, too abstract, self-effacing, and self-other asymmetric. In section 3, I consider and analyse virtue ethics' account of right action and of motives in order to argue in section 4 that, if these inadequacies are indeed problems affecting traditional ethical theories, virtue ethics does not avoid these problems either—particularly because of its basis in the concept of virtues and its heavy reliance on *phronesis*. I show that another ethical theory, limited moral pluralism, has the same advantages of not being overly reductive, algorithmic, or abstract, and being self-other symmetric, and that virtue ethics does not avoid self-effacement as it claims to. I also question here whether self-effacement and self-other asymmetry should be considered problems when evaluating moral theories. Finally, I suggest in section 5 that virtue ethics is open to further criticisms of indeterminacy and lack of explanatory power.

1 Introduction

Contemporary virtue ethics has been presented as a response to inadequacies in more traditional theories. Virtue ethics claims that an action *A*, performed in certain circumstances, is *obligatory* if and only if *A* is an action that a virtuous person, acting in

[*] Norah Woodcock completed her undergraduate studies at McGill University with an honours BA in philosophy and classics. Her research has focused on the intersection of ancient biology and metaphysics, and she is also interested in moral philosophy and feminist theory. She will be beginning her doctorate in classical philosophy at Princeton in the fall.

character, would not fail to perform in the circumstances in question (Timmons 2013, 280). Likewise, a wrong action is one the virtuous person would not do, and an optional action is one the virtuous person might do. Virtue ethics thus defines right action in terms of character, so it is agent-centred rather than act-centred. It is also a pluralistic moral theory rather than a monistic one: it posits more than one factor of intrinsic moral relevance that explains the rightness or wrongness of an action, where these factors are irreducible to any underlying principle.

In this paper, I argue that virtue ethics fails to avoid the inadequacies of traditional act-centred ethical theories, and brings with it problems of its own. To do so, I first present the main problems afflicting traditional act-centred ethical theories, which virtue ethicists claim their agent-centred approach avoids: being too reductive, algorithmic, and abstract; self-effacement; and self-other asymmetry (section 2). Having considered in further detail virtue ethics' account of right action and of motives (section 3), I argue that virtue ethics is *not* more promising than traditional theories. First, it cannot claim advantages over all other theories (section 4). Limited moral pluralism is also not too reductive, algorithmic, or abstract, and can be self-other symmetric, while virtue ethics is also subject to the same problem of self-effacement. Moreover, self-other asymmetry, and possibly self-effacement, do not have to be problematic for ethical theories. Second, virtue ethics is open to the additional criticisms of its indeterminacy and lack of explanatory power because of its basis in the concept of virtues and its heavy reliance on *phronesis* (section 5).

2 Problems with traditional act-centred ethical theories: motivating virtue ethics

The first reason why someone might turn away from contemporary non-virtue-based ethical theories is dissatisfaction with the attempt to make moral judgments by applying abstract principles to particular concrete cases (289–90). Hursthouse (1999) refers to this project as codifiability, and says that many ethicists have since dismissed the idea that ethics can be '*as* codifiable as used to be commonly supposed' because of the 'gap between the abstract principles and the complex particularity of concrete moral situations' (39–41). First, as a pluralistic theory, virtue ethics avoids the criticism of being too reductive to account for our complex moral lives, and thus seems to have an advantage over certain forms of consequentialism (Timmons 2013, 290). It also seems to have advantages over most deontological theories (excluding Rossian limited moral pluralism), as it avoids being too algorithmic, instead giving an essential role to moral judgment (moral wisdom, *phronesis*) (Hursthouse 1999; Timmons 2013). Finally, despite appealing to abstract principles making reference to what an ideal virtuous agent would do, it makes its principles more concrete by specifying particular virtues and is therefore not as abstract as many ethical theories (again, with the exception of limited moral pluralism) (Timmons 2013, 290–91). Virtue ethicists are also opposed to the usual focus on deontic categories of actions, believing that these are the wrong terms and objects of evaluation to be emphasising in moral theory. For instance, Stocker (1976) argues that, by focusing on abstract principles such as duty, obligation, and rightness, contemporary ethical theories limit their scope to 'a dry and minimal part' of ethics, and so fail as ethical theories by ignoring the inner realm of motives and how these relate to values (455). Virtue ethics concentrates on that inner realm, using people (their character traits and dispositions) as objects of assessment. In so doing, it uses aretaic terms (virtue- or vice-based terms) as the terms of assessment, rather than deontic ones. Virtue ethics is thus able to be less abstract and algorithmic than most other pluralist theories.

Traditional moral theories can also be criticised for being externality-ridden, as they do not examine our inner lives as virtue ethics does. Stocker claims that, since traditional theories are 'externality-ridden', they do not recognise '*people*-as-valuable' (460). For Stocker, ethical theories that do not incorporate our motives are undesirable, because a person who adopts their values and principles as her own will either lack important phenomena in her life such as genuine love and friendship, or will suffer from 'moral schizophrenia' (455). On the one hand, if a person adopts the values of a traditional contemporary ethical theory as her own values and is motivated by these values (so that there is harmony between her values and motives), her motives will preclude genuine relationships like love and friendship (455). This is because Stocker thinks that in such relationships, the other person must be valued or loved for her own sake, as an end in herself (456–61). If love is motivated by values such as rightness, duty, or obligation, or even by love itself or happiness derived from love, then the beloved is ultimately loved for the sake of values, which precludes genuine love (456–57, 461). On the other hand, if a person adopts the values of a traditional ethical theory as her own values and is *not* motivated by them, then she suffers from moral schizophrenia: a lack of harmony in her moral life that comes from not being moved by what she values, and not actually valuing the values by which she is moved (453–54). Therefore, according to Stocker's argument, traditional moral theories either preclude harmony between values and motives or preclude genuine relationships, both of which are necessary for a good life (455). Keller (2007) calls this problem self-effacement: a self-effacing theory is one that seems to require that what makes actions right (the values) is not what agents should be motivated by (the motives) (221). Virtue ethics seems to avoid this dilemma, because it explains right action 'in terms of the virtues, and hence of motives', so a virtue ethicist's values should be in harmony with her motives (224).

Another principal problem of traditional moral theories for Slote (Slote 1997) is their self-other asymmetry, which also stems from their mistaken focus on deontic evaluations of actions rather than on our inner lives (175). Judgments we make about the deontic or specifically moral category of actions change depending on whether we are referring to others or to ourselves (for instance, saying that it is obligatory to benefit others but not to benefit ourselves, or that it is morally better to benefit others than to benefit ourselves). Slote, however, points out that this self-other asymmetry is inconsistent with the partiality of our common sense morality as well as many deontological theories, because they judge that we have more obligations towards close ones than towards strangers (1997, 181–82, 185–86). Furthermore, self-other asymmetry is problematic for Slote as it devalues and degrades moral agents, treating the agent's 'pursuit of her own well-being as lacking the [...] positive moral value one assigns to her pursuit of others' happiness' (185–87). Consequentialism avoids this problem because it is impartial, but this impartiality makes it unfairly demanding and leads to an agent's interests being overwhelmed by those of others, similarly degrading or devaluing the moral agent (188-90). Slote suggests that a virtue-based ethical theory can remedy this problem, since it will not be based on fundamental deontic or specifically moral concepts, but on aretaic concepts instead (181, 186–88). In particular, he proposes a common-sense virtue ethics based on ordinary thinking about what is admirable and counts as a virtue, which would, for example, allow for finding self-benefiting traits to be admirable but not morally so (186-88).

These criticisms of traditional ethical theories motivate virtue ethics, which is presumed not to be open to them—and if these are indeed problems for the other theories that virtue ethics can avoid, then it has significant points in its favour.

3 Virtue ethics as an alternative to traditional theories: right action and motivation

Virtue ethics characterises right and wrong action in terms of facts about a virtuous agent: the right act is what the virtuous agent would do. In its account of right action, then, virtue ethics appeals to the hypothetical choices of an ideal agent who possesses the virtues (relatively fixed character traits or dispositions, deemed aretaically good or admirable) and does not possess the vices (Timmons 2013, 270–71, 279–80). Virtue ethics thus relies fundamentally on aretaic concepts and defines rightness only in relation to them, if deontic concepts are used at all (Oakley 1996; Timmons 2013). The basis for saying that an act is morally good or right is the aretaic classification of a character trait as a virtue, so facts about virtues and virtuous agents are more basic than facts about right action (Timmons 2013, 278). For instance, if a virtuous agent would perform an act (in the circumstances), that means that it is aretaically good, which implies that choosing that act would be a morally good decision and its performance would be right. Deontic concepts are therefore fundamental neither to action assessment nor to action guidance, since they are derived from aretaic evaluations (Slote 1997, 2000).

Another key feature of virtue-ethical theories is that they must give an account of the virtuous agent to whom it appeals, by specifying the virtues and explaining how the virtues are determined—that is, by giving content to its theory of value (or theory of the good) (Timmons 2013, 279–80). There are two general approaches for doing so. One is an Aristotelian approach wherein virtues are grounded in an intrinsically good, fundamental concept such as *eudaimonia,* so that virtues are determined by which traits further and constitute a concept like flourishing. The other is a non-Aristotelian approach that takes virtues to be themselves intrinsically valuable, such as Slote's common-sense virtue ethics, which derives virtues from common-sense views about which traits are admirable (Oakley 1996; Timmons 2013).

Virtue-ethical theories can also vary on what their account of right action says about an agent's inner life—her motives, traits, and dispositions. The virtue-ethical criterion of right action can be articulated as follows: an act A is right iff it is the act that a virtuous agent V, acting characteristically, would perform under the given circumstances C. This criterion can be interpreted in multiple ways. *Action-centric* accounts claim that a person who performs A (in C) performs the right action if A is what V would have done, regardless of her own motives and dispositions at the time (Oakley 1996, 135–36). Here, all the emphasis is placed on the act: a person performing A with disharmony between her motives and values would still be doing the right act. *Action-and motive-centric* accounts strengthen this criterion by claiming that a person who performs A (in C) performs the right action iff A is what V would have done and she has the same virtuous motives and dispositions as V would have in performing the action. Oakley argues that virtue ethics must be understood in this more demanding way: 'acting out of the appropriate motives and dispositions is *necessary* for right action' (136). Acting out of virtuous motives is not however *sufficient* for the *action-and motive-centric* account, because it 'allows for the possibility that an action done out of good motives ... may fail to reach the appropriate standard of excellence which one is normatively disposed to uphold' (138). Here, the nature of the act performed still matters for right action. Such is not the case for *motive-centric accounts,* which claim that a person who performs A (in C) performs the right action iff she has the same virtuous motives and dispositions that V does in performing A. Acting out of the appropriate motives is here *sufficient* for right action; the act itself is not considered, only the motive behind it. Motive-centric accounts are often used in agent-based theories, where moral judgments of acts come *only* from evaluations of traits and motives (Slote 1997, 209). No matter how virtue ethics qualifies its account of right action, the virtuous agent is presented as an ideal to be emulated, intending that we 'seek to *be* virtuous agents' (Keller 2007, 224).

4 Virtue ethics versus more traditional theories: is it really preferable?

One of the apparent advantages of virtue ethics discussed earlier was that it seemed to avoid the problem of self-effacement. Keller (2007) argues that virtue ethics is actually subject to this criticism as well. According to Keller, since the right act is that which a virtuous agent would do under the given circumstances, in doing that act one would be motivated by a thought like 'a fully virtuous person would [do this] ... And I want to do what the virtuous person would do' (Keller 2007, 226). For example, say that in aiming to be morally good, I perform an act that expresses the virtue of generosity (as it is what the virtuous agent, being generous, would do). My motivation to perform this act lies in the fact that it is what a virtuous agent would do, meaning I am not being moved by my own generosity (which would look something like, 'this person needs help, and this act would help them').

Such a criticism, however, applies only to virtue ethics as understood by the action-centric account (228). On the *action-and motive-centric* account of right action, virtue ethics is not self-effacing, as what it values (what rightness is based on) *does* include the person's motives: in acting how the virtuous agent would act, and so performing the act that expresses virtue, a person is motivated by the virtuous agent's motives (the virtues) (228). By an *action-and motive-centric* account, when someone is motivated from the X reasons by which the virtuous person would be motivated, then she is 'motivated as the [virtuous] person would be motivated' without being required 'to have any explicit thoughts of the virtue itself, or of the fully virtuous person' (228).[1] A virtuous person, possessing the virtue of generosity, need not think about what a virtuous agent would do in her circumstances; they would simply act out of generosity, thinking something like 'this person needs help, and this act would help them'. As Oakley says, possessing a virtue 'requires internalising a certain normative standard of excellence … a virtuous agent will have certain … normative dispositions, which need not always be consciously formulated or applied, but which will govern and shape their motivations and actions' (1996, 137). As such, the *action-and motive-centric* account of right action can avoid self-effacement. But this approach is also available to non-virtue ethical theories. For instance, a consequentialist theory could adopt a similar account of right action that includes motives: it could say that a person who performs a generous action, and thereby produces the best consequences, acts rightly if they are motivated by generosity, by their desire to help someone, which produces the best consequences (Keller 2007, 230). Since such a strategy for avoiding disharmony between actions and motives is available to any ethical theory, virtue ethics has no advantage over theories in this respect.

Another response to the objection of self-effacement that virtue ethicists could appeal to would be to argue that self-effacement is not a problem ethical theories need

[1] What I am referring to as an *action-centric* account here corresponds with a *de dicto* reading of virtue ethics, while what I am referring to as an *action-and-motive-centric* account corresponds with a *de re* reading (Keller 2007, 228). Drawing on Bernard Williams for this distinction, Keller explains that reading 'what the virtuous person would do' *de re* means that we understand the virtuous person's actions to include their motives, i.e., the virtuous dispositions that motivate their actions (whereas a *de dicto* reading would allow for the same action to be right when it is not motivated by these virtuous dispositions) (228).

to avoid. One appealing feature of virtue ethics is the Aristotelian idea that 'one who is learning to be virtuous may find it useful to have the explicit motive of emulating the virtuous person' (227). If someone's values are not in complete harmony with her motives, because she is a virtue ethicist motivated by the idea that 'such an act is what a virtuous agent would do', we do not need to see this as a problem for the theory—such people are working towards true virtue and moral harmony, and indeed the theory will not be self-effacing for virtuous people (227). The *action-centric* account of right action thus seems preferable: it allows an action which the virtuous agent would do, but performed out of motives the virtuous agent would not have, to be right—as it is the *same action* the virtuous agent would do. Nevertheless, virtue-ethical theories employing this account can still articulate through aretaic evaluation a difference between the inner states of the non-virtuous agent and the hypothetical virtuous one, as any account of right action is only derivative from the main focus of the virtue-ethical theory, the aretaic assessment of character. This approach also allows a vicious person to do the right action out of deplorable motives, rather than not distinguishing between the badness of the motives and the goodness of the action (it can recognise that such a case is different, as to outcomes but not as to virtues or vices involved, from when a vicious person does the *wrong* action out of deplorable motives).

One consequence of this kind of response, however, is that if self-effacement is not a problem for virtue ethics, then it is not a problem for other ethical theories either. Proponents of these theories can also say that being moved by thoughts such as 'this act will produce the best consequences' or 'this act will respect others as ends in themselves' is an acceptable way for people to think about their motivations, while learning how to fully and harmoniously embody what they value. This all shows that virtue ethics does not have the advantage of not being self-effacing while other theories are, and if in fact self-effacement is not a problem for virtue ethics, then it need not be a problem for other theories.

One main objection to virtue ethics is that it is indeterminate; that is, because it does not provide an algorithm for moral decision-making (which above was given as an advantage), it cannot 'yield real guidance' (Timmons 2013, 292) . But, as we have seen, virtue ethics can provide both action assessment and guidance: A is the right act because it is what a virtuous agent would do, and deciding to do A is the 'morally correct decision' because it is what a virtuous agent would decide to do (Hursthouse 1999, 51). Hursthouse argues that each virtue generates a prescription and each vice a prohibition (for example, honesty generates the rule 'be honest' and dishonesty generates the rule 'do not be dishonest'); she calls these rules, derived from our account of virtues, 'v-rules' (29, 37–39). Moreover, whichever way a virtue-ethical theory explains the virtues, it must give an important role to moral wisdom (*phronesis*) for 'interpret[ing] the rules and … determin[ing] *which* rule' should be applied (41). We are expected to have some moral wisdom for identifying which traits are virtues and thereby generating our list of v-rules, and a significant amount of moral wisdom will be needed for applying them to particularly difficult situations (Hursthouse 1999). So virtue ethics does provide some kind of guidance here, in that a person with enough moral wisdom will be able to intuit what a virtuous person, who is morally wise, would do.

The claim that morality is not codifiable enough for there to be any kind of mechanical, general procedure for applying the v-rules can be taken as an advantage of the theory (as discussed above), because it recognises the 'complexity of moral phenomena' and so is not too algorithmic, reductive, or abstract (Hursthouse 1999; Rawls 2009; Timmons 2013). If that is true, then virtue ethics is not alone in recognising the complex texture of moral life in this way, as non-virtue-based theories such as limited moral pluralism may do so as well (Hursthouse 1999; Timmons 2013). Virtue ethics' account of right action is also still open to the objection of indeterminacy, since it relies so heavily on moral wisdom. A virtue ethicist could argue that this is simply a necessary feature of an ethical theory and that 'we cannot plausibly expect more determinacy from the principles of a plausible moral theory' (Timmons 2013, 292).

But while all ethical theories rely on some measure of intuitive moral evaluations to some extent, the extensive reliance on *phronesis* and consequent indeterminacy that we see in virtue ethics (and in limited moral pluralism) poses a serious methodological problem (Hursthouse 1999, 33). This problem comes out in Rawls' (2009) criticism of intuitionism, where he claims that our moral intuitions are 'influenced by our own situation' and 'strongly colored by custom and current expectations', and that intuitionist theories provide no criteria, other than cultural mores, for morally evaluating these (35–37). As is suggested by Aristotle's concept of the vicious person's ignorance of the universal (of what is good and bad), vicious people can believe that there is nothing bad about actions expressing vices (*Nicomachean Ethics* 1110b25-30, 1150b30-37). So although 'there is nothing necessarily irrational in the appeal to intuition', it is necessary that we try 'to reduce direct appeal to our considered judgments' so as to reduce the threat of moral relativism (41). Otherwise, as Heathwood (2007, 798) observes, our ethical theory 'leaves bigots and zealots on their own to intuit their preferred answers'. So virtue ethics' indeterminacy is actually one of its *disadvantages*, albeit one that limited moral pluralism has as well.[2]

Finally, the last claimed advantage of virtue ethics discussed earlier is that it can be self-other symmetric, while more traditional theories cannot be. According to Slote (1997), common-sense virtue ethics is self-other *symmetric*: non-moral virtues (traits outside of 'the sphere of morality' given aretaic value by common thinking) are included in common-sense virtue ethics in a symmetrical and balanced relation with moral virtues. However, given that some of the traits we find admirable are moral and some are non-moral (such as intellectual virtues), why can our ethical theory not say that some traits are morally relevant and some are not (Timmons 2013)? In distinguishing between moral and non-moral virtues, Slote (1997) refers to the former

[2] This disadvantage could be accepted as unfortunately necessary due to the facts of moral reality, though, if virtue ethics were otherwise superior to other ethical theories.

as 'other-benefiting' and the latter as 'self-benefiting'; other-benefiting virtues are *ethically* relevant, whereas non-moral, self-benefiting virtues are *aretaically* relevant, but need not be considered within the sphere of morality. For instance, though we may aretaically admire a person who promotes her own self-interest, a person who sacrifices herself for someone else instead is more morally admirable. We do not need a virtue-ethical theory to recognize such a distinction—a consequentialist can find a person's commitment to a project aretaically admirable while recognising that she acted wrongly in pursuing it rather than sacrificing her life. So, while non-moral, self-benefiting virtues are undoubtedly an important part of our lives, and ethical theories would do well to incorporate them into their accounts for more nuance, this is not a reason to favour virtue ethics in particular over other ethical theories. Ethical theories that '*permit* us to seek our own well-being (within moral limits) ... as a mere *concession* to agents' well-being' may just be reflecting how some important areas of our lives are non-moral (187).

Moreover, Slote proposes a self-other symmetric theory because he believes that the self-other asymmetry of traditional ethical theories downgrades the moral agent. This criticism is a variation of the over-demandingness criticism used against consequentialism, applied to all non-virtue-ethical theories. But not all non-virtue-ethical theories are that demanding; for example, the limited moral pluralism of W. D. Ross (2002) includes a *prima facie* duty to improve our own condition in respect of virtue or of intelligence alongside that of beneficence, and other deontological theories could have the option to include such rules. Since limited moral pluralism can take self-benefiting seriously (as there are no absolute constraints to always outweigh it), if I object that it is still too demanding, it will seem like I just do not want to be concerned with benefiting others. Likewise, the agent's self-concern seems to be valued *too* highly in Slote's (Slote 1997, 193–94) suggestion that common-sense virtue ethics consider other people 'as a class or category', rather than one-on-one. Despite allowing for more balance between concern for ourselves and for others so as not to downgrade the agent, thinking of others as a class rather than as individuals downgrades *other agents*, whose individual interests also matter. While I am not claiming that ethical theories should be impartial, this is another area where we should question our intuitions; perhaps we should instead follow the conflicting intuition that Slote (1997) mentions, in which our common-sense *moral* thought treats permission to self-benefit as a *concession* to agents' well-being. Self-other asymmetry is thus not necessarily a problem, and since other ethical theories can also recognise non-moral virtues, virtue ethics cannot claim self-other symmetry as an advantage.

5 Additional disadvantages of virtue ethics as a moral theory

Another serious weakness of virtue ethics is its lack of explanatory power: what is it about virtues that make actions that express them right? Since a character trait's aretaic goodness (which makes it a virtue) 'bestows upon the action flowing from it the property of rightness,' virtue ethics needs to explain 'why this character trait is good' (Timmons 2013, 295). By the Aristotelian approach for grounding virtues, 'the goodness of certain character traits' is explained 'in terms of their contribution to human *eudaimonia* or flourishing' (Timmons 2013, 296). But if a trait's goodness is explained by its contribution to *eudaimonia*, and that goodness is what makes the act that flows from it right, then why not 'explain the rightness of an action ... directly in terms of its contribution to human flourishing?' (296). Moreover, virtue ethics is considered to be a pluralist theory, with the virtues being 'irreducibly plural intrinsic goods', but if traits are considered good because they promote *eudaimonia*, it seems that the virtues are reducible to the single, more fundamental intrinsic good of *eudaimonia* (from which the virtues are then derived) (Oakley 1996, 140).

By this account, then, virtue ethics looks like a form of monism. The account of what makes an action right would be that it promotes *eudaimonia*, with the virtues as intermediary stages that are instrumental to that promotion, and it would no longer be clear that the theory, which is supposed to be virtue-based, needs to include virtues at all in its account of right action. If this is the case, then we do not need, as Timmons (Timmons 2013, 296) says, 'to first explain the goodness of traits in terms of flourishing and then explain the rightness of action in terms of the goodness of traits; we can explain both the goodness of traits and the rightness of action directly in terms of flourishing'. Therefore, even without delving into the problems of defining a concept like *eudaimonia* and avoiding circularity in doing so, the Aristotelian approach cannot provide an acceptable—truly virtue-*based*—explanatory account of right action for virtue ethics (Slote 1997, 207; Timmons 2013, 295–96). Doing so would 'deprive virtue ethics of its distinctive character' (Timmons 2013, 209).

Alternatively, for virtue-ethical theories that take the non-Aristotelian approach, the goodness of character traits is 'an unexplained brute fact': 'certain character traits just are intrinsically good and ... their goodness need not be further explained' (296). Either the virtues are simply 'grasp[ed] through intuition' as 'self-evident truths', or we can reasonably suppose that a trait is a virtue when this claim is 'supported by the body of our considered moral beliefs'; that is, by internal support (296). While claims about *which* traits are virtues do have internal support, this explanation for *why* these traits are good is unsatisfactory. As Timmons notes, there are ways of plausibly explaining why a trait like benevolence is good: since we can come up with explanations for why a particular virtue is good (for example, benevolence moves one to help people in need, showing respect for them as persons, and reduces suffering), it is unconvincing to claim that their goodness is a brute fact. Only relying on intuition is undesirable for the reasons discussed above, and here it is called into question by 'the sort of constructive criteria that are said not to exist' (Rawls 2009, 39). Virtue ethics therefore cannot give an adequate explanation of why acts are right or wrong.

6 Conclusion

I have argued that contemporary virtue ethics, as a response to perceived problems in more traditional ethical theories, does not in fact avoid such problems itself, and brings with it additional issues. Proponents of virtue ethics claim that deontological or consequentialist theories suffer from being too reductive, algorithmic, abstract, self-effacing, and self-other asymmetric, and that virtue ethics can avoid these problems and should therefore be preferred over the more act-centered ethical theories. These reasons, however, do not hold as satisfactory advantages for virtue ethics. If virtue ethics avoids being too reductive, algorithmic, or abstract, then Rossian limited moral pluralism does too, so these are not reasons to choose virtue ethics over other contemporary ethical theories. With respect to the other apparent advantages, either consequentialists or deontologists can make use of the same tools as virtue ethicists to avoid the problems of self-effacement and self-other asymmetry, such as including a person's motives in their account of right action or recognising self-benefitting as aretaically valuable, or these need not be considered problematic for any ethical theory.

Moreover, in trying to be less reductive, less algorithmic, and less abstract than other ethical theories, virtue ethics relies too heavily on *phronesis* and intuition, causing it to be too indeterminate as an ethical theory. An additional problem to which virtue ethics is subject, and several more traditional ethical theories are not, is a lack of explanatory power: it is unable to explain what it is about virtues that makes actions expressing them right, without reducing the virtue-ethical theory to monism (and a monistic theory would not be *virtue*-based, but, e.g., *eudamonia*-based), or appealing to intuition and brute fact.

What were presented as advantages for virtue ethics are therefore not advantages for it after all, and so are not convincing reasons to prefer a virtue-ethical approach to a deontological or consequentialist one. Virtue ethics furthermore has additional disadvantages in its indeterminacy and lack of explanatory power. As such, although virtue ethics brings up valuable considerations about the inner lives of agents that could be used to supplement and refine consequentialist or deontological theories, a virtue-ethical approach to moral theory is not more promising than traditional act-centered moral theories.

References

Aristotle. 1999. *Nicomachean Ethics.* 2nd ed. Translated by Terence Irwin. Hackett.

Heathwood, Chris. 2007. Book Note on *Moral Theory: An Introduction,* by Mark Timmons. *Ethics* 117:797–98.

Hursthouse, Rosalind. 1999. *On Virtue Ethics.* Oxford University Press.

Keller, Simon. 2007. 'Virtue Ethics is Self-Effacing'. *Australasian Journal of Philosophy* 85 (2): 221–31.

Oakley, Justin. 1996. 'Varieties of Virtue Ethics'. *Ratio* 9 (2): 128–52.

Rawls, John. 2009. *A Theory of Justice.* Harvard University Press. Revised edition.

Ross, W. D. 2002. 'What Makes Right Acts Right?' In *The Right and the Good,* edited by Philip Stratton-Lake, 16–64. Clarendon.

Slote, Michael. 1997. 'Virtue Ethics'. In *Three Methods of Ethics: A Debate,* edited by Marcia W. Baron, Philip Pettit and Michael Slote. Wiley-Blackwell.

Stocker, Michael. 1976. 'The Schizophrenia of Modern Ethical Theories'. *Journal of Philosophy* 73 (14): 453–66.

Timmons, Mark. 2013. *Moral Theory: An Introduction.* 2nd ed. Rowman & Littlefield.