The True and the Good

BY

David Black

A dissertation submitted to the
School of Graduate Studies
Rutgers, The State University of New Jersey
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
Graduate Program in Philosophy

Written under the direction of
Douglas Husak and Ernest Sosa
and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

October, 2019

ABSTRACT OF THE DISSERTATION

The True and the Good

by David Black

Dissertation Director:

Douglas Husak and Ernest Sosa

Ethics and epistemology are both normative fields. Philosophers talk about what we *should* do, what we have *reason* to believe, whether truth or pleasure is the ultimate *good*. Since they are both normative, we should expect that there are structural and substantive analogies between the two fields. In this dissertation, I develop and explore four of those analogies.

In "Epistemic Punishments", I argue that an aspect of our epistemic practice looks a lot like the informal, social punishments we impose on each other. Moreover, epistemic punishments can be justified in the same way that the more familiar social sanctions are. "Liberal Neutrality and False Beliefs" argues against a core tenet of liberalism. Though the state need not respect disagreement about fact, liberal neutrality says that the state must respect disagreements about value. Because of the similarities between action and belief evaluation, liberal neutralists cannot maintain this asymmetry.

In "Praiseworthiness (and Knowledge) from Falsehood", I present a counterexample to the claim that an action is praiseworthy when it is done for the right reasons. This counterexample runs parallel to an epistemic one. Cases of knowledge from falsehood show that you can know p even if you don't believe p for the right reasons. I conclude that both knowledge and praise require only that your reasons are "good enough".

Finally, "Moral Swamping" poses a moral version of an epistemic problem. The swamping problem challenges us to explain why it's good to form justified beliefs. A true belief, based on good evidence, is no more true than the same belief based on superstition. A false belief, based on good evidence, is no less false than one picked from a hat. I argue that there is a similar puzzle in explaining the value of freedom. An agent who pursues the good freely need not do a better job than one who is compelled. An agent who pursues the bad freely won't necessarily perform better than one in the grips of addiction, manipulation, or irrationality. Why, then, should we care about freedom at all?

Acknowledgements

This dissertation would not be possible without my committee members, Branden Fitelson, Miranda Fricker, Douglas Husak, and Ernest Sosa. Each taught me something invaluable, that I hope to take with me in the future.

Branden taught me that philosophers are just people. They have their own hopes and fears, passions and peeves. Doug taught me that it's ok not to get it right. I won't have all the answers on my first try, and probably not on my second or third or twelfth, either. I can tell a person that I don't understand them, without it being a snarky way of saying that they're wrong. From Miranda, I learned to write about what I think is interesting, where I think I can make a contribution. I don't know where the trends are going to go, and chasing them isn't worth it.

It's almost insulting to say that I "acknowledge" Ernie's help. I do not know how to express the gratitude, respect, and admiration I feel for him. His work is obviously a major inspiration for my project. It was the first place that I encountered the subtle and fascinating analogies between action and belief. He taught me how to find the points of connection between them, and how to build out from there. I don't think I could ever repay the resources, time, and effort he offered to me.

Most importantly, he showed that a good philosopher can be thoughtful, kind, gracious, and humble. Good people can be good philosophers. Since I started grad school, I have seen these qualities in others. But Ernie demonstrates them in every interaction we've had. Ernie is a role model.

This dissertation would be possible, but not worth reading, without Will Fleisher and Jimmy Goodrich. They were always willing to talk with me or read a draft. Their feedback was always helpful. I have benefitted tremendously from working

with them. I hope they have gotten something out of working with me. Jimmy and Will are excellent colleagues and even better friends.

Additionally, I would like to thank the members of Ernie's dissertation group over the years: Bob Beddor, Laura Callahan, Marilie Coetsee, Megan Feeney, Carolina Flores, Danny Forman, Georgi Gardiner, and Chris Willard-Kyle. Special thanks to Lisa Miracchi and Kurt Sylvan, who welcomed and supported me when I was in my first year.

Special thanks also to Mercedes Diaz.

There are many others that deserve my thanks. I have listed some of them here, in alphabetical order. Thank you to Austin Baker, Nick Beckstead, Ray Briggs, Ben Bronner, Liam Bright, Tim Campbell, Sam Carter, Kenny Easwaran, Andy Egan, Simon Goldstein, Veronica Gomez, Alex Guerrero, Mike Hicks, Tyler John, David Anton Johnson, Ben Levinstein, Howard McGary, Andrew Moon, Pamela Robinson, Daniel Rubio, Eli Shupe, Holly Smith, Philip Swenson, Larry Temkin, Nick Tourville, Beth Valentine, Peter van Elswyk, and Steve Woodside.

Dedication

To my mom. You taught me the value of an education. No one has worked as hard as you so that I could have one.

To Arm, my love, for everything.

Table of Contents

List of Figures

Chapter 1
Introduction

Ethics and epistemology are both forms of normative inquiry. Morally, you should help people in need. Epistemically, you should believe that the earth is round. Pleasure is a moral good. Knowledge is an epistemic good.

I am not the first person to notice these parallels.[1] I came to them through learning about virtue epistemology, and the claim that epistemic normativity is just one species of performance normativity.[2] We are encouraged to evaluate beliefs the way we evaluate other performances, like an archer's shot.

This dissertation is motivated by the further thought that moral evaluation is also a kind of performance normativity. Shots can be accurate, adroit, and apt. Beliefs can be true, competent, and amount to knowledge. Actions can be right, virtuous, and praiseworthy. But we don't need to adopt the performance normativitiy approach to see interesting relationships between ethics and epistemology; action-evaluation and belief-evaluation. At times, analogies between the ethical and the epistemic are clear. These are worth developing, to see how far they extend. At other times, the two fields diverge. This is no less interesting. Just as we want to know how beliefs are like actions, we want to know how they are different.

The four papers in this dissertation are not a defense of performance normativity,

---

[1]For a good survey in epistemology, see Riggs (2008). For a good example in ethics (or philosophy of law), see Duff (2007, ch. 11)

[2]See Sosa (2007, 2009, 2011). For dissent, Chrisman (2012)

virtue epistemology, or the the existence of analogies between action- and belief-evaluation. I will not offer a performance normative account of ethical evaluation.[3] Nor will I take any of these claims for granted, and apply them in order to solve problems. Each of the papers is inspired by an analogy between ethics and epistemology. In writing them, I have been careful not to assume that epistemology and ethics are the same in structure or substance.

The claim of the dissertation as a whole is modest. When you are doing normative work, it's interesting and worthwhile to think about other normative fields. When faced with an epistemological problem, see what resources the ethicists have developed to handle an analogous one. When giving an account of an ethical property, figure out what property plays the same role in epistemic evaluation and how epistemologists account for it.

My conclusion is offered in the spirit of advice to other researchers. I'm not sure how to convince someone that a piece of advice is good or that something is worth doing. It would be inhuman to try to argue that a piece of advice is correct. We know good advice by its fruits. I will demonstrate the benefits of moving between the moral and the epistemic worlds by doing good philosophy, with the analogies between them always in the background.

In this introductory chapter, I will explain each paper and the analogy between ethics and epistemology that motivates it.

One branch of social epistemology addresses the social systems and institutions surrounding our epistemic practices. We want to know how well existing social institutions promote the epistemic good, and whether they are just. Chapter 2, "Epistemic Punishments", and Chapter 3, "Liberal Neutrality and False Beliefs", discuss the informal social and the formal political apparatus that we have developed for dealing with epistemic concerns.

Chapter 2, "Epistemic Punishments", argues that there is an informal institution

[3]Though see Mantel (2013).

of punishing people who promote the epistemic bad of false belief. I argue that epistemic punishments are as just, morally and epistemic, as other informal punitive practices.

Here is how the parallels between belief and action fit in: In our daily lives, as socially-situated agents, we want to promote true belief. What tools do we have to do this? Are they just? This is important because people with false beliefs are likely to make themselves and others worse off, epistemically and practically.So we reflect on the informal institutions we use to improve people's actions. One method is by punishing them, in more or less formalized ways. This line of thought raises the issues that "Epistemic Punishments" addresses.

Socially, we treat each other in punitive ways when faced with a culpable wrongdoer. Culpability is a matter of showing sufficient disregard for important moral values. In informal settings, punishment often means being marginalized in a community, losing status and some benefits of membership in the community. When marginalization is used to punish, there is usually a social stigma attached to it. For example, you might lose status in a tabletop gaming group because they moved the meetings to Wednesday, and you can't make Wednesdays. But socially and morally, that's very different from being kicked out of the group because you lost your temper and keyed somebody's car.

I argue that we manage our social lives as epistemic agents in a similar way. In effect, we punish people who show sufficient disregard for the epistemic good by intentionally, knowingly, recklessly, or negligently spreading falsehoods. As potential informants, people who culpably spread falsehoods are marginalized from our community of inquirers. We do not place as much trust in them or give them as much space to contribute to shared inquiry. When this marginalization is used punitively, it is often accompanied by stigmatization: "crackpot","conspiracy theorist", and even "liar" are loaded, stigmatizing terms. Socially and morally, this is different from being left out of shared inquiry for more innocent reasons.

I argue that the epistemic punishments need not constitute epistemic injustice, when they are deserved. It would be epistemically unjust to marginalize someone as an inquirer because of their race, but it is not epistemically unjust to marginalize a potential informant for a history of deception. This is the same mechanism whereby it is unjust to marginalize an agent because of their race, but not because of their track record of wrong-doing.

In Chapter 3, "Liberal Neutrality and False Beliefs", I argue against a traditional liberal tenet, neutralism. According to neutralism, there are some issues that the state should remain neutral about as a matter of principle. Even if it would do more good than bad, neutralists think there is still something objectionable about an official state religion. Oddly, though they are fine with teaching that the earth is round in public schools, they do not want schools to teach that gods do not exist. Neutralists cannot justify treating belief in gods and belief in a flat earth so differently.

In that chapter, I pose a challenge for neutralists. Setting aside forward-looking considerations, explain the politically relevant difference between religious controversy and the controversy over the shape of the earth. I suggest that the neutralist needs to say something along the lines of: Even when a person's religious or moral commitments are wrong, their decision to pursue them is worth respecting. They've got a legitimate interest in worshipping the god that they believe in, and they should be left free to do so. On the other hand, the decision to pursue the life of a flat earther, climate change denier, anti-vaxxer, or conspiracy theorist is not worth respecting.

If the neutralist is going to make good on this reply, they need to explain why flat earth beliefs are not worth respecting. Their best bet for drawing the line between the respect-worthy and the disrespectable is to appeal to the notion of reasonsableness. Reasonable commitments, like belief in a god, are worth our respect. Unreasonable commitments, like belief in the flat earth, need not be respected.

The neutralist is left with a dilemma. If the standards for reasonableness are low, then belief in the flat earth is reasonable. The liberal state cannot teach an adequate

science curriculum. If the standards are high, then belief in a god is unreasonable. It is within the bounds of liberal legitimacy to teach atheism in school.

The upshot for social epistemology is easier to see if we run the argument of chapter 3 backwards. I argue that, since liberal neutrality conflicts with teaching that the earth is round, liberal neutrality cannot be a requirement of justice. A committed neutralist could instead take my argument to show that the public school system is unjust, since it conflicts with neutrality. My argument can be turned around to condemn most of our political apparatus for promoting the epistemic good over the epistemic bad.

The analogy between belief and action is clear. We consider belief as just one more type of action or performance that the state might have an interest in improving. We expect liberal states to tolerate a broad range of bad actions. People should be left to live their lives as they want, even when we could do some good by manipulating them or disrespecting their choices. On the other hand, when we consider the varied and colorful world of conspiracy, we expect states to tolerate a rather narrow range of bad beliefs. Why is this one type of bad action, false belief, treated so differently from other bad actions? Chapter 3 grew out of my attempt to answer this question.

The second half of the dissertation turns from social epistemology to normative ethics. Chapter 4, "Praiseworthiness (and Knowledge) from Falsehood", argues against a popular account of praiseworthy action and suggests an alternative. Chapter 5, "Moral Swamping", poses a challenge: explain why it matters whether an action was performed freely.

Chapter 4, "Praiseworthiness (and Knowledge) from Falsehood", picks up with Markovits (2010)'s coincident reasons thesis. According to the CRT, an action is praiseworthy just in case the motivating reasons that the agent performed it coincide with the moral reasons there were to perform it. This spells out the idea that it's important to do the right thing for the right reasons. For example, Kant's grocer gives correct change, but only because they know they will go out of business if they

don't. Their action lacks moral worth because their motivating reasons are selfish. A grocer who gives correct change in order to treat their customers fairly acts with moral worth. Their reason, "Giving correct change is fair", is a good moral reason to give correct change.

The CRT, together with the orthodox view that moral reasons are facts, implies that praiseworthy action cannot be motivated by falsehood. If your action is praiseworthy, you must be motivated by moral reasons. And moral reasons must be facts. A similar line of thought has held sway in epistemology. Counter-closure[4] says that the premises in an inference leading to knowledge must be facts. One way to argue for this is to say: If you gain inferential knowledge, your inference must be based on good evidence. And good evidence consists of facts.

There are cases where agents seem to gain knowledge by inferring from falsehood. I present analogous cases that support the idea of praiseworthy action from falsehood. Agents who make small mistakes can still act with moral worth, even if they aren't moved by the reasons that make their actions right.

Most of the paper discusses attempts to accomodate the counterexamples on behalf of CRT and counter-closure. I argue that these either fail or are more trouble than they're worth. The take-away is that epistemic and moral agents don't need to track their reasons perfectly. Knowing is compatible with doing a "good enough" job of tracking the evidence. Praiseworthy action requires you to track the reasons "closely enough" to exhibit virtue, but no more than that.

Chapter 5, "Moral Swamping" presses a version of the swamping problem against autonomous action. Considering how freedom is supposed to work, it is hard to explain why it matters so much morally. The original swamping problem[5] was targeted at reliabilism about epistemic justification.

---

[4]Contemporary interest in this issue seems to have started with Warfield (2005).

[5]The contemporary discussion begins with Zagzebski (1996), although it has also been tied to the problem from Plato's *Meno*.

Reliabilists say that justified belief is belief produced by a process that reliably-enough produces true beliefs. This makes justified belief out to be similar to a cup of coffee produced by a reliable coffee machine. If the coffee is tasty, it is no tastier for having been produced reliably. If the coffee is bad, it is no better for having been produced reliably. Yet we treat justified true beliefs as if they are more valuable than unjustified true beliefs. Justified falsehoods are supposed to be better than unjustified falsehoods. Reliabilism looks unequipped to explain why this is.

It has since been recognized[6] that this is a problem for any theory of justification that makes it out to be instrumental to securing the truth. I argue that the best existing accounts of autonomy make it look instrumental to securing another good. Part of what makes an action free, autonomous, or voluntary is that it was made by an agent who was sufficiently competent to decide well.

Ethicists face the same three broad choices for responding to the swamping problem that epistemologists do. First, they could offer a deflationary explanation of autonomy's value. We care that people choose freely, but only because this means they are likely to choose better. Second, they could go primitivist. If an agent autonomously harms themselves, anti-paternalism means we cannot intervene. But if an agent harms themselves while drunk or otherwise impaired, we can stop them. This is a primitive fact of morality, and there is no explanation for why we should treat these cases differently.

Last, ethicists could try to derive the value of autonomy from other values. Epistemologists have made similar attempts, trying to derive the value of justified belief from the value of true belief. This would provide a satisfying solution to the swamping problem, but I argue that current accounts do not give us any obvious way of making progress.

Originally, when I started working on the swamping problem and the counterexamples to counter-closure, I turned to ethics. We want to give an account of when

[6]Pritchard (2010)

a belief can amount to knowledge, even though the agent has flubbed the evidence. So let's think about how an action can be praiseworthy, even though the agent has flubbed the reasons. We want to explain why competent, justified beliefs are epistemically important. So let's think about why capacitated, autonomous decisions are morally important.

Unfortunately, things didn't work out as I hoped. Rather than finding moral solutions that could be ported over to solve epistemic problems, I found a new batch of moral problems. That's close enough to progress in philosophy for me to feel good about.

Each of these papers has its own point to make. Chapter 2 claims that we have a practice which can help stop the spread of falsehood, and this can be done in an epistemically just way. Chapter 3 argues that states can decide controversial matters of value, just as they do with facts. Chapter 4 shows that the leading theory of praiseworthy action needs to be modified. Chapter 5 poses an explanatory challenge for moral theories.

Additionally, there is a reason why I chose to write these four papers, and not four different ones. Taken together, these papers tell a story of how my views have developed over the years. Even at my most ambitious, I couldn't claim that this dissertation constitutes an argument for consequentialism. There are too many gaps and choice points left open. Still, I hope this little bit of intellectual autobiography helps make this dissertation more cohesive and gives a sense of narrative to the chapters.

I've always been attracted by the straightforward logic of simple consequentialism: More good is better, more bad is worse, and that's about all there is to it. For whatever reason, there are two kinds of cases where this line of thought really doesn't appeal to me. First, we evaluate beliefs in a process-centric way. The honorifics that we use for beliefs (e.g., reasonable, justified, knowledge) seem to be less a matter of their content. How the agent came to believe them is more important. Second,

consequentialism seems to me to fail when autonomy is at stake—especially when it comes to paternalism and sexual consent.

These two kinds of cases have made me think that manner matters, and not just outcome. When I started this dissertation, I was especially concerned with epistemic evaluation. In particular, I was interested in the normative properties that attach to justification. Does justification mean that it's epistemically permissible, obligatory, praiseworthy to hold a belief? Two developments in epistemology led me to suspect that justification was a hypological notion—mostly about praise- and blameworthiness—rather than a deontic one.

First, authors in the knowledge-first program (especially Williamson 2000, Littlejohn 2012) argued that our intuitions and theories about false justified belief have more in common with accounts of excusable action than with justified action. Comparison with Duff (2007) and Gardner (2007) confirmed this. Second, a connection between justified true belief and praiseworthy action seemed to me to be central to answering the swamping problem.

If epistemic evaluation mostly helps us to identify praiseworthy and blameworthy beliefs, I felt like I could make sense of it. I'm happy with the main ideas of the consequentialist account of praise and blame. If epistemic evalaution is part of that same toolbox, I can see how it fits into my picture of the world. To test this view, I made some predictions from it. If justification is like an excuse for believing the false, we should be willing to treat unjustified beliefs as if they are unexcused actions.

This prediction is somewhat borne out for small-scale interpersonal interaction. As I argued in "Epistemic Punishments", we are (and should be) willing to treat peers with unjustified beliefs in a way that's similar to how we informally punish culpable wrong-doers. If this ultimately gets cashed out in consequentialist terms, all the better. Moreover, in "Praiseworthiness (and Knowledge) from Falsehood", I argue that both praiseworthy action and knowledge require a similar "close enough"ness

of reasons. If epistemic evaluations track a kind of praiseworthiness, and if praiseworthiness matters in a consequentialist-friendly way, this is exactly what we would expect. That's two predictions for two.

On a political scale, the claim that unjustified beliefs are or should be treated like unexcused actions fares quite miserably. This is partly because there is no one way that states treat unjustified beliefs, and no one way that states treat blameworthy action. The state devotes a lot of resources to preventing some unjustified beliefs (e.g. through public schools). It takes no action against other unjustified (religious or moral) beliefs. Those beliefs aren't the state's business. On the other hand, the state prosecutes some unexcused actions (e.g., by criminalizing them). But they leave others alone. The state is not coming after you for not calling your mom on her birthday, no matter how bad your excuse is. That's not the state's business.

"Liberal Neutrality and False Belief" came out of my attempt to make sense of this state of affairs. Is there a relationship between the beliefs that aren't the state's business and the actions? In the end, I conclude that forward-looking reasons dictate how the state should manage false belief. That was good enough for me.

Meanwhile, as I read the literature on autonomy—especially autonomy under oppressive socialization—I became attracted to a normative competence view.[7] "Moral Swamping" explains a problem that I noticed for views that give a special place to freedom or autonomy. As in the other cases, I think a forward-looking view properly balances the theoretical considerations.

Taken together, these papers form a pattern. There are analogies between ethics and epistemology. A forward-looking or consequentialist theory seems (to me) best able to predict explain how these two areas relate to each other. I have convinced myself that this inference to the best explanation is good enough for me to favor consequentialism. But I don't expect that non-argument to convince anyone else.

---

[7] Wolf (1990), Stoljar (2000)

I came to these four papers by thinking about analogies between ethics and epistemology, and my goal is to illustrate that this is worth doing. If I convince you that each paper was worth writing, I've done what I meant to.

Chapter 2
Epistemic Punishment

As social creatures, we rely on each other to share information and divide the burdens of inquiry. When informants offer truths, things run along smoothly. But false testimony disrupts the flow of information. The audience may take action to prevent or discourage the informant from testifying again in the future. In this paper, I want to explore some of the ethical issues involved in disallowing someone from participating in the collective search for truth.

To start, let's compare the following two cases. Sarah gets on Twitter, and lays out some facts about her economic condition. She proposes an explanation for how things got to be the way they are, and suggests some questions that are raised by her experience, so that we can come to better understand economic issues. She gets dozens of responses calling her a liar and a host of slurs. Going forward, she notices that there's a group which seems dedicated to posting insulting, derailing, and uncooperative responses to her, no matter the subject. Exhausted and upset, Sarah stops posting to Twitter. She avoids any political or economic discussion on social media entirely.

Like Sarah, Lars gets on Twitter and testifies as to his economic conditions. Unlike Sarah, Lars's testimony is false. He fabricates a story of hardship, and proposes a Jewish conspiracy as an explanation. He lies, in order to drum up support for his fringe political position. Though some people are convinced, others see the lie for what it is. They openly call him a liar and encourage others not to trust him or include him in conversations. His reputation destroyed, Lars decides not to post

anymore.

Sarah has been wronged. In Dotson (2011)'s words, she has suffered from testimonial smothering: she decides not to testify, in order to avoid the hardships she can expect. It is also likely that she has been subject to Fricker (2007)'s testimonial injustice. Because she is a woman, her audience has unjustly discredited her testimony. And these are only two of the ways in which Sarah has been mistreated. On the other hand, Lars hasn't been wronged. Though his feelings might be hurt, and he has lost the opportunity to contribute to the conversation, his bad reputation is well-earned. What accounts for the moral difference between these two cases?

There is an obvious difference: misogyny is a bad reason to exclude a person from inquiry, and a bad track record is a good reason. But this just pushes the question back a step. What makes a bad track record a good reason to so exclude someone? I will argue that denying people the opportunity to participate in shared inquiry can be an epistemic punishment. Bad track records give us reasons to exclude liars in the same way that we have reason to punish other kinds of wrong-doers. The central claim in this paper is that we sometimes do and morally may engage in epistemic punishment. This is what explains why Lars hasn't been wronged, epistemically or otherwise.

This immediately raises two questions. First, in what sense are epistemic punishments *punishments*? This issue is addressed in §§1–3. In §1, I explain the sense in which I am using the term "punishment". §2 presents some candidate punishments, and candidate behaviors that merit punishments. There, I also argue that the proposed punishments fulfill two of the conditions for a behavior to count as punitive (viz., they are unwelcome, and inflicted in response to alleged culpable wrong-doing). §3 completes the argument, and shows that the candidates fulfill the last two conditions of punishment.[1]

[1]The third condition is that punitive responses are condemnatory in character. The fourth depends on the particular theory of punishment we are working with, and concerns what punishment

The second question is: in what sense are epistemic punishments *epistemic*? §4 attempts to answer this question, and fills in other details of the view. To give a first pass for now, I say these actions are epistemic punishments because they concern how we treat each other as potential informants and collaborators in inquiry. Navigating these relationships is an important part of our lives as epistemic agents. I think this is good enough to merit the term "epistemic". Moreover, there is a case to be made that the punishments are epistemic in a "deeper" sense.

§5 concludes with some avenues for future work on epistemic punishment. In this paper, I mostly focus on more intimate, interpersonal practices, but there are important and difficult questions about how these extend to impersonal institutional or state-sanctioned practices.

2.1 What is a Punishment?

I'm claiming that there are parts of our epistemic practices that are punitive, and that we have good moral reason to punish in this way, at least in some cases. The first step in arguing for this claim is to settle which actions count as punishments. That is not a question I can hope to settle here. Instead, I will try to remain as neutral as possible between competing theories and point to some widely-recognized features of punishment. This section explains the sense in which striking a liar off your list of potential informants (for instance) can be punitive.

"Punishment" brings to mind official legal proceedings. However, I focus on the informal, interpersonal practices of punishment. Hampton (1992) gives a good example. After a blizzard, Alex spend hours shovelling out her car. While she is gone, Brad, who lives in a neighboring building, steals her parking spot. Brad didn't do anything to get rid of the snow in one of the other spots. When Alex sees this, she partly buries Brad's car under the snow. This is a mild, informal punishment, and

---

is "for". Retributivists, for instance, can disagree with consequentialists about the purpose of punishment.

cases like this are familiar from everyday life. Maybe Alex's action is better called a "sanction" or "hard treatment", or even just "blaming". If "punishment" is reserved for legal (or otherwise official) actions, then I will be arguing for epistemic sanctions or epistemic hard treatment. At any rate, it is the informal phenomenon that I mean to be talking about.

One of the informal ways people punish is by excluding someone from their social circles. For instance, Charlotte insults Damon. Damon stops inviting her to parties, refuses to go to events that he knows Charlotte will be at, and encourages their mutual friends to do the same. This can be punitive. When I say there are epistemic punishments, I am saying that some of what we do, as epistemic agents, is punitive in the same way that shunning someone can be. People rely on each other to share information and learn. The way information is shared within a community—the epistemic facts about that community—matters, morally. If Sarah is considered incompetent to testify by everyone in her community, for prejudicial reasons, she has suffered an epistemic injustice. Lars does not. We collectively decide how people fit in, epistemically, to our communities. One person can be widely regarded as an expert on some topic, and another can have almost no voice. Sometimes, marginalizing someone epistemically is a punishment, as other kinds of shunning are.

I focus on two ways of epistemically punishing someone. First, an individual can refuse to elicit their testimony (on some topic), and encourage others to do the same. This is what happened to Lars. Colorfully, he has been excluded from the community of informants. The second kind of punishment involves excluding someone from the community of inquirers. We don't just exchange information with each other. We work together in inquiry. In shared inquiry, people suggest new lines for investigation, propose hypotheses and explanations, and contribute ideas to the debate or discussion. By denying someone opportunities to do those things, we can punish them. For brevity, I will use the word "marginalization" to refer to the ways that we deny a person the opportunity to contribute to collective inquiry, either by

testimony or otherwise.

Saying that marginalizations are exclusions from the community of informants and inquirers is meant to be suggestive, on analogy with excluding someone in other ways from a community. But I don't mean to rest too much argumentative weight on these phrases. In §§2–3, I'll argue that marginalization can be punitive, directly; without appeal to communities of informants and inquirers.

Of course, I could not claim that marginalization is always an epistemic punishment or is always permissible. It is not clear that Sarah has been punished,[2] and if she is it is certainly not permissible. Even when done for good reasons, excluding someone from a community is not always a punishment. It could be that, after years of conflict, Fernando decides that Evan's continued presence in his life is just not good for him. Fernando refuses to interact with him, at least to some extent, and in some ways. If this is done dispassionately, and in order to "get it over with", so to speak, it may not be a punishment. It seems that part of what distinguishes punitive action is the reason it's performed. Denying people the opportunity to participate in inquiry should be no different.

So, to make the case that some of our epistemic practices are punitive, we need to determine what makes a practice punitive in the first place. There is no once-and-for-all consensus on this point. We're better off looking to the central identifying features of punishment, and going from there. I discuss four.[3]

First, in punishing, the punisher imposes a burden or withholds a benefit from the punished. Alex puts Brad in a position where he has to dig his own car out of the snow. If the rest of the community goes along with Damon, they collectively withhold the benefits of membership in that community from Charlotte. The important thing here is that punishments are typically unwelcome by their recipient.[4]

---

[2]Though I would argue that she has. Her tormenters punish her for the supposed wrong of "testifying while being a woman".

[3]Compare Bedau & Kelly (2017), Walen (2016).

[4]Cases where an especially motivated wrong-doer welcomes and even demands punishment are

Second, punishments are inflicted in response to alleged culpable wrong-doing. Stealing someone's car imposes a cost on them, but this hardly counts as punitive unless you think they've wronged you in some way. It's possible to harm innocent people in all kinds of ways, but you can't *punish* someone while knowing that they haven't done anything. Moreover, it is typically wrong to impose such a burden or withhold a benefit from an innocent person. Alex couldn't, permissibly, bury a stranger's car under snow just because she felt like it.

Third, a punishment condemns the alleged wrong-doer, or their action. Condemnation helps to separate punishment from other unwelcome actions that respond to wrong-doing (Feinberg 1965). This may explain why Evan's is not a case of punishment. When Fernando cuts Evan out, he may be too exhausted by the whole ordeal to properly condemn him. His chief motivation will be getting away from Evan and moving forward. If condemnation is completely off the table, Fernando's treatment of Evan looks less like a punishment.

The last element of punishment is the most contentious. Put briefly, whether an action counts as punitive depends on what it is for. One theorist may say that the purpose of punishment is to discourage people from doing wrong. A second will say that punishments are meant to communicate something to the punished. A third that punishments exact retribution, and a fourth that all of these are important purposes of punishment. More generally, we can identify punitive actions by the role that punishment is supposed to play in a broader social, moral, or legal theory.[5]

Settling the purpose of punishment is related to, but distinct from, justifying punishment. Two people may agree that we should attempt to rehabilitate wrong-doers and convince them to be better people, even in ways that are unwelcome to them. If the first believes this counts as punishment, they will say that punishment,

---

understood to be marginal.

[5]The condemnatory aspect of punishment can also be thought of as part of the role of punishment. Since it is widely agreed that punishments condemn, I treat it separately from the more controversial theories of punishment's purpose.

in the form of rehabilitation, is justified. If the second thinks that there is something in the concept of punishment that makes it retributive, they will say that punishment is unjustified, and only rehabilitation is justified.

Punishment's purpose is also related to its justification in a more subtle way. For example, if we think that punishment is supposed to encourage reconciliation between wrong-doer and victim, straightforward hard treatment will fall short. Ideally, punishers would also explain to wrong-doers how they've done wrong, who they must (attempt to) reconcile with, and why they are being punished. Without these accompanying actions, it's hard to see how just punishing someone could encourage reconciliation. I intend the theory of epistemic punishments to reflect this fact. Punishment may need to be accompanied by other actions (like giving the wrong-doer a chance to apologize or correct the wrong) in order to be justified.

I want to remain neutral between the different proposed purposes of punishment. Similarly, in the interest of neutrality, I will mostly stay silent on which accompanying actions are necessary or appropriate. In §3, I argue that a good case for the existence and permissibility of epistemic punishments can be made along many of these lines.

2.2 How and When to Punish

Our goal is now clearer. Sometimes, by marginalizing a person, we punish them. This section addresses the first two elements of punishment: it is unwelcome, and it is inflicted in response to (alleged) culpable wrong-doing. In the course of this discussion, we also have a good opportunity to go into more detail about the exact mechanisms by which we epistemically punish each other. We'll start there, since this will help to make the case that marginalization is unwelcome.

2.2.1 Participating in Inquiry

In trying to live our lives, we often find that we need to cooperate with others to find out what the world is like. This can happen in highly formalized settings: a committee requests an official report from a team of researchers or some academics co-author a paper. But the most familiar cases are largely informal. You ask some friends the best way to get to the movie theater. Some of them suggest answers, others object, the conversation turns to when the construction on Kalorama will be finished, and why it's so far behind schedule. Over the course of a normal conversation, part of what we do is exchange information, but we also propose new lines for investigation (by e.g., asking questions), suggest hypotheses and potential explanations, and offer arguments regarding what's been said so far. These are some of the many ways that individuals participate in shared inquiry.

The punishments that I discuss in this paper are ways of preventing people from participating in shared inquiry.The simplest way to deny wrong-doers the opportunity to contribute will be just not asking them for their opinion. This doesn't go very far in preventing them from testifying, unsolicited. But often enough, people don't answer questions they aren't asked. They don't spontaneously take part in other people's conversations. While not asking may have a mild incapacitative or deterrent effect, it is unlikely to effectively promote those goals or any other proposed job for punishment. By analogy, suppose that Charlotte and Damon have an existing relationship. If Damon just doesn't invite Charlotte to his next party, it is all too easy for Charlotte to misinterpret Damon's action as a minor oversight. She may show up anyway (they've always been such good friends!), and she is unlikely to reflect on how she's acted or be deterred from acting the same way in the future.

A natural next step is to pointedly ask for others' opinions, but not the wrong-doer's. So, in a conversation where Florence is present, Giana may say "Herbert, Isaac, what do you think about vaccines?". Since Florence is standing right there,

she is likely to get the message: her participation is not welcome. When you are purposefully and obviously excluded from a conversation, this is experienced as insulting. Moreover, the person who treats you this way intends it to be insulting.

Increasing the harshness of the treatment, a punisher can pretend that a wrong-doer has not spoken, and continue the conversation as normal. So, when Jake says that the earth is flat, Kah can proceed to make whatever point she was making, as if Jake hadn't said anything at all. As in the case where a contribution is not solicited in the first place, this will be more effective, and more unpleasant, when done pointedly. Kah can wait a beat, to make it clear that she heard what Jake said, before refusing to engage with it. Going further, Kah can laugh it off as a joke, and refuse to treat it as a serious contribution. She can respond "Right! Can you believe that some people really think that?".

Marginalizing can be done in a more confrontational way, as well. You can talk over someone whenever they start to speak, making it more difficult for them to get their point across. At an extreme end would be to explicitly confront a person with their wrong-doing: "You're a liar. I don't care what you have to say."

I discuss these ways of treating someone because they have two features. First, the person subjected to such measures will usually find them unpleasant. Some people may not be hurt by such behavior, for example if they don't respect the person who treats them this way. But normal humans, as a matter of fact, do not like to be marginalized. It is insulting and can be humiliating. I am not sure which of these actions withhold benefits and which impose burdens. Many people want to have good standing within their social circles. We like to talk to each other, and make contributions to shared inquiry and conversations. When interlocutors make it more difficult to participate, we can say that a person loses out on the benefit of having a voice in the conversation. Additionally, to the extent that they actively insult or humiliate, these actions harm the wrong-doer.

In the end, I am not sure that it matters so much whether the epistemic punishments withhold benefits or impose burdens; each seems to have elements of both. It is enough, for my purposes, to note that they are unpleasant and unwelcome. If we wish, we could make progress on this question by considering the parallel non-epistemic treatments: to what extent does telling someone you don't want them around anymore withhold a benefit? impose a burden? This establishes the first element of punishment: unwelcome treatment.

The second reason I present these as my epistemic punishments is because they are modelled on epistemic injustices that have already been discussed in the literature. Each of these actions is presented as a way of making it more difficult for someone to testify (and thereby add information) or contribute in other ways to shared inquiry (propose explanations or hypotheses, suggest new lines of inquiry, etc.). When people are treated unjustly with regards to their testimony, they suffer from testimonial injustice.[6] Hookway (2010) offers unjust ways of excluding people from shared inquiry that aren't testimonial. Though he doesn't name it, I suggest we call it "inquisitory injustice".[7]

Relying on the existing literature offers two benefits. For one, it illustrates a recipe for finding more epistemic punishments. First, find an epistemic injustice. Second, find cases where the usually-unjust action is permissible. Third, see if punishment explains the difference in permissibility across cases. More significantly, it gives us a shortcut to establishing part of the second element, specifically that treating someone in these ways is typically wrong.

Fricker describes several ways in which testimonial injustice wrongs those who

---

[6]The phrase "testimonial injustice" is most closely associated with Fricker (2007). I intend it to apply to the various injustices that target an agent's information-sharing capacities, including Fricker's but also Dotson (2011)'s, Mills (2007)'s, and others.

[7]I don't mean to suggest that there is a clean break between the testimonial and wider inquisitory considerations. Actions taken to prevent people from testifying are likely to prevent them from contributing to inquiry in other ways as well, and the punishments I've proposed are suited for double duty.

are made to suffer it. I will focus on only two. First, when Sarah's testimony is prejudicially discounted, she "is undermined... in a capacity essential to human value" (Fricker 2007, p. 44), either the capacity to form true beliefs or the capacity to communicate those beliefs sincerely. Testimonial injustices reflect judgments that the testifier is either incompetent or insincere. Since these capacities are at the core of who we are as epistemic agents, undermining or insulting agents in this respect "cuts deep" (44). Second, Sarah suffers "exclusion from the community of epistemic trust" (45). As social creatures, we have a significant interest in being members of this community.

If Fricker's second explanation of the wrong is correct, then it is appropriate to talk about exclusion from the community of informants, as I have. We should expect the analogy between epistemic punishment and other punishments that cut wrong-doers out of a community to be tight. However, this is ultimately unnecessary. Whatever explains why testimonial injustice is wrong will help to explain why marginalization is typically wrong, when done for bad reasons.

Let's turn to another kind of epistemic injustice. Hookway (2010) provides the following case:

> The ... example involves a poor teacher whose behavior in dealing with a student can be seen to manifest a kind of epistemic injustice ... The teacher is engaging in discussion with her pupil ... However, when the student raises a question which is not a request for information, and is apparently intended as a contribution to continuing debate or discussion, then the teacher makes a presumption of irrelevance and ignores the question ... In this case, the student is not treated as a potential participant in disucssuon. (155)

In this case, the student is denied the opportunity to participate in inquiry by raising questions and suggesting avenues for discussion. In line with Hookway's diagnosis, we can identify the wrong here by noting that we are not just sources or

recipients of information. Active inquiry is just as central to our epistemic lives, wherein we raise questions, offer explanations and counterexamples, develop theories, and otherwise contribute to investigation (157). When the teacher undermines the student with respect to these capacities, it can cut just as deeply as when she does it with respect to their "informational" capacities. Although I find the explanation of the wrong that Hookway suggests plausible—it parallels Fricker's first account of the wrong of testimonial injustice—we don't have to commit to it. Again, it is enough to note that treating another person in this way is typically wrongful.

This only goes half of the way to establishing the second element of punishment; that it is typically wrongful. To get the second element, we also need to show that these actions are taken in response to some alleged culpable wrong-doing. The remainder of this section offers some examples of culpable wrong-doing that merit epistemic punishment.[8]

Before moving on, though, two points. First, there's no reason to suppose that marginalization can or should function as a punishment only when the punisher themself has been wronged. Seeing how Charlotte has treated Damon, you might start to treat Charlotte in the same ways. Or, knowing how Lars has intentionally misled a friend, you might marginalize him, too. This is just to say that, sometimes, you can epistemically punish someone even though you personally haven't been a victim of their wrong-doing. We rely on each other to get the job of punishment done.

Relatedly, just as with any type of informal punishment, marginalization will be more harmful and more effective as more people are willing to do it. Being ignored or confronted by one person may hurt someone and encourage them to reflect on their behavior. But it can also be dismissed as a single person's being rude or holding a grudge. When everyone in a group treats you as if your contributions are irrelevant or disvaluable, you are more likely to take their message to heart, or at least think about what's led to this point. This is just to say that, as in other areas of social life,

[8]In §4, I address the sense in which the wrongs "merit" the punishments.

group action can have a bigger impact than a unilateral decision.

2.2.2 Epistemic Wrongs

Here, I present some kinds of culpable wrong-doing, which I call epistemic wrongs. In this section, I argue for a descriptive claim: we can expect that the epistemic wrongs will be met with marginalization.

In discussing culpable wrong-doing, I break each instance into two parts. First, the wrong done. The kinds of wrongs I treat here are ways of giving bad information. In particular, causing someone to believe falsehoods through false testimony, or offering false testimony that is not believed by the hearer. Second, the degree to which the agent is culpable for the wrong, whether they intentionally, knowingly, recklessly, or negligently commit the wrong.

This is based on the criminal framework of dividing a crime into *actus reus* and *mens rea* elements. The first targets the harmful or wrongful conduct, and the second the mental state of the alleged wrong-doer.[9] Since we are dealing with informal wrongs and punishments, a legal framework may be seen as inappropriately formal. I stick with the actus/mens distinction because it gives us a clear and familiar way of finding culpable wrong-doing. This approach has its drawbacks: what I say implies that a sincere testifier with well-justified but false beliefs does wrong, though non-culpably. Some may balk at this claim. We don't need to get into this controversy here. What I need for my purposes is that when both elements are present, agents are culpable for wrong-doing. I trust that my arguments can be translated into your favored theory.

To make things concrete, suppose you ask a group of peers for their opinion on

[9]Beth Henzel points out that there are at least two places for mens rea to figure into the wrongs involved in changing another's beliefs. First, an agent might host mens rea with respect to the truth-value of their interlocutor's belief—they might intend that or be reckless regarding whether the belief is false. Second, they might host it with respect to how they change the belief. They might intentionally communicate by saying so, or recklessly communicate it by carelessly leaving a note out in the open. I will only discuss the first of these places for mens rea, and assume that when a speaker changes a hearer's beliefs (or attempts to), they do so intentionally. Future work can focus more on the second place for mens rea.

some current event; say a recent shooting. You would be wronged if someone caused you to believe falsely, by way of false testimony. Your goal, in posing this question to others, is to reach the truth about the circumstances surrounding the shooting: who committed it, why, and what is to be done next? You have an interest in believing the truth on shooting-related matters, and as long as your interest is legitimate, you would be wronged.[10] Moreover, having a false belief may lead you to protest for, vote for, or donate to the wrong things, frustrating further interests of yours.

There are many ways to get someone to believe the false through false testimony. When done intentionally, the testifier has told an effective lie. If Lars is trying to shore up support for a fringe political position, he can lie to you. He intends that you should believe falsely on this matter, because he knows his political position won't win out on its own merits. He is culpable for the wrong he has done. Moving down the scale of culpability, it is possible to knowingly but not intentionally testify falsely. Lars may not particularly care about what you end up believing. Instead, he wants his political allies to hear him parroting the party line. He knows his testimony isn't true, and he knows you'll end up with a false belief. But in testifying, all he wants is his confederates to hear him say the falsehood. This also seems to count as a lie, but more importantly Lars is again culpable.[11]

Reckless testimony is easy to imagine: In the hours after a shooting, reports are conflcting and constantly-changing. Rashaad has his preferred news source, which is generally reliable, and has read their recent report about the motives of the shooter. He tentatively believes the source, but knows that there is a decent chance that new

---

[10]Your interest may be illegitimate; you might want to find the most painful way of torturing an innocent person. In cases where spreading falsehoods is justified by other considerations, the testifier won't be liable to punishment. It will be impermissible to punish them, epistemically or otherwise. However, you will see your interest in the truth as legitimate and so will see yourself as wronged. It is possible for you to punish the testifier, and this may even be rational.

[11]There are, of course, many ways to getting others to believe falsely intentionally or knowingly. Some may count as lying, others misleading, or a third category entirely. Some may be better or worse than others, or wrong in additonal ways as well (Saul 2012, ch. 4). I leave it to other theorists to develop those claims.

evidence will come out within the next few hours. If he confidently testifies about the shooter's motives, he has been reckless. He has ignored the fact that there is a substantial risk his testimony will be false, that you will subsequently believe him, and that your interests will be set back.

The lowest grade of culpability is negligence, wherein an agent does not know (there is a substantial risk) that their action will be harmful. However, they should have known, or a reasonable person similarly situated would have known. Nelly's preferred news source, FactBattles, has a horrible track record. They have often been caught out in lies and conspiracy. Nelly has her own rationalizations for why FactBattles is nevertheless reliable, but they strain credibility. Still, Nelly sincerely believes what FactBattles says, and testifies to you that no shooting has occurred. It was a hoax perpetrated by the deep state. Nelly does you wrong, if you accept her testimony, but we cannot say she did so intentionally, knowingly, or recklessly.

She sincerely believes she is telling the truth and takes FactBattles to be an authoritative source. Still, given their bad track record and her flimsy rationalizations, she should have known, and a reasonable person would have known, that this is a baseless conspiracy. Nelly has been negligent.

Going forward, it would make sense for you not to give these people as much of a voice. Once you find out that their testimony was false, and that you have been misled, you will no longer regard them as experts on the matter. At the very least, you probably wouldn't elicit their testimony on this topic again. There is still a substantial choice about how far to go in punishing them, and this will depend on many factors including how far the testimony was from the truth, how culpable the testifier was, and the testifier's own track record. For instance, you may decide that an ideologue like Lars needs to be shut down immediately. Once you uncover his lie, he shouldn't be given a place to speak on current events. You might be more patient with Nelly. If she uses bad sources once or twice, you might let it slide. But if she cites FactBattles confindently and constantly, you would begin to discourage

or prevent her from contributing. This parallels non-epistemic wrongs: we are less lenient with a person who harms intentionally than one who does so negligently.

For what I've said, we've only gotten to discouraging further testimony from the bad testifiers. We'd like examples where people are prevented from contributing to inquiry in other, non-testimonial ways. They aren't too far off. When an ideologue like Lars or a dupe like Nelly suggests a new line of inquiry, we can expect a cold reception. An interlocutor's suggestion that there's a link between gun violence and economic conditions might lead to interesting and fruitful discussion. When our bad inquirers suggest a link between gun violence and the lizard people controlling the government, others will not follow that thread. When someone has been a bad testifier (and so a bad participant in collective inquiry), we are less willing to allow them to participate in the conversation in other ways.

At this point, we have four kinds of culpable wrong-doing, and we can expect that people will impose the epistemic punishments on wrong-doers. Moreover, the severity of the treatment varies in just the way that severity of punishment does. The worse the harm, the greater the culpability, or the poorer the track record of the wrong-doer, the more severe punishment we can expect. This shows that the epistemic punishments fulfill the second element of punishment.

We could stop here, but there is another group of epistemic wrongs that I want to consider. Oftentimes, people offer false testimony that their hearers don't believe. The audience might be better informed, and so see the falsehood for what it is, or they may distrust the speaker in the first place. Even when false testimony doesn't change anyone's beliefs, I think we can expect people to respond in epistemically punitive ways. However, it's harder to make the case that there is culpable wrong-doing here. Since no one ends up with a false belief, no one's interests are set back.

Imagine that Lars, Rashaad, or Nelly gives their bit of false testimony. For whatever reason, you don't believe them. You haven't come to a false belief, but we should

still expect that you will respond in the ways I have claimed are epistemic punishments. This poses a puzzle for me: people do (and presumably should) deny others the opportunity to participate in shared inquiry, even when there is no culpable wrong that they are responding to. To support the claim that these responses are punitive, I need to point to culpable wrong-doing.

In response, I suggest an analogy with attempted murder. When unsuccessful, the attempt may not harm anyone. Still it is pretty clear that attempted murder is wrong. There are many theories that try to close the gap between the wrongfulness of successful attempts (murder) and mere attempts (attempted murder). I don't want to be committed to any of them here. As far as I can tell the view I develop is consistent with all of them. Attempts are characteristically accompanied by intention: often, if not always, when you attempt to ϕ you intend to ϕ. This accounts for one grade of culpability for false-but-ineffective testimony.

However, when agents testify to falsehoods knowingly, recklessly, or negligently, it is not clear what to say about them. "Negligent attempts" are not recognized by existing criminal law, and there's an air of oxymoron to them. Instead, we can try to account for these cases as endangerments. Knowingly, recklessly, or negligently doing things which endanger others is wrong, and agents are culpable for it.[12]

I've described eight epistemic wrongs. We have identified two actus elements (attempting to or successfully getting others to believe falsehoods) and four mens elements (intention, knowledge, recklessness, and negligence). In response to each of these, we can expect that agents will impose the epistemic punishments on testifiers. As argued earlier in this section, being treated in the ways characteristic of epistemic punishment is unpleasant and typically wrongful. So we have established that

[12]Chiao (2010) argues that we should see attempts in general as a species of endangerment. Others may be resistant to the general move, though want to count what I just called "negligent attempts" as varieties of endangerment. This doesn't obviously help: few Anglo-American juridictions recognize negligent endangerment (Cahill 2007, for a discussion of the resistance to criminalizing non-intentional attempts).

epistemic punishments possess the first two features of punishment: they are unpleasant responses to (alleged) culpable wrong-doing, that would otherwise be typically wrongful.

In the next section, we turn to the last two features of punishment. Punishments condemn the agent, or their actions. And punishments have some additional purpose, depending on the particular theory of punishment we're working with. Once we have settled these two features, we can also argue that it is permissible to punish, and not just that it happens.

2.3 The Purpose and Justification of Punishment

Let's take stock of what we've done so far. I've said that there is a class of epistemic wrongs, and that they can be perpetrated with different grades of culpability. We've also seen a couple of proposed epistemic punishments. The existing arguments from the epistemic injustice literature give us reason to think that marginalizing someone is typically wrong, but we can see why epistemic wrongs would be met with the epistemic punishments. Moreover, I've argued that marginalization can have two important features of punishment. First, it is unwelcome, and imposes a burden or withholds a benefit. Second, it can be inflicted in response to (alleged) culpable wrongdoing.

However, we have unwelcome ways of dealing with wrong-doers that aren't necessarily punitive. If I intentionally damage your car, you'd be within your rights to call up my insurance company and get compensated. Of course, I wouldn't want you to do this, since it will raise my insurance premium. So you have responded to my culpable wrong in a way I find unwelcome. But this isn't punishment. You are not "out to get me"; it might not matter to you whether a wealthy benefactor pays for your car, so that I bear no cost at all. These interactions are part of corrective, not retributive, justice. Or recall what happened to Evan. Fernando ended their

friendship in part because of Evan's culpable wrong-doing. While this can be done as a form of punishment, it can also be done out of sheer exhaustion. Fernando isn't concerned with pointing fingers and determining blame; he just wants to move on.

The last two elements of punishment distinguish it from other responses to wrong-doing. First, as Feinberg points out, punishments condemn agents or their actions. This is why Evan has not been treated punitively. In cutting him out, Fernando doesn't condemn him. While condemnation is recognized by most theorists to be a part of punishment, they disagree on what else punishment is for. We may punish in order to gain some forward-looking benefits, like deterring the bad testifier or others from their behavior. Punishers may be seeking retribution for the wrong done. Or punishment may be a key part in communicating to a wrong-doer that they've done wrong. Most likely, these all have some place in our actual practices of punishment, and there may be further motives besides.

In this section, I show how marginalization can condemn bad testifiers, and can serve the other ends of punishment. Since the purpose of punishment is closely related to its justification, I will also argue that there are moral reasons to epistemically punish. Of course, my argument on this point can be no better than the argument in favor of punishments in general. If you think punishment is never justified, then I won't be able to convince you that epistemic punishment is.[13]

2.3.1 Condemnation

According to Feinberg (1965), condemnation has two aspects. First, it expresses disapproval of the wrong. Second, it expresses resentment and other reactive attitudes, which arise in response to the wrong.[14] In denying bad testifiers the opportunity

[13]Similarly, if seeking retribution or deterrence is never justified, then epistemically punishing for those reasons isn't either.

[14]I couldn't hope to give a complete list of the reactive attitudes which are expressed in condemnation. The list will likely be very close to the reactive attitudes which are tied to blame: Feinberg cites anger and hostility, and something like gratitude clearly has no place in condemnation.

to contribute to inquiry, we sometimes condemn them. This is most obvious with the more confrontational forms of epistemic punishment. Calling someone a liar to their face effectively conveys disapproval of what they've done. Adding that you don't care what they think is a direct way of expressing resentment, disappointment, anger, and even hostility. Even the less overt epistemic punishments can condemn. The cold shoulder, in the right place and time, is a powerful sign of disapproval and resentment.

This immediately distinguishes epistemic punishments from non-puntiive ways of marginalizing.[15] You wouldn't consult your six year-old niece on economic matters, and if she were insistent on being party to a discussion, you might ask her not to speak. The same can be said for adults who lack expertise on some matter. Suppose Guillaume forms beliefs responsibly and wishes to share them in a group discussion. If the other inquirers have access to better evidence or are better able to interpret the evidence, he might find himself with very little room in the conversation.

Treating others in this way is familiar and legitimate, but it is hardly punitive. The difference, I suggest, is that in asking your niece to keep quiet, or in not spending as much time with Guillaume's contributions, you are not condemning them. You might well disapprove of your niece's or Guillaume's contributions (since they waste time), but it would be abnormal to resent them for it or get angry about it. Some words that we use to marginalize, "liar", "crackpot", "conspiracy theorist", "shill", and the like, are vitriolic. They condemn in a way that other ways of marginalizing do not.

So far, this gets to the descriptive issue, of whether we do epistemically punish. Since we do sometimes condemn agents in marginalizing them, this establishes the third element of punishment. However, this leaves the normative issue untouched. Is condemnation justified? In this respect, marginalization seems no different than

[15]I am grateful to Laura Callahan, Will Fleisher, Doug Husak, and Pamela Robinson for pressing me on this point.

other kinds of informal exclusion. Compare Lars the liar with Charlotte, who insulted Damon. When their actions are harmful enough, and they are fully culpable, negative reactive attitudes seem fully appropriate and worth expressing. Condemning a negligent testifier like Nelly seems less appropriate, but this is as it should be. It is harder to justify punishing the negligent, unless their negligence was extreme or often repeated. As Nelly exhibits a lengthier or more egregious track-record of relying on FactBattles, disappointment, and even anger and resentment, begin to feel more appropriate.

Still, the question of why condemnation should be expressed through punishment remains. What reason is there to cut people out of the conversation in a condemnatory way? Why not condemn Lars on Monday, with a strongly worded letter, and then on Tuesday treat his contributions like we would Guillaume's? I don't have an answer to this. In general, it is hard to point to a reason that we condemn by hard treatment, rather than separate the two. At any rate, the question is no harder when it comes to Lars than Charlotte.

2.3.2 The Point of Punishment

Having established that marginalization can condemn, we have only one element left in arguing that there are epistemic punishments. It is impossible to punish for no reason whatsoever. In punishing (as opposed to just harming), the punisher hopes to accomplish something.

Consequentialists point to the forward-looking benefits of punishment: wrongdoers are made unable to re-offend, or are deterred, or even rehabilitated. Consequentialists don't have a monopoly on these purported benefits, of course, and non-consequentialists can appeal to these purposes as well as others. Perhaps punishers seek retribution in one form or another. Most likely, any combination of these can motivate a punisher, and may have a part to play in justifying hard treatment.

I will consider these proposed purposes of punishment, and argue that marginalization (in some cases, together with an accompanying act) can accomplish them.[16] Moreover, if marginalization can accomplish the purposes of punishment, then this justifies epistemic punishment, at least in some cases: agents are culpable for wrongdoing, so we have reasons to punish them. Marginalization is an effective way of satisfying those reasons, and so it is justified.

Forward Looking

When we punish, we do so at least partly in the hopes that some good will come of the way we've treated the wrong-doer. We hope that our punishments will ultimately be for the best. There are standardly three sorts of forward-looking reasons people punish.

We have already briefly discussed one forward-looking reason to epistemically punish: incapacitation. In incapacitating a wrong-doer, we try to prevent them from doing wrong, or at least make their future wrongs less harmful. Since Damon refuses to spend time with Charlotte, she is unable to hurt him again in the future. Lars, Rashaad, and Nelly are similarly incapicated. With less license to contribute to inquiry, it is more difficult for them to mislead their fellow inquirers. Unless they are interrupted or spoken over, nothing prevents them from offering bad testimony anyway. However, given their bad reputations and low standing in their communities, their actions are unlikely to cause harm. No one will believe what they say. This shows that marginalizing a potential informant can incapacitate them, and so marginalization can fulfill the first role for punishment.

In the criminal law, deterrence is an oft-cited reason to punish. Knowing what awaits them if they do wrong, the wrong-doer and other potential wrong-doers are

[16] I choose these families of views about punishment because they are familiar and popular. See Bedau & Kelly (2017), Wood (2010*a*,*b*). I think a case for the existence and permissibilty of epistemic punishment can be made along, e.g., communicative grounds (Duff 2001, von Hirsch 1993), though for reasons of space I cannot pursue them here.

disincetivized. Effectively, this is what happened to Lars in the case we started with. Since he doesn't want to put up with the insults and the hassle of posting, Lars is deterred from sharing bad information. If others can expect to be similarly marginalized, they will be deterred from spreading falsehoods. In trying to deter, it will usually be appropriate to explain to the wrong-doer why you are punishing them. They will not know which behaviors to stop, and so will not be properly disincentivized, unless they know what they've done wrong. This is already a part of our practice: in condemning a bad informant, we are usually ready to cite specific instances of wrong-doing as the basis for marginalization.

Lastly, we might punish in an effort to rehabilitate. Hard treatment clearly gives agents a prudential reason not to do wrong; this is the idea behind deterrence. But, we can hope, punishment can inspire people to change their ways on a deeper level. They'll come to see why their action was wrong and who it hurt, and will become more sensitive to the moral reasons not to do wrong. Hard treatment by itself may sometimes accomplish this: if things go well, Charlotte's exclusion from her circle of friends will give her the opportunity to think about how she got to this point. Reflecting on what's happened, she might resolve to change her ways.

This is optimistic, to say the least. Hard treatment, by itself, is unlikely to inspire a person to change their life. If we punish in the hope of rehabilitating, the punishment will have to be accompanied by things like explanations of the wrong, and some arguments that it was wrong, and that the wrong-doer should change. I think we sometimes see this in marginalization. Imagine that your cousin has been roped into vaccine denial, and they begin to testify to all sorts of falsehoods, convincing other members of your family. Moreover, your cousin has been reckless. The next time you are all gathered together, you won't let your cousin have free rein to say whatever they want. You will try to marginalize them, and may go so far as to condemn their bad testimony.

But you don't have to stop there. You can explain why your cousin's evidence

doesn't provide a good argument against vaccines. You can explain what the evidence in favor of vaccines is, and why the dominant medical methodology is better than the anecdote-based methodology used by anti-vaxxers. Though you are willing to marginalize your cousin for now, you also try to get them to see the truth and change their behavior. You try to rehabilitate them as an informant.

Retributive

Aside from forward-looking reasons to punish, retributive reasons are the most familiar to philosophers. To many, punishment seems to be an intrinsically appropriate, deserved response to wrong-doing. In retributing, we "get back at" the wrong-doer. I think that marginalization, when done in a spirit of condemnation, does satisfy the felt need to retribute against bad testifiers.

Part of the reason why people give others a bad reputation is because, I think, they see it as deserved. Given their bad track record, the testifier no longer deserves to be admired or respected.[17] They no longer deserve a place at the table in collective inquiry. The pursuit of retribution can also explain why people go to such great lengths to condemn and marginalize others on platforms like Twitter. Lars's detractors may not expect much in the way of forward-looking benefits: it's all too easy for him to make a new account and escape his bad reputation. But, they feel, he deserves to have his lies exposed for what they are. It would be inappropriate to let him go, even if he will just make a new account and keep spreading lies. If this is part of what motivates people in marginalizing others, they are seeking retribution to some extent.

This helps to establish that people sometimes do seek retribution through marginalization, and now we can ask whether it's justified to retribute in this way. Unfortunately, there is no agreed-on retributivist explanation for why culpable wrong-doing justifies punishment. It's clear that it would be inappropriate to throw a parade

[17]At least, as a testifer. They may be a good friend in other respects.

for a murderer, but it's much harder to explain why. I can't make progress on this point while remaining neutral between competing forms of retributivism, other than to point to the felt appropriateness of marginalizing a bad testifier. So, I'll briefly illustrate one familiar theory of what justifies retributive punishment as a proof-of-concept.

Morris (1968) argues that culpable wrongdoers fail to shoulder the burden that we share in making the world a good place to live. At the same time, they enjoy the benefits the rest of us can offer. Wrong-doers are, effectively, free riders. Punishment removes that benefit, which hasn't been fairly earned. Whatever the merits of this theory in general, it doesn't look so bad when it comes to bad testifiers. We rely on each other in inquiry. Bad testifiers don't shoulder their fair share of the burden, in making sure that their contributions are true. Still, they seek the benefits of participating in shared inquiry, including the status that goes along with being seen as trustworthy. Marginalizing and condemning bad testifiers removes this benefit. If this is a good reason to punish, then there is a good reason to epistemically punish.

Since marginalization can accomplish the goals of punishment, it fulfills the fourth condition for punitive action. In the right circumstances, marginalizing someone is an unwelcome response to wrong-doing. It condemns bad informants, and tries to reap forward-looking benefits or exact retribution. Therefore, marginalization can be an epistemic punishment. Insofar as condemnation, deterrence, and the like are worthy goals, epistemic punishment will be justified. This completes the argument for the existence and permissibility of epistemically punitive practices.

At this point, I have explained what epistemic punishments are, why we should think they exist, and why it is permissible to epistemically punish. There are many details to fill out, and I will start to do so in the next section. Before moving on, I want to address one more point about the justification of epistemic punishment. In showing that epistemic punishment is justified, I pointed to the goals that the practice, in general, can accomplish. This is very different from justifying any particular instance

of epistemic punishment.[18]

In any given case, we'll have to consider many factors to decide the extent to which someone should be epistemically punished, just as we always must when it comes to punishment. Cutting Lars's tongue out would be an effective, condemnatory way of margianlizing him. But that's clearly not justified. It is a wildly disproportionate response to his wrong-doing. Even the ways of marginalizing that I've suggested can go too far. If Nelly is slightly negligent on one occasion, calling her a liar and constantly telling her to shut up is not justified. On top of being disproportionate, it will be too harmful for Nelly, given the forward-looking benefits you can expect from marginalizing her.

A particular instance of epistemic punishment may be unjustified for another reason. As I argued, epistemic punishment has a deterrent function. But deterrence can go too far. If Rachel sees that people are harshly condemned for making small errors, she may rationally decide not to contribute in the first place. It won't be worth it for her to take that risk. Rachel might have a lot of good information, incisive explanations, or interesting questions to share. Deterring people like Rachel is bad for the goals of shared inquiry. Epistemic punishment is unjustified when it creates a chilling effect on shared inquiry.[19]

2.4 Details of the View

I've argued that marginalizing someone can be a way of epistemically punishing them, since it can possess the four key aspects of punishment. This is only the first step in developing a theory of epistemic punishment. In this section, I give more details and answer questions about what epistemic punishments are and how they work.

---

[18]Compare Husak (2000).

[19]Compare Dotson (2011).

2.4.1 Why "Epistemic"?

For the most part, I've focussed on the punishment part of "epistemic punishment", and payed much less attention to the epistemic part. In what sense are epistemic punishments epistemic? I say that marginalization can be an epistemic punishment because it concerns how we manage collective inquiry. It is, and should be, a part of our practices when we are engaging in shared inquiry and determining how to treat others as epistemic agents.

Is that enough to make these punishments "truly epistemic"? I'm not sure it matters. That seems to bear on whether I'm doing social epistemology or applied ethics, but not much else. Additionally, I'm not entirely sure what the question is even asking. The kinds of agents that I've said deserve epistemic punishments have failed in epistemic ways. Lars, Rashaad, and Nelly spread falsehood.[20] Their assertions violate the proposed epistemic norms: their testimony isn't known, true, or justifiably believed. In Lars's case, it isn't even believed.[21] And we have good epistemic reason to punish them. Insofar as they are incapacitated, deterred, or rehabilitated, marginalization helps to promote the truth and avoid error. If all this is not enough for the punishments to count as "properly epistemic", I don't know what else is wanted.

At any rate, recall the reason we posited epistemic punishment in the first place. We started with the observation that Sarah suffers from epistemic injustice, and Lars does not. In many ways, epistemic punishments are the flip side of epistemic injustices. This objection could be just as well posed against epistemic injustices, and I am satisfied if epistemic punishments stand or fall together with them.

---

[20]Goldman (1999, §3.4) argues that we have epistemic reason to care about the aggregate level of accuracy in a community and that others have an intrinsic, epistemic interest in believing the truth.

[21]See for instance Williamson (2000) and Lackey (2007) on norms of assertion.

2.4.2 Culpability and Marginalization

A different objection targets the need for positing epistemic punishments in the first place. We started with the puzzle of why Sarah's treatment was wrong but Lars's wasn't. I proposed that punishment explains the difference. But, the objector can say, we don't need to appeal to punishment, and culpability doesn't really play a role in explaining the difference. Sarah was helping to spread truths, Lars wasn't. Sarah was treated in ways that don't help to spread truth, and Lars was. We should marginalize people to the extent that they threaten to spread falsehoods or prevent us from discovering the truth. This consequentialist explanation does everything we asked without appealing to culpability or punishment. We don't need to add anything new to our theories.

First, this argument doesn't establish that epistemic punishments don't exist or aren't permissible. If you are a consequentialist then you can explain every instance of punishment along the lines of the previous paragraph. That is what punishment looks like for consequentialists; they don't conclude that there's no such thing as punishment. It is, in general, hard to find a place for culpability to fit into a consequentialist theory.

This won't be a satisfying response for non-consequentialists. We still have the puzzle: it seems permissible to marginalize people when their testimony would spread falsehoods, even if they aren't at all to blame for it. As discussed in §3, that may be true. Even then, culpability determines whether and how much to condemn a person when marginalizing them. Testifiers with false but justified beliefs should be marginalized, but not in a condemnatory way. Additionally, it is not clear to me that culpability is irrelevant in deciding whether to marginalize. Consider the following case:

> **Dangerous Knowledge**: Simon is a devoted and conscientious scientist. He has discovered that there is a slight correlation between race

> and working memory. The evidence strongly suggests that this correlation isn't genetic, and is entirely due to environmental factors. Simon takes care to frame and phrase his results accurately when speaking to the public.
>
> But, irresponsible science journalism and dishonest politicans lead the public to get a different idea. The belief that black people are intrinsically less intelligent starts to spread. You are the dean of Simon's university. The cheapest, most effective way to turn the tide of falsehood is to fire Simon and write a letter officially disavowing him and his work.

Should you fire and condemn Simon? Maybe. If there are no other good options, this may be the best course of action. But you shouldn't do so too quickly or too gladly. It is worth it to look into alternatives, like public information campaigns, even if those would be more expensive or less effective. Though condeming Simon may be, ultimately, right, there is something regrettable about it.

On the other hand, if Simon were falsifying data to make it look like black people were less intelligent, your decision would be much easier. Firing and disavowing Simon would be an obvious choice. Notice that in deciding how to treat Simon you are partly determining how he can participate in shared inquiry in the future. If he comes out of this as a disgraced scientist, it will be much harder for him to find another job. He would be silenced. Cases like **Dangerous Knowledge** show that decisions about marginalization do depend on whether agents have engaged in culpable wrong-doing.

2.4.3 Other Wrongs

I have treated the spreading of falsehood as the main wrong meritting punishment. However, we have seen that marginalization is also, normally, wrongful. For this reason, it would be natural to claim that these are further wrongs, meritting epistemic punishments themselves. For instance, epistemically punishing those who commit

epistemic injustices. I am open to this claim, though I don't want to be committed to the idea that those who commit epistemic injustices should necessarily be marginalized. Other punishments may be more appropriate. This is something that will have to wait for a different paper.

Additionally, I've ignored the wrong of merely believing the false. When a person believes what is false, they don't promote the truth and avoid error. They do go wrong. If they do so recklessly or negligently, then my account seems to imply that they merit epistemic punishment.[22] Intuitively, though, this seems wrong. Imagine a person like Lackey (1999)'s creationist teacher. The teacher believes that the theory of evolution is false, but only ever testifies that evolution is true because they don't want to lose their job. For similar reasons, the teacher never testifies against a biological fact, despite their beliefs. It wouldn't be right, I think, to marginalize the teacher, if they never even attempt to spread falsehood.[23] Yet, it seems I am committed to this claim.

I have two lines of response available. First, if the teacher merely believes falsehoods, and never testifies to them, there is no forward-looking benefit in punishment. If the teacher is good at their job, punishment may itself promote the false. In general, when a form of punishment provides no forward-looking benefit and poses a risk of harm, we have reason not to punish in that way. This gives an explanation for why we should not marginalize those who *merely* believe falsely, while leaving open

---

[22]Intentionally or knowingly believing the false may be possible. People who voluntarily take hallucinogenic or dissociative drugs, or even drink alcohol, might intend or know that they will come to have some false beliefs while inebriated. The complication is that, in these cases, you are doing something *now* that diminishes your capacity to believe the truth *later*. At the moment of belief-formation, your diminished capacity may mean you are not responsible for your poor epistemic performance. I want to avoid the kettle of fish which is determining culpability in "tracing" cases, where the agent is responsible now for the fact that they won't be responsible later.

On a different note, Sutton (2005)'s known unknowns, are realistic examples of knowingly believing what you don't know, which is a related way of going wrong.

[23]There is a sense in which the teacher does attempt to spread falsehoods. They attempt to get people to believe certain claims, and they believe that those claims are false. However, since the claims are in fact true, this is an impossible attempt. I do not have the space to explore impossible attempts here. See Westen (2008) for an overview of some of the issues here.

the possibility that they should be punished in other ways.

However, I want to offer a stronger line. In merely believing falsely, the teacher only wrongs themselves. It is hard to justify punishments for wrongs that are done *only* to the agent themselves.[24] This is the case even for higher grades of culpability, like knowledge or intention. Suppose, for instance, that out of self-hatred, I cut my own arm off. If I survive, should I be punished? The issue here is not just that punishment is unlikely to have forward-looking benefits (if I hate myself that much, hard treatment might encourage me and others like me), but that I don't seem to deserve punishment in the first place. When agents harm themselves with mens rea, paternalistic intereference may be warranted. For instance, we might control the evidence they have access to in hope that they change their mind.[25] However, paternalism is different from punishment. Most significantly, the bulk of the reasons in favor of paternalistic interference must concern benefit to the agent. Reasons that favor punishment are offered, largely, in terms of the well-being of people other than the agent, potential victims.

2.4.4 Other Punishments

Suppose that we have the opportunity to convince a wrong-doer of a falsehood. This would harm them, and they certainly wouldn't welcome it, so it seems like a candidate epistemic punishment. But that doesn't seem right. Getting liars or negligent testifiers to believe falsehoods is not justice, but some odd form of petty revenge.[26]

I agree with the main thrust of the objection—that it is not typically appropriate to spread falsehoods to wrong-doers—and the view that I offer can deliver this result. Getting a wrong-doer to believe falsehoods is unlikely to have forward-looking benefits. Especially if we do not also marginalize them, this is likely to spread falsehoods

[24]See Husak (2013) and Mill (1859)

[25]See Ahlstrom-Vij (2013) on epistemic paternalism.

[26]Thanks to Laura Callahan, Simon Goldstein, and Pamela Robinson for discussion on this point.

to other, innocent people. Marginalization is likely to provide greater forward-looking benefits, and for that reason should be preferred. So, I do not think it is in-principle wrong to get offenders to believe falsely, just that it will typically be sub-optimal.

There may be some extreme cases where it does make sense to get offenders to believe falsehoods. Suppose that the wrong-doer has built up a really terrible reputation. Others in the community judge the wrong-doer to be anti-reliable. They become less confident in what the offender says. In this case, getting the wrong-doer to believe, and then testify to, a falsehood would actually promote true belief. If the benefit is great enough (and the punishment proportionate), then we may have reason to punish in this way.

2.4.5 Lex Talionis

I have proposed that we epistemically punish in response to epistemic wrongs. So it looks like I am saying that the punishment should fit the crime. To be explicit, I don't think that epistemic punishments must only be inflicted in response to epistemic wrongs, or that other kinds of wrongs cannot merit epistemic punishment. I think that epistemic punishments are likely to have the greatest forward-looking benefits when performed in response to the epistemic wrongs.

In any particular case, I haven't offered a reason in-principle to suppose that marginalization will be the best way of promoting truth and avoiding error, or is an intrinsically fitting response to wrong-doing.[27] In the other direction, epistemic punishment may be appropriate in response to other kinds of wrong-doers. Suppose you discover that a colleague abuses their children. You might refuse to work with that person, cite their work, invite them to conferences, and find other ways to deny them the opportunity to participate in shared inquiry. This can very well be a fitting

[27]Megan Feeney suggests that communicative theories might help to explain why epistemic wrongs merit particularly epistemic punishments. By punishing in an epistemic way, we affirm as a community that we properly value the epistemically valuable.

punishment, and just as effective as other forms of punishment. For these reasons, I do not think we should adopt lex talionis for epistemic punishments.

This is some reason to think that epistemic punishment need not be punishment in kind, that we shouldn't adopt an epistemic lex talionis. But I have only offered one case against it. If you have arguments that, in general, punishments should fit their wrongs, it would apply to epistemic punishments just as well. For reasons of space, we cannot fully explore whether epistemic punishments are particularly fitting.

For all I've said, marginalizing bad informants is just one permissible form of epistemic punishment. We could have reason to epistemically punish in other ways, and we could have reason to punish bad informants in non-epistemic ways. In this section, I've argued that culpability should make a difference to who and how we epistemically punish. And I've explained some ways that epistemic punishment shouldn't go. We shouldn't punish people who merely believe falsehoods but are still helpful in shared inquiry. And it will usually be wrong to punish wrong-doers by getting them to believe the false. This gives us a better idea of how epistemic punishments fit into a broader theory of punishment, and what other kinds of epistemic punishment we could engage in.

2.5 Conclusion

I have argued that a part of our epistemic practice has all the hallmarks of punishment: alleged wrong-doers are subject to unwelcome, condemnatory treatment. And we are sometimes justified in epistemically punishing. To conclude this paper, I want to point to two avenues for future work, that are raised by my claims here.

It turns out retributive justice has a place in our epistemic practice. It is natural to wonder whether other kinds of justice have similar epistemic analogs. There might be epistemically corrective practices, wherein wrong-doers are compelled to repair the damage they have done. We can ask whether there is such a thing as epistemic

self- or other-defense. I suspect that we can permissibly engage in Dotson (2012)'s contributory harm and Pohlhaus (2012)'s willful hermeneutic to protect the innocent from an active threat. But those issues wil have to be explored in a different paper.

I have said that we can and do epistemically punish. But I haven't said who "we" are. In this paper, I've focussed on interpersonal examples involving small groups of agents, and argued about how they can and should act. However, research teams, schools, and other institutions have as a goal promoting truth and avoiding error. Insofar as they can harm and condemn, it seems possible for them to epistemically punish. But institutions have different obligations and abilities than individuals, so it is still an open question whether official epistemic punishment could ever be permissible.

In modern liberal states, we leave most of the punishment up to the state itself. Can the state justly epistemically punish?[28] The idea that the state could or should get involved in determining whose beliefs are true and who should be able to participate in inquiry raises issue of censorship, free speech, and political legitimacy. Given how much influence the state can have over the flow of information, and the promotion of true or false beliefs, the issue of state-sanctioned epistemic punishments is interesting and urgent.

[28]I am extremely grateful to Liam Bright, Beth Henzel, and Howard McGary for discussion on these issues, which I hope to return to in future work.

Chapter 3
Liberal Neutrality and False Beliefs

There are limits on the actions states can take. The state should not fund a multi-billion dollar project to better understand and combat a global decrease in religiosity. The state should not criminalize moral harms, with expert ethicists called in to testify about the extent to which the defendant corrupted the victim. Public schools should not teach Kantianism as fact, with other moral views presented as failures, like Lamarckian evolution is. Let's call these three policies the "bad policies".

There are other things the state should be allowed to do. It is vital that states spend a lot of money understanding and counteracting climate change. We do and should criminalize physical harm, with expert doctors testifying about the nature and cause of the harm. Public schools teach that the earth is round, and they shouldn't stop anytime soon. These are the "good policies". More generally, the state can decide which sources of information are good. The state is empowered to recognize some people as experts, some scientific methods as reliable, and some informants as trustworthy. The state can decide what is good or bad science, real or fake news.

In this paper, I argue that a prominent strain of liberalism cannot have things both ways. Neutralists, who believe that there is something wrong in-principle with the bad policies, cannot also support the good policies. Instead, we should abandon neutralism in favor of an instrumentalist or perfectionist liberalism.

If the bad policies are bad and the good policies are good, there must be a normatively relevant difference between them. I will argue that neutralists cannot identify a relevant difference. My argument proceeds in several steps. §§1 and 2 clarify the

target of my argument. §1 discusses and sets aside the forward-looking differences between the two policies. §2 looks more closely at what's at stake, and what the liberal would need to say to respond to the challenge of this paper. In §3, I argue that three responses to the challenge fail. Their failures are instructive, and allow us to generate a template for responses. The bad policies set back legitimate interests, and the good policies do not.

§4 discusses two applications of the template: people have an interest in autonomy and an interest in being respected by the state. We see that the good policies set back these interests, so the neutralist must argue that these interests are illegitimate. §5 considers two strategies for delegitimizing the interests: (i) ways of life based on false commitments do not deserve protection and (ii) practices and commitments that are not important to people do not deserve protection. We see that these attempts fail. §6 argues that reasonableness is the neutralist's best hope for distinguishing between the two slates of policies. However, the neutralist faces a dilemma. If the standards for reasonableness are too low, we cannot uphold the good policies. If they are too high, we cannot condemn the bad policies.

§7 concludes by briefly exploring what a non-neutral liberalism is like. Unfortunately, the dialectic surrounding neutrality has a lot of choice points. Though I have tried to do a lot of signposting and summarizing, it is easy to get lost in the argumenatitve fray. For your convenience, I include a flowchart on the next page. It is my hope that this mystical symbol will help guide you at each point in the paper.

3.1 The Easy (Correct) Answer

In this section, I will discuss some ways of distinguishing between the good and bad policies that are, broadly speaking, forward-looking. After this section, these strategies will be put aside until §7.

The worst thing about the bad policies is that they just aren't worth it. Religiosity

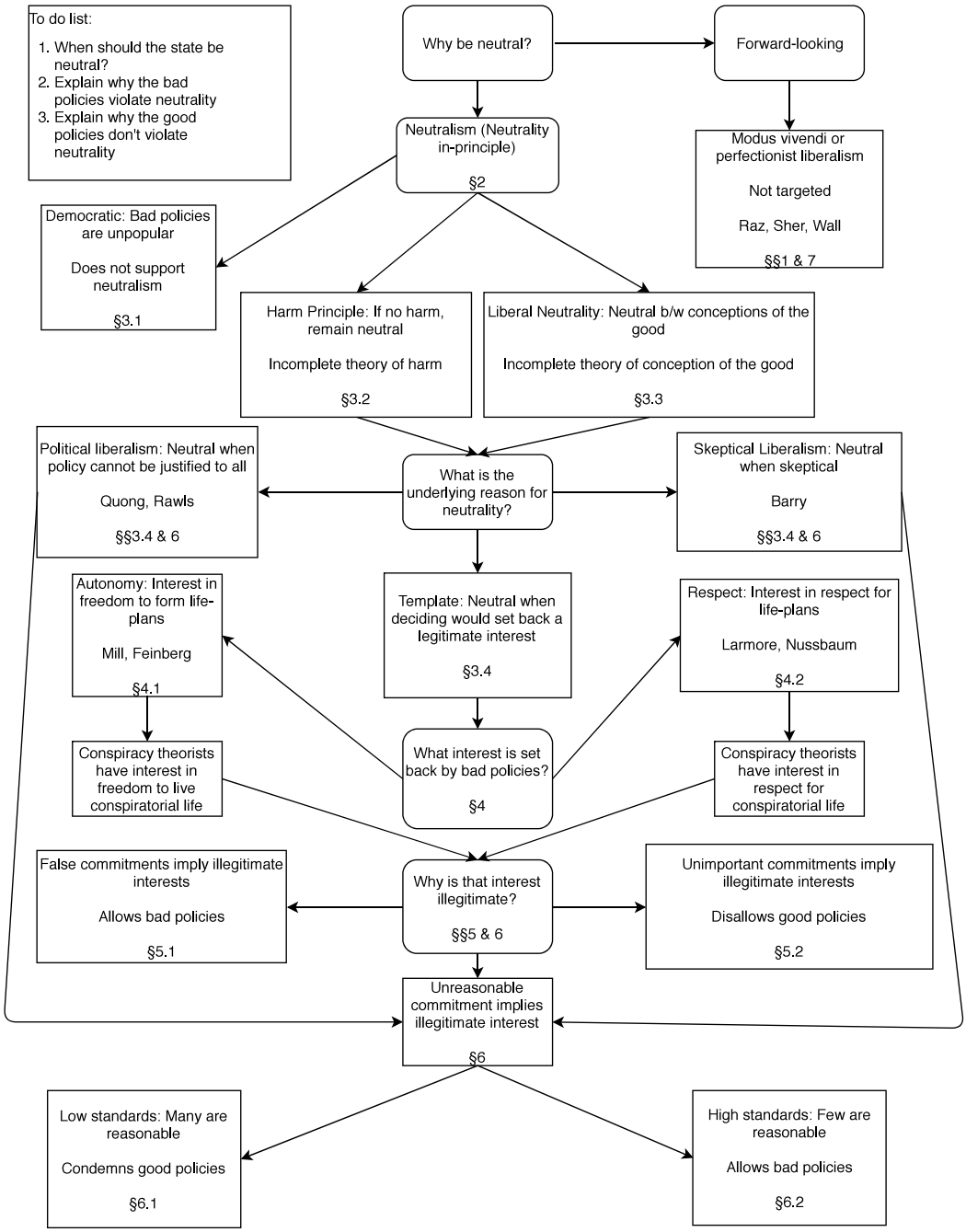

Figure 3.1: A Map of the Argument

isn't valuable in the first place. All the money the state could spend "fixing" the problem would be better spent elsewhere. To the extent that moral harm is real, the state has more important things to do with its resources. Kantianism is false, so the state should not teach it as fact.

On the other hand, climate change is a real threat to us and future generations. It's worthwhile to handle violent crime appropriately, and provide students with accurate information about the shape of the earth. A lot is at stake in deciding between good and bad science or real and fake news. We can explain the difference between the bad policies and the good policies easily, by reference to forward-looking, instrumental considerations.

To get a feel for this family of strategies, consider how a utilitarian would try to distinguish the good from the bad. For the utilitarian, what makes a policy (or anything else) bad is that it doesn't maximize utility.

Given what we know about ourselves and our history, contemporary liberal states would almost certainly not maximize utility by enacting the bad policies. A lack of religiosity isn't intrinsically bad, from the utilitarian point of view, in the first place. Moreover, states do not have a great track record of enforcing moral and religious norms in a way that makes people's lives better. We could expect civil unrest, and even violence, in the face of a state-sponsored religion. We should similarly expect a state-sponsored moral theory to work out poorly.

On the other hand, the good policies (apparently) help to maximize utility. Unchecked climate change would be a disaster. Insofar as the criminal justice system makes people's lives better, criminalizing assault is a good thing. A utilitarian argument that schools should teach the earth is round is not as straightforward, but it isn't beyond imagining. This gives us a familiar, tidy way of drawing the line between good and bad.

Of course, this is not proprietary to utilitarians. Consequentialists of any stripe can appeal to the promotion of their favored values to distinguish between policies. If

true belief is intrinsically valuable, for instance, then it is easier to see why teaching that the earth is round is justified.

This way of drawing the distinction is not even necessarily consequentialist, though to see this takes a bit more work. For example, suppose you think that there is a non-consequentialist duty to punish the culpable, and a prohibition on punishing the innocent. As a non-consequentialist, you don't want to just minimize the number of innocents punished. That would lead to things like punishing a known innocent now in order to prevent three innocents from being punished later. Still, all else equal, the fewer innocents punished the better.

On the view we're discussing, criminalizing moral corruption is not inherently objectionable. Putting a law on the books does not constitute punishing anyone, culpable or otherwise, so it doesn't directly violate any non-consequentialist duties. However, given the flaws of any legal system, having such a law will make it very likely that innocents will be punished. For that reason, it is a bad piece of legislation.

So, according to this family of arguments, to evaluate a policy we must look forward at its likely consequences, broadly construed. I think this the right way to account for the difference between the good and bad policies. However, I doubt that most liberals will agree that this is the whole story. Since we are only evaluating policies on a forward-looking basis, whether the bad policies are objectionable is a contingent matter.

To put the point another way, *if* a state-sponsored morality or religion held promise from a forward-looking point of view, then this family of arguments would allow that the state should enshrine that morality or religion. This would not be acceptable for most liberals: freedom of religion and freedom of conscience are usually thought of as foundational liberties. Even if state officials knew that a given moral theory were true and could do some good by forcing it on others, it would be objectionable to wield the state's power this way.

Following this line of thought, many liberals[1] hold that the bad policies are objectionable in principle. When it comes to morality and religion, they say, the state should not take sides. The state has no place in our churches or bedrooms. On these issues, it should remain as neutral as possible. We can call people who have this view neutralists.

Neutralism: The bad policies are objectionable in principle.

That said, not all liberals are neutralists. Modus vivendi liberalism holds that a familiar liberal scheme of rights is, as a matter of fact, justified on broadly consequentialist grounds. Roughly, a state that recognizes liberal rights is the best way we have developed so far to keep people from killing each other. Modus vivendi liberals do not think the bad policies are in principle objectionable.

There are also varieties of perfectionist liberalism that don't need to respond to my challenge. Perfectionist liberals claim that there are good arguments for liberal rights which go beyond the modus vivendi liberal's. For example, Raz (1986) argues for a perfectionist liberalism founded on the value of autonomy: recognizably liberal rights are necessary for respecting autonomy. However, perfectionist liberals like Raz, Sher (1997), and Wall (1998) can allow that the bad policies are not objectionable in principle.

Most liberals will want to go farther than the modus vivendi crowd, or this strain of perfectionism. They will say that there is something objectionable about the bad policies, even if forward-looking considerations favored them. The majority of this paper argues that, if neutralism is true, then the same objection can be mounted against the good policies.

---

[1]Dworkin (1978), Feinberg (1984*b*), Kymlicka (1989), Larmore (1987), Mill (1859), Quong (2010), Rawls (1971, 1996)

3.2 The Challenge

In the previous section, I discussed one (forward-looking) way of distinguishing between the good and bad policies. If we take this line, then under different circumstances, a state-sanctioned morality or religion would be justified. My challenge is addressed to those liberals who find that line of thought unsavory. Here, I make my challenge explicit, and elaborate on what the neutralist would have to do to answer the challenge.

To state my challenge informally, neutralists want to treat the bad policies as different in kind from the good policies. They say the state should not take sides when it comes to moral harms or religiosity. But the state may take sides in deciding that climate change is real, that biology is right and phrenology is bunk, and that some sources of information are unreliable. If the neutralist cannot justify this difference, then they must abandon the good policies.

If the neutralist goes down that path, then they say that the state may not take any action against climate change, may not criminalize assault, may not bar unqualified people from offering "expert" testimony, and may not teach the earth is round. That is a false theory of the state's power. The state should have the power to decide what counts as good or bad science; to determine whether a report is fake news. The state is not bound by "both sides"ism about every issue.

So, how can the neutralist justify treating the good policies so differently from the bad policies? I will argue that they can't. If the bad policies are objectionable in principle, so are the good policies.

In stating the challenge more carefully, it will be helpful to have a toy neutralist view on the table. Consider

> Liberal neutrality: In principle, it is objectionable for the state to promote some controversial conceptions of the good over others.

The thought is that there are some topics that are just not the state's business

one way or the other. When it comes to controversies over conceptions of the good, the state must remain neutral. It need not remain neutral on issues that don't involve conceptions of the good.[2] By teaching Kant in schools, for instance, the state sides with Kantians on their conception of the good. Teaching any other moral theory as fact would involve a similar taking of sides, and thus would violate liberal neutrality. The other bad policies can be condemned along similar lines.

I challenge the neutralist to accomplish the following set of tasks. First, give an account of what issues the state must remain neutral about, and what issues the state may decide on. We saw the account offered by liberal neutrality: the state must remain neutral between conceptions of the good, but it need not be neutral about other things. Ideally, this will be accompanied by an explanation of why the state must be neutral about the former but not the latter. What is special about controversial conceptions of the good, for instance, that requires the state to remain neutral between them? Second, explain why the bad policies are objectionably non-neutral. For instance, "which character traits are vices and which are virtues?" involves deciding between conceptions of the good. So liberal neutrality says the state cannot decide that a given character trait is a virtue. However, this is exactly what the state does when it attempts to promote religiosity. In this way, the proponent of liberal neutrality can explain why this policy is objectionable.

Third, the neutralist must explain why the good policies are not objectionable. There are two ways of doing this. First, they could argue that a policy does not involve any issues over which the state must remain neutral. For example, the liberal neutralist could argue that choosing a science curriculum involves promoting controversial conceptions of the true but *not* of the good. So it is acceptable for the state to take sides on what the best science curriculum is. Alternatively, the neutralist could argue that although a good policy does involve some issues over which the state must

[2]It is unfortunate that this principle has come to be named "liberal neutrality". As we'll see, a liberal could be a neutralist without believing in liberal neutrality; they don't have to put the same kind of stock into the fact/value distinction that liberal neutrality seems to require.

remain neutral, the proposed policy is neutral. For example, the liberal neutralist can admit that criminalizing assault means deciding in favor of some conceptions of the good. But, they could go on to say, states are required to remain neutral only between *controversial* conceptions of the good. And it is not controversial that assault is bad. Therefore, this policy counts as neutral and is not objectionable.

I will argue that the neutralist cannot accomplish all three tasks: successfully completing the first two makes the third impossible. In the next section, I will explain why liberal neutrality in particular cannot be the whole story. For now, liberal neutrality shows us the "shape" of the challenge, and responses to it. It will also help us in making some points of clarification.

First, I have credited neutralists with the view that the bad policies are "objectionable", and it is not at all clear what that means. It could be that individuals would do something morally wrong by voting for a bad policy. Or maybe it is permissible (though admittedly scuzzy) for citizens to vote for the bad policies, but it is morally wrong for legislators to propose and then enact them. Executives of the state might have a moral duty not to enforce or otherwise act on non-neutral policies. We could say that states with such policies are unjust or illegitimate, even if this does not imply anything about individual government officials.

It's likely that if you object to the bad policies in one of these ways, you will find them objectionable in at least one other way. I am ecumenical: my arguments do not depend on which reading we take. Whatever it means for the bad policies to be "objectionable", I will argue that the neutralist is committed to saying the good policies are similarly objectionable.

Second, to test whether a given state action violates liberal neutrality, we need to know when an action counts as promoting a controversial conception of the good. For this reason, liberal neutrality is often given three readings: neutrality in effect, intention, or justification. Neutrality in effect holds that an action counts as promoting a controversial conception when it has effects which happen to favor some conceptions

and not others. This is overly demanding. Whenever the state acts, there will be winners and losers. For example, recognizing a five-day work week gives people the free time to go to religious services, and so it may lead to an increase in religiosity among the citizenry. This should not count as promoting a controversial—religious—conception of the good.

Neutrality in intention and justification are more plausible. They hold, respectively, that the state promotes a conception of the good when its action is intended to promote some conception or when it justifies its actions by reference to such a conception. We see that neutralists of any kind are able to appeal to the same distinctions in explaining when a policy is neutral or not.

To give an example that doesn't appeal to liberal neutrality, some Christians claim that a literal interpretation of the Bible entails that the earth is flat. A neutralist may (but doesn't have to) approach the round earth curriculum as follows. The state must remain neutral about religious matters. In choosing a round earth curriculum, the state touches on an issue that it must remain neutral about. It is likely that teaching that the earth is round will cause people not to adopt that particular strain of Biblical literalism. This policy is not neutral in effect, since it happens to promote one side of an issue. What matters is that the state can justify its actions without wading into religion, and that it does not intend to promote any (lack of) religion. That is what we find when it comes to teaching the round earth in schools, so the round earth curriculum counts as neutral.

In later sections, I will argue that some forms of neutralism imply that the round earth curriculum is non-neutral. I mean this in the senses of neutrality of intention and of justification. It would be unfair of me to object to neutralism on the grounds that the good policies are non-neutral in effect. The neutralist can grant this, and go on to say that it is irrelevant.

Turning to the third and final point of clarification: the neutralist claims that the bad policies are objectionable in principle, but they can also agree with much of what

forward-looking theorists have to say. Neutralists can agree that there are forward-looking reasons against the bad policies and forward-looking reasons in favor of the good policies. This potentially opens up a line of response for the neutralist. They could try to argue that the good policies *are* objectionably non-neutral. However, they agree with forward-lookers that there are strong forward-looking reasons to enstate them. These reasons outweigh the objections of principle that they have against the policies.

Neutrality is upheld because the bad policies are in-principle objectionable. So are the good policies. But, all-things-considered, the good policies are good and the bad policies are bad. For example, given the likely catastrophic effects of unchecked climate change, the forward-looking reasons for doing something about it are overwhelming. This may be a rare case where states would be justified in violating even very strong liberal rights. The neutralist has thus avoided answering the challenge. I do not think this strategy will work in all cases, though. To bring this out, I focus on the contrast between the Kantian and round earth curricula in public schools.

The goods that we gain by teaching the true shape of the earth are moderate, at best. Having a true belief on the matter does not impact most people's lives. It isn't as if flat earthers are living horrible lives. If you agree with my imagined neutralist that the round earth curriculum does violate liberal rights, I think you'll find it hard to argue that the good of teaching round earth justifies violating them. Still, if you are willing to give the neutralist a bit more breathing room, this paper can be read as an extended argument that the in-principle reasons against teaching the true morality are as strong as those against teaching the true geography.

If the neutralist says the in-principle reason to be neutral about the shape of the earth is only pro tanto and overridden in this case, then the reason is quite weak. This is because the forward-looking reasons to teach the true geography are fairly weak themselves. Correspondingly, the in-principle reason against teaching the true morality are weak. I think we would gain a lot more if people believed the truth

about morality than we do by having them believe the truth about the shape of the earth. Therefore, a properly designed (and bloodless, etc.) moral curriculum would be all-things-considered justified, despite its in-principle objectionability.

Technically, that's a version of neutralism. But it isn't the neutrality that many liberals seem to want. They want freedom of conscience and religion, period. Not freedom of conscience and religion, until we can design a moral or religious curriculum that is as good as second-grade science. At that point, we might as well just abandon neutralism altogether.

So, the curriculum issue is a good case study. The forward-looking reasons in favor of an effective geography curriculum are weaker than the reasons in favor of an effective (bloodless, etc.) moral curriculum. A robust form of neutralism claims that the in-principle reasons against the moral and religious curricula are much stronger than the reasons against teaching that the earth is round. I will argue that the neutralist cannot make good on this claim.

Now that we have gotten clear on the challenge, we can start to look at attempts to answer it.

3.3 Answering the Challenge: Three Attempts

This section has two goals. First, we'll see how three different attempts to answer the challenge fail: appeals to democracy, liberal neutrality, and the harm principle. A "death by a thousand cuts" argument against neutralism would be unilluminating. So we will take a step back and think about how the neutralist could respond more generally. This will serve as the basis for the discussion in the remainder of the paper.

3.3.1 Democracy

There is plenty of overlap between liberals and those who support democratic governance. The state is constrained by the will of the people, and the bad policies

are insufficiently reflective of their will. It is not entirely obvious how an appeal to democratic values allows the neutralist to accomplish the first task of the challenge. Presumably, it would be along the lines of "It is objectionable for the state to non-democratically decide some controversy". Unless the polity has somehow empowered the state to resolve an issue, it may not attempt to do so.

This looks like it will give a decent explanation of why the good policies are unobjectionable. The vast majority of people do want the state to criminalize some actions, set enviornmental policies, and design public school curricula. People might disagree with the exact policies that the state ends up going for, but that's a normal part of the democratic process. People accept that the state should have these powers, and that is good evidence that they are (or would be) granted to the state on a democratic basis.

However, this does not allow us to explain why the bad policies are objectionable. Suppose that 70%, a healthy majority, of the population were Kantians. After a fair election, we have a majority Kantian legislature. They decide that Kantianism should be taught in schools. Although it is democratic, it is still objectionably non-neutral.

It wouldn't be worthwhile for the democratic to try tweaking their view. There is a deeper (and familiar) problem at play here. Liberals, as a rule, believe that democratic decision-making is constrained by liberal ideals. No democratic majority could legitimately enslave a minority or prohibit a religious belief. Democratic processes might determine how the state should exercise its power, but it does not determine what powers the state should have in the first place.

The neutralist's problem with a state morality isn't that the state would choose a morality that didn't sufficiently reflect the will of the people. Rather, the state shouldn't have that power in the first place.

3.3.2 The Harm Principle

An old liberal favorite for limiting state power is

The Harm Prinicple: The state should act only so as to prevent harm.

This statement of the harm principle is too simplistic,[3] but it will do for our purposes. Though not stated in terms of neutrality, there is a straightforward way of applying the principle here. "It is objectionable for the state to decide an issue, unless harm-prevention is at stake."

This correctly categorizes some of the policies. A lack of religiosity does not itself harm anyone, so the state may not decide to take action against it. On the other hand, climate change and violent crime are harmful. The harm principle does not forbid the state from taking appropriate action against them. Correctly identifying witnesses as experts is necessary in courts, which are there to deter (and so prevent) harm. Agencies must decide which sources are reliable if they are to do their jobs effectively; the CDC cannot effectively prevent the harm of disease unless it can decide whether vaccines cause autism. When the state decides between good and bad science, or real and fake news, harm-prevention is at stake.

So the suggestion is that the good policies (are supposed to) prevent harm. Meanwhile, the bad policies do not appropriately target harm, so they are objectionably non-neutral. Unfortunately, this strategy does not apply as straight-forwardly as we might hope. On some views, corrupting a person's character harms them; it is good for a person to be virtuous, so you make their life worse by influencing them to be vicious. In that case, taking your friends out for a night of drinking, gambling, and general carousing is a harm to them. Legislation that punishes the life of the bon vivant would save the fuddy-duddies from their corrupting influence. Presumably, though, we don't want the state criminalizing fun *just because* it sets a bad example for others. Insofar as we recognize moral harm, we have to say that it is not the right kind of harm to trigger the principle.

Things get even trickier when it comes to the school curricula. Teaching students

[3]For one, most will want to reformulate it to exclude paternalistic intervention to prevent harm to the self. See Mill (1859) for Feinberg (1984*a*) for canonical statements.

to be Kantians does not clearly prevent harm to anyone, so it seems to be (correctly) forbidden by the harm principle. By the same token, though, teaching students that the earth is round does not clearly prevent harm to them. Flat earthers are not more prone to falling off Australia into the void of space. Perhaps we can say that teaching the earth is round prevents people from coming to believe falsely, which is bad for them. In that case, teaching the true moral theory in school would protect students from the harm of false moral belief.

My point is not to argue that no version of the harm principle can do the job. Rather, I am arguing that we're going to need a new, more carefully articulated version of the principle if we want to mark the difference between the good and bad principles. Some things (like false geographic beliefs) will have to count as harms but others (false moral beliefs) can't, for the purposes of the new and improved harm principle. It isn't clear how to formulate a suitable principle.

3.3.3 Liberal Neutrality

Let's return now to the principle of liberal neutrality: it is objectionable for states to promote controversial conceptions of the good. This is initially promising, as discussed above. Neutrality refers to "conceptions of the good", whereas many of good policies are based on promoting a conception of the facts. We want the state to act on the word of scientists and reliable informants. Agencies do this by hiring advisors, commissioning research, and through similar means. Courts do this by allowing experts to testify as to matters of fact. Schools do this by teaching the scientific consensus.

All these policies mean that the state is promoting a certain conception of the facts. The world is more or less as scientists think it is. Their methods are sound, for the most part. These are controversial claims: people are, unfortunately, taken in by conspiracy theories, bad science, and fake news. But (so says the liberal neutralist) it is alright for the state to take sides in that controversy. Insofar as any value is being

promoted, it is the value of truth. And this is not a controversial value. We can all agree that the state should act on the truth, even if we disagree about what the truth is.

So far, so good. But it is not clear what a "conception of the good" is for liberal neutralists. In traditional forms of liberalism, the state is also supposed to be neutral on certain matters of fact.[4] Whether God exists or not is a question of fact, but it is one that the state is supposed to stay out of.

"Conception of the good" is usually understood broadly enough to include religious beliefs. I don't think this is too much of a stretch. The way a citizen sees things, it is part of a good life for them to believe that God exists. And so belief in God is part of their conception of the good. The state should not involve itself in this matter. This won't help the neutralist unless they can explain why belief that the earth is flat is not part of some people's conception of the good. If they value truth and think that the earth is flat, then this belief seems to be part of their conception of a good life.

Liberal neutrality is running into the same problems that the harm priniple faced. They need to work out a version of the fact/value distinction so that religious and moral beliefs count as matters of value, but the shape of the earth doesn't.

This is not a knockdown argument against liberal neutrality or the harm principle by any means. However, I think this points to a flaw in these two neutralist strategies. Suppose that after a good ten rounds of Chisholming, the neutralist finds a way of understanding "harm" so that false geographic beliefs counts as a harm. We'd still be left wondering why that specific notion of harm is the relevant one for neutrality. Let's say that the liberal neutralist gerrymanders a notion of "value" so that questions of God's existence are questions of value, but questions about the best scientific methodology aren't. This doesn't tell us anything about why the state should care

[4]On the other side, it's worth pointing out that the state may not be neutral between all conceptions of the good. In prosecuting assault, it takes a stand that assault is actually bad/wrong. I take it that "controversial" helps to do the work here.

that the one question is technically a matter of value, while the latter is a matter of fact.

The harm principle and liberal neutrality fail to answer the challenge. Even if they could get the extensions right, and I doubt they can, they aren't able to explain why the state should be neutral when it should be. A different tactic is needed.

3.3.4 A Template

Here, I suggest a template for answers to the challenge. It is based on the explanatory failures of liberal neutrality and the harm principle. The idea is that what explains why the bad policies are objectionable isn't going to be radically different in kind from what makes other policies in-principle objectionable.

Different theorists will disagree about what, in general, makes a policy in-principle objectionable. Some will say that it violates a right. Others will say that it gives people grounds to lodge a certain kind of complaint, or to make a certain kind of claim. Perhaps it wrongs them in some way. Or they have some legitimate interests that are set back by the policy.

I am of the opinion that these languages are inter-translatable. People are wronged when their legitimate interests are set back, and their rights serve to protect their legitimate interests. Claims and complaints are grounded in the impact a policy would have on people's legitimate interests. I do not have the space to argue this point in detail here. For that reason, and because I think talking about interests allows for cleaner phrasing, I will use that language for the rest of the paper. If you think that rights-talk (or whichever) is importantly different from interest-talk, I invite you to translate my arguments into your preferred language. I do not believe it affects my arguments, but for all I've said it could open the door to better answers to the challenge.

At any rate, we can see the template for answers taking shape. To complete the first task, the neutralist should identify some specific interest I. When the state take

sides on an issue, it sets people's interest in I back. However, when that interest is illegitimate, it is unobjectionable for the state to take sides. For instance, in criminalizing assault, the state takes side on the issue of the value of assault. This impedes certain freedoms and sets back the interest some have in acting in ways that constitute assault. However, assaulters' interest in this freedom is illegitimate, so it is unobjectionable for the state to take sides here.

To complete task two, they need to explain how the bad policies set I back and why the interest is legitimate in this case. As before, task three can be completed in two ways. The neutralist can argue that I would not be impacted by the policy. Since I is not at stake, the state is effectively remaining neutral. Alternatively, they could argue argue that, although people's interest in I is set back, it is an illegitimate interest in this case. The state is taking sides, but this is a case where the state is allowed to take sides.

Framing the discussion around interests may not account for every neutralist out there. Unfortunately, I cannot address them in full detail in this paper. Here are three examples worth pointing out.

Political liberals, like Quong (2010) and Rawls (1996), argue that the state should only do things that can be justified to all reasonable citizens. As a rule, political liberals believe that a liberal form of government can be supported without reference to any substantive moral theory. They will probably resist framing their position in terms of legitimate interests or moral rights, which would require substantive moral commitments. Jønch-Clausen & Kappel (2016), Kappel (2017), Kappel & Jonch-Clausen (2015) have already addressed the political liberal take on this issue. They argue that political liberals cannot appeal to the strong fact/value distinction that they apparently need. I think there are some problems with their arguments, but they are largely correct. At any rate, since political liberals rely on the notion of "reasonableness", they will be subject to the arguments of §6.

A related argument grounds neutrality in skepticism. Barry (1995, 161–173) pursues this strategy. Since we should be skeptical of our moral and religious beliefs, we cannot justify encoding them in law. Given that skepticism about value, they may not want to appeal to any particular interest in supporting their position. After all, we can be skeptical about whether that interest is worth protecting. Without getting into the details of Barry's argument, he relies on the idea of reasonable disagreement or reasonable skepticism. I believe the skeptical liberal use of "reasonableness" gives rise to problems similar to the one I am pressing here. We'll pick up with them in §6.

Finally, I will completely ignore views which support neutrality on non-cognitivist or expressivist grounds. On this view, since there is no fact of the matter about God or morality, the state may not say as much. But there is a fact of the matter about the earth's shape, so the state may say so. I think this line of argument is hopeless, but I do not have the space to address it here.

3.4 Two Interests: Autonomy and Respect

Let's recap. The harm principle and liberal neutrality have a hard time separating the bad policies from the good. Even if they could get the extensions right, they leave the distinction unexplained. I suggested that the neutralist does best by responding to the challenge in terms of interests. Identify some legitimate interest set back by objectionably non-neutral policies, and argue that the good policies do not set back any legitimate interests. This could be because the good policies do not set back any interests, or because the interests they set back are illegitimate.

With that in mind, here is the structure for the rest of the paper. In this section, I will discuss two interests that are often implicated when talking about neutrality. People have an interest in their own autonomy, and they have an interest in being respected by their states. The neutralist can make good on the claim that the bad policies set these interests back, and so are objectionable in-principle. However, I

will argue, the good policies set back very similar interests. For example, while the Kantian curriculum impedes one form of autonomy, the round earth curriculum impedes another. Moreover, this is a kind of autonomy that people have an interest in.

If the arguments of this section are right, this closes off one avenue that neutralists have for defending the good policies. They do set back some interests. The neutralist must argue that, although the good policies set back interests, these interests are illegitimate. In §5, I discuss two arguments that the relevant interests are illegitimate. They both fail.

§6 considers the neutralist's best strategy (as I see it). Though people do have interests that are set back by the good policies, it is unreasonable for them to have these interests. Since the interests are unreasonable, they are illegitimate. Therefore, it is not in-principle objectionable for the state to set back these interests. This would vindicate the good policies. I argue that the neutralist faces a dilemma. If the standards for reasonableness are high, then the bad policies are unobjectionable. If the standards for reasonableness are low, then the good policies are objectionable. This completes the argument against neutralism.

3.4.1 Autonomy

For my money, an appeal to autonomy provides the best defense of *a* version of neutrality and explains its proper scope. There is reason to be suspicious that the proper statement of neutrality should apply to all state action. The principle of neutrality is, presumably, justified by appeal to some value or other. If the state does something to affirm neutrality (say, its constitution separates church and state) the state will thereby affirm the values that support neutrality. But neutrality is controversial; for one, I am disputing it. So (the argument goes) it is impossible and undesirable to remain neutral on all controversies.

Autonomy resolves this puzzle. Properly respecting autonomy requires that we

allow people free reign over some areas of their lives: "the inward domain of consciousness...thought and feeling...opinion and sentiment...liberty of tastses and pursuits; of framing the plan of our life to suit our own character; of doing as we like... freedom to unite" (Mill 1859, I). Autonomy, either as a primary liberal value or one important value among others, counts in favor of neutral policy-making when it comes to the domains where an individual should be left free.

Of course, giving decent coverage of the relationship between autonomy and liberalism would take a lot more space than I have. We can't hope to pin down exactly what autonomy is or its moral import. Even if we help ourselves to a specific conception of autonomy, there's still plenty of work to get from there to any solid political conclusions. Still, we can look at what appeals to autonomy have in common and give a general outline of how they work.

People who believe that autonomy is important believe that each individual has a certain sphere of life over which they have sovereign right. They, and no one else, gets to decide what goes on in that sphere. For Mill, as we saw above, an individual is sovereign over their consciousness, tastes and pursuits, and the unions they form. Whatever its boundaries are, individuals have a range of options within that sphere, and they have a legitimate interest as to what goes on in it. This might give rise to an absolute right to sovereignty, or it may be only one consideration, overridable in dire circumstances.

This provides a handy explanation of why the criminal law should be neutral.[5] By imposing criminal penalties on (say) public advocacy of consequentialism, the state effectively closes off an option that should be left to individuals. People have a legitimate interest in being free to be consequentialists, and criminal sanctions set that interest back.

This doesn't quite get us to an explanation of why the Kantian curriculum is bad. People would still be able to deny Kantianism, and may not face penalties for it.

[5]This was Feinberg's primary concern.

Similarly, a massive public campaign combatting the rise of atheism doesn't have to force anyone to be religious. The state may try to convince people of religion's value, run PSAs, or grant tax rebates to those who can demonstrate their piety. But that is a far cry from forcing anyone to be religious or Kantian.

We can extend the explanation of neutrality about the criminal law in a couple of ways. For example, "closing off" an option is a matter of degree. We know that some people are willing to face criminal prosecution for the demands of their conscience, so criminalization doesn't entirely close those options off. Perhaps the Kantian curriculum makes non-Kantian options more difficult for individuals, and so partly closes it off. If the value of options is important (Raz 1986), then we can say the Kantian curriculum makes non-Kantian options less valuable than they should be. I think there are plausible mechanisms by which state action can make these options less valuable or more difficult. We'll pick up on that later, but for now let's grant it to the neutralist.

The point is that an autonomy-based condemnation of the Kantian curriculum will have to identify some people whose autonomy is threatened by the policy. This threat to autonomy will involve the state somehow interfering with what goes on inside an agent's legitimate sphere of control. A case can be made that the Kantian curriculum is inimical to the autonomy of three kinds of agents. First, it could undermine student autonomy. Teaching students that Kantianism is true exercises influence over an important moral commitment which is within students' sphere of choice. It sets back students' interest in choosing their own values.

Second, dissenting teachers are put in a position where they must publicly advocate for values that they reject. The option "keep my job and don't advocate for Kantianism" has been closed off. Their legitimate interest in being able to live with integrity has been set back. Last, dissenting taxpayers fund a project that they think is worthless. The option "spend that money on something that aligns with my non-Kantian values" has been closed off. Taxpayers have a legitimate interest in deciding

what moral system to advocate for, and that interest has been set back.

So it looks like an appeal to autonomy can explain why the bad policies are objectionable. people have legitimate interests, grounded in their autonomy, which are set back by the bad policies. However, the good policies also interefere with people' autonomy.

Consider the round earth curriculum. A major point of teaching that the earth is round in schools is to influence how students form beliefs, so that they come to believe in the true shape of the earth. The state is exercising influence over an empirical commitment that students make. The options "keep my job and don't advocate for round earth" and "spend that money on something that aligns with my flat earth beliefs" have been closed off from dissenting teachers and tax-payers, respectively. Both policies tread on people's spheres. By interfering with the autonomy of people, the good policies set back their interests.

If autonomy is going to win the day here, the neutralist must argue that the relevant interests are illegitimate. For example, teachers who believe in the flat earth do have an interest in being free to pursue a life of flat earth integrity. For them, this means not espousing a round earth. However, they have no legitimate interest in this freedom. For comparison, a racist might claim that to live with integrity, they need to incite violence against people of color. Disallowing that means impeding their autonomy, and setting back their interests. Though they've developed an interest in hate speech, we don't need to worry about setting that interest back. It is an illegitimate interest.

3.4.2 Respect

A second line of argument in favor of neutrality appeals to respect for persons (Larmore 1990). Like autonomy, there are millions of ways of spelling out what respect is, and what constraints it places on politics. We'll try to stay at a high level, and talk about respect in a way that is friendly to all its proponents.

Respecting persons requires that we respect their choices, even when we disagree with them. This explains why paternalistic interference is unjustified; it disrespects the paternalized person's choices. Moreover, grounding neutrality in respect for persons explains why neutrality in effect is less important than neutrality in justification or intention. If I happen to accidentally step on your toe, I haven't disrespected you. If I do so intentionally, then that does disrespect you as a person. If I justify my action by citing the fact that it will hurt you, that's an even worse form of disrespect.

So respect for persons looks like a solid basis for neutralism. This isn't quite enough to explain the problems with the Kantian curriculum. Since people are not actively prevented from rejecting Kantianism, the curriculum isn't obviously paternalistic. There's no person whose choices are obviously disregarded or disrespected. In line with Nussbaum (2011), we should say that respect for persons requires that we don't publicly denigrate them and their commitments. Objectionably non-neutral policies denigrate some people, their choices, or ways of life. Thus, people are disrespected. Since they have a legitimate interest in their being respected, the policies are objectionable.

This gives a tidy explanation of why the Kantian curriculum is unjustified. By teaching Kantianism as fact, the state effectively says that non-Kantians are mistaken. Their moral commitments are less worthwhile than their Kantian neighbors'. The state, apparently, tolerates their presence. But it cannot treat them "as fully equal ends in themselves".[6] On this view, the bad policies are not bad because they close off options. They are bad because they denigrate and disrespect persons who take the state's non-preferred option. With a mind to autonomy, this could also be used to explain how the state makes options less valuable than they should be.

However, this same denigrating, disrespectful message is sent to flat earthers by the round earth curriculum. There is a very straightforward sense in which we, as members of the polity, do not respect flat earthers or their beliefs. To compare

[6]Nussbaum (2011, 22).

someone to a flat earther is to insult them. In general, whenever the state decides on a matter of fact, it changes the dynamic between those who support the state and those who don't. "Conspiracy theorist" and "crackpot" are stigmatizing terms that we apply to people who reject official (usually state-backed) accounts of events.

Of course, we should not blame the state for the fact that conspiracy theorists are stigmatized. The fact that state action need not be neutral with regards to effect is important here.[7] Even if the state can foresee that deciding on some matter will change how people treat dissenters, this could still be unobjectionable as long as its intentions and justifications are neutral. But when the state decides to teach that the earth is round, runs information campaigns, or names some people as experts, its aims and justifications aren't neutral. In presenting p as true, the state is intentionally portraying belief in p as the better option. It can only justify the decision by reference to the fact that p.

Granted, the state may not intend to stigmatize p-deniers, and it does not justify its action by reference to the fact that it will stigmatize p-deniers. But this cannot be the relevant measure of a violation of neutrality. The state intentionally portrays p as true, and justifies its action by reference to the fact that p. This is not a way of respecting people's commitment to $\neg p$.

Imagine that the state officially announces, "There is a God. But we encourage citizens to treat their misguided atheist neighbors well." If the agents of the state are sincere, we have no reason to believe they intend to disrespect atheists. But they do disrespect atheists, and they do this by intentionally portraying atheists as wrong, and justifying their announcement by reference to the "truth" of theism. If this constitutes disrespect of atheists, then when the state recognizes an official account, it disrespects conspiracy theorists, believers of pseudo-scientists, and consumers of fake news.

In this section, I discussed two interests that theorists have appealed to in order to

---

[7]Thanks to David Enoch for pressing me to elaborate on this point.

argue for neutralism. Ciizens' legitimate interests in autonomy and respect explain why the bad policies are objectionably non-neutral. I have argued that the good policies set these same interests back, though for different people. The round earth curriculum disrespects and interferes with the autonomy of flat earthers as much as the bad policies do to non-Kantians.

Admittedly, this argument is not air-tight. Neutrality might be necessary to protect other interests that I have not considered here. I do not have a conclusive argument that this cannot be the case. This section advances my case in two ways. First, autonomy and respect are the interests usually cited in support of neutrality. If there is some other interest that supports neutrality without also condemnig the bad policies, neutralists haven't been clear about what it is.

Second, and more importantly, my examples illustrate a pattern. It is entirely possible for a theocratic or Kantian state to treat atheists or non-Kantians in just the same way that states should treat flat earthers, anti-vaxxers, and other conspiracy theorists. Whatever atheistic/non-Kantian interests the theocratic/Kantian state sets back, we should expect unobjectionable governments to set back conspiracy theorists' parallel interests.

Since the good policies set back some people's interests, the neutralist must claim that these interests are illegitimate. That way, there is no in-principle objection against setting these interests back. In the next two sections, I argue that the neutralist cannot make good on this claim.

3.5 Delegitimizing Interests

As we saw in the last section, neutralists need a way of disqualifying flat earth and other conspiratorial beliefs from the protections that neutrality promises. Being a teacher who doesn't advocate for Kantianism as part of their day job is an option agents should be free to pursue, protected by autonomy. Or rejecting Kantianism is a

position worthy of the state's respect, that it should not denigrate. On the other hand, being a teacher who doesn't advocate round earth in class is not one of the options protected by autonomy. And accepting flat earth is not worthy of the state's respect. The neutralist needs to explain why the interest people have in non-Kantianism is legitimate, though flat earth interests are illegitimate.

In this section, I discuss two attempts to delegitimize the interests that arise from belief in a flat earth and other conspiracy theories. The first disqualifies them because they are false. The second holds that conspiratorial beliefs are not as important to people as moral or religious beliefs. We can dismiss both of these fairly quickly. In the next section, I address the view that only reasonable beliefs and practices deserve the protection due to a legitimate interest.

3.5.1 Truth

The earth is round, not flat. Flat earthers believe in something false. This is a tempting line of thought. Surely, the fact that flat earthers are *wrong* makes a difference to what the state can do.

So the neutralist could claim that a citizen's interest in ways of life based on falsehood are illegitimate. This line of argument falls flat immediately. The fact that a project of yours is premised on a falsehood cannot, by itself, delegitimize your interest in pursuing that project. Otherwise, the state could teach the true religion or morality. Since there is no God, people's interest in religious ways of life is illegitimate—not worth protecting—on this view.

3.5.2 Not Important

Some interests are more important than others. It is very important that the state respects my autonomy to decide whether and whom to marry. It is much less important that the state respects my and my spouse's autonomy when we decide whether

to blast obscene music at the reception. We can say the same thing about respect. Denigrating my choice in spouse is very disrespectful. Denigrating my choice in public obscenity may be disrespectful, but it's hardly momentous.

Perhaps flat earth and other conspiracy theories fall on the less important side of the spectrum. Then the protections afforded by autonomy or respect would be weaker, and so would the reason to remain neutral about the truth of the matter. Ahlstrom-Vij (2013) gives this kind of argument against epistemic autonomy, the freedom to believe whatever seems most likely to an agent, given whatever evidence and methods that agent chooses.

That form of epistemic autonomy is not directly relevant to what we are doing here. The state isn't preventing people from forming beliefs or penalizing them for their beliefs. Rather, we are concerned with the choices people make surrounding their beliefs. Can they arrange to live a life in line with what they believe? Or have some important options been closed off for them? Is the state disrespecting them and their plans?

Still, suppose Ahlstrom-Vij can show that there is no legitimate interest in being free to form beliefs by whatever means an agent deems fit. He will have gone a long way to showing that the interest people have in living by those beliefs is illegitimate. He says:

> These liberties [that we have a legitimate interest in] pertain to activities central to figuring out how to live one's life and to forging strong and meaningful bonds with others. In so doing, it is not just the quality of the outcome that matters, but also the process by which outcomes—be they optimal or not—are realized. (Ahlstrom-Vij 2013, p. 95-96)

Ahlstrom-Vij thinks that we have a legitimate interest in the autonomy to make decisions central to our life plans and our relationships with others. Extending his argument, we can say that agents have a legitimate interest in having such decisions

be respected. He argues that forming a belief on a scientific question doesn't have these properties.[8] To his list, I would add a third property. When a decision has a special meaning for agents' self-conceptions, both as individuals and as members of the polity, this helps to legitimate the interest agents have in the autonomy to make that decision and have it be respected.[9] A person's moral or religious convictions make a big difference to their life plans, relationships with others, and self-conceptions. That is why they are afforded such extensive protections on the ground of autonomy or respect.

But we can say the same thing about belief in a flat earth and other conspiracies. Whether you buy into conspiracy theories is important for figuring out how to live your life, which meaningful bonds to forge, and forming and maintaining a self-conception. I think there are good reasons why we should expect descriptive beliefs to matter in this way to any creatures remotely like us. But I need only the weaker claim that, as things stand, descriptive beliefs (and in particular conspiratorial beliefs) are important to people.

Here, I am not pointing to the mundane fact that descriptive beliefs play a role in means-end reasoning, and so have a large influence on how we conduct ourselves. Descriptive beliefs run much deeper than that. Tight-knit communities form around and against conspiracies. Flat-earthers and anti-vaxxers seek others of their kind. It is easy to find people online describing the rifts that arise in families and social circles when a member adopts anti-vaxx or flat earth beliefs. Rationalist societies of various stripes uphold their adherence to the scientific method and consensus, and

---

[8]It is clear from his discussion that he considers epistemic autonomy to be akin to epistemic independence: forming a belief without the influence of others. I do not have the space to discuss the reasons against this understanding of autonomy. Suffice it to say, a person can be autonomous and rely on others' judgments or assistance.

[9]In a more Kantian vein, we could add that our capacity to form moral or religious convictions is tied up with the fundamental features of who we are as humans/rational agents/persons. I think the recent literature in social epistemology, especially in the epistemic injustice literature, shows that the capacity to form and share descriptive beliefs is similarly fundamental (Fricker 2007). Our interest in having those capacities respected is just as legitimate.

associate with each other on that basis. People choose which relationships are strong and meaningful to them, where this is founded in part on what they believe the world is like.

For some, acceptance or rejection of flat earth, anti-vaxx, and other conspiracies is a point of pride. There is intense vitriol between the two sides over these descriptive matters. In the case of anti-vaxx and many other conspriacy theories, it is easy to point to the moral values at stake. Each side believes that lives hang in the balance. But there is no similar issue when it comes to the shape of the earth. Those who reject orthodoxy see themselves as iconoclasts, bravely standing up for the truth. Only they are enlightened enough to follow the evidence where it leads. On the other side, passionate defenders of the scientific consensus see themselves as especially rational. They uphold the correct methods for figuring out what the world is like.

When the state takes a stand on some matter of fact, it changes the social meaning of publicly advocating for a belief one way or the other. You set yourself either in line with or in opposition to the establishment.[10] Controversial descriptive beliefs are at the basis of how we understand ourselves and our relation to the broader polity.

These points together show that people figure out how to live their lives and form life plans in light of their descriptive beliefs. The fact that some people make a career out of advocating for or debunking anti-vaxx, flat earth, and conspiracy theories helps to strengthen this point.

So I grant that the strength and legitimacy of our interests can vary across different areas. If the neutralist wants to, they can claim that the more central a choice or commitment is to someone's life plan, relationships with others, and self-conception, the more important it is for the state to remain neutral about it. But if lives based around belief in a conspiracy theory are not worthy of respect, or if those choices aren't worthy of the protections of autonomy, it is not because they don't make a difference to life-plan, relationships, and self-conception. If these factors are what

[10]See Raz (1982) on this point, in connection with moral and religious belief.

legitimate agents' interests in having their moral values respected, they also legitimate conspiracy theorists' interests in being respected.

The neutralist needs a different way to argue that flat earthers' interests are illegitimate.

3.6 Reasonableness

In this section, I offer a final attempt to deligitimize the interests of flat earthers and other conspiracy theorists. People take an interest in all kinds of silly things, and this means state action is inevitably going to set back some interest or other. This doesn't mean the state can never act. If I make unreasonable demands on the state, the fact that its actions set my interests back counts for less.

Reasonable people can disagree about religion and morality. But it is unreasonable to believe that the earth is flat, that vaccinations are dangerous, or that astrology and phrenology are real sciences. This doesn't mean that flat earthers are to blame for their bad beliefs, or that they are bad people. We're all unreasonable about some things, and geography is where they shine.

The state need not respect unreasonable commitments or the choices based on them. This is a promising path for the neutralist. Liberals often restrict their principles using reasonableness. Political liberals hold that states should justify their actions to all reasonable citizens. Skeptical arguments for neutrality point to the reasonable doubts people should have regarding their religion or moral code. Moreover, arguments for autonomy or respect for persons don't cite the fact that people have the capacity to choose their commitments and life plans willy-nilly. The important capacity is the one to live in a way that is somehow responsive to reasons.

These observations together suggest that reasonableness should play a central role in limiting state power. The state must remain neutral when deciding an issue would infringe on the autonomy to make a reasonable decision. Or the state's deciding would

constitute disrespecting a reasonable commitment that people have made. Moving away from the language of interests, political and skeptical liberals can say that neutrality requires the state not to decide issues on which reasonable people could disagree.

So, the neutralist needs some way of saying that, whereas many different religious and moral commitments are reasonable, belief in conspiracy theories and defunct sciences is unreasonable. If they can secure this claim, they can argue as follows:

> It is reasonable to deny Kantianism in favor of (say) consequentialism, so we cannot expect all reasonable people to be Kantians. People have a legitimate interest in being free to live up to the demands of their consequentialist consciences. They have a legitimate interest in respect for their commitment to that moral code. Therefore, teaching Kant as fact in schools is unjustified.
>
> On the other hand, it is unreasonable to deny that the earth is round. We can expect all reasonable people to believe in the round earth. Some citizens might have an interest in living up to their flat earth beliefs, and this interest is set back by the round earth curriculum. This interest, however sincerely and deeply held by flat earthers, is illegitimate; it is based on an entirely unreasonable commitment. It is not a freedom worth protecting, or a decision worth respecting. Therefore, there is nothing in-principle objectionable about teaching that the earth is round.

At first, this looks like a slam dunk for the neutralist. As I've already admitted, it is unreasonable to buy into bad science and fake news. The problem is that "reasonable" is said in many ways.[11] I know a person who spends all their disposable income on video games, anime, and collectible figures. I think their life would be better if they developed some broader interests, learned a new skill, picked up a hobby. It is

[11]Here, I have benefitted from the discussion in Worsnip (2016).

unreasonable that this is all they want to do with their time and money, with their life.

But hey, they're happy. They pay their bills and taxes, and they aren't hurting anyone. Despite the fact that they waste an unreasonable amount of money on what amounts to garbage, who am I to force them to do otherwise? The neutralist should not say that my friend is being unreasonable in the sense that matters for neutrality. We would not accept the state passing legislation to the effect of "Citizens are only allowed to spend $50 a month on pop culture collectibles. It's unreasonable to spend more than that, and if that upsets you, then you're a loser nerd." Just as it isn't the state's place to decide whether Kantianism is true, it isn't their place to decide what hobbies are worthwhile.

With that in mind, the neutralist cannot conclude that belief in a flat earth is unreasonable in the sense they need to make their argument work. To properly answer the challenge, they need to spell out a sense of "reasonable" such that: (i) whether a commitment is reasonable is relevant for neutrality, (ii) denying Kantianism is reasonable, and (iii) denying round earth is unreasonable.

This is a delicate balancing act. If the standards for reasonableness are low and easy to achieve, then conspiracy theories will be reasonable. On this line, both the bad and the good policies will be objectionably non-neutral. If the standards for reasonableness are high and difficult, then the good policies will be acceptable. However, many religious and moral commitments won't be reasonable, and so policies like the bad ones become acceptable.

3.6.1 Low Standards

Rawls (1996, Leture II) has a not-very-demanding notion of reasonableness. For Rawls, a reasonable person accepts a basic scheme of liberal rights and justice. Nazis are unreasonable. Second, reasonable people are "ready to propose principles and

standards as fair terms of cooperation" between free and equal people. Third, reasonable people are rational, in a thin sense. They abide by the strictures of means-end reasoning. They form beliefs and make decisions in a consistent, coherent manner. We could say that their preferences are transitive, or that they are representable as expected value maximizers, or whatever.

Last, reasonable people accept the burdens of judgment. They understand that evidence on important matters is complex and conflicting and can be weighed in different ways. They accept that how a person interprets evidence can depend on life experiences they might not share. When making any decision, there are many reasons pulling in all different directions, and no person or institution can be expected to take proper account of all of them. To make a long story short, in accepting the burdens of judgment, reasonable people accept that finding points of agreement in politics is *hard*.

I think it is obvious that conspiracy theorists can buy into a basic form of liberalism and accept the burdens of judgment. There is nothing inherently illiberal about believing the earth is flat or that astrology is real. If Rawlsian standards are going to disqualify conspiracy theories from the protections of neutrality, it will have to be because they are irrational.

This move doesn't work. Conspiracy theorists often enough have a coherent picture of the world. Given what they believe on some matters, their other beliefs make sense. As before, let's focus on flat earthers. I will argue that their beliefs hang together in the right way, to count as rational. We might try to say that flat earthers are irrational because they disregard the mountains of evidence from scientists. But flat earthers have developed an entire epistemology, zeteticism, that explains why the evidence supports a flat earth.

They hold that the immediate evidence of the senses is an especially strong, often dispositive source of evidence. Since the earth appears to be flat, that is very strong evidence that it is flat. They have argued that the flat earth can account for much of

the evidence that laypeople can collect in their free time. For instance, when you are higher up, the horizon is farther away. We know this is because your higher vantage lets you see farther around the curve of the earth. Flat earthers chalk this up to facts about visibility and air density. Expert-provided data is suspect because scientists are in on the conspiracy. They have the chance to doctor photographs and mess with instruments so that they appear to show the curvature of the earth.

My point is not to defend flat earth or other conspiracies. They are unjustifiable. But they are not incoherent. There is *a* way of weighing the evidence so that it supports belief in a flat earth. It is a bad way, but it's there nonetheless. I assert that we will find this to be the case when it comes to anti-vaxxers, climate change deniers, and consumers of conspiracy theories and fake news generally. Their beliefs are not incoherent. If you weighed the evidence the way that they do, it would make sense for you to believe as they do.

To try to generalize the point beyond flat earth, for a belief to be rational in this thin sense is largely a matter of a belief's being justified by an agent's own lights. Having been convinced by good evidence, it is tempting to think that the sheer weight of it should tell against a conspiracy theory, no matter how bizarrely they weigh the evidence. But we disagree with conspiracy theorists on what the evidence even is. Where we see years of data and records, they see a pile of lies put out by conspirators to hide what they've done. By the lights of the conspiracy theorist, the testimony and data provided by our experts is worthless. In this thin sense of rationality, we should expect conspiracy theorists to be no more irrational than anyone else.

If it's easy to be reasonable then belief in a flat earth is reasonable. The interest in this belief and its attendant actions will then be legitimate, and so worthy of the protection of autonomy or respect. The state will have a reason to remain neutral about whether the earth is flat. So, while this line condemns the bad policies, it does not provide support for the good policies.

3.6.2 High Standards

We can raise the standards for reasonableness, so as to exclude flat earthers and other conspiracy theorists. I will argue that this leaves us with a view that is neutralist in letter, but not in spirit.

As we saw, the problem with conspiratorial beliefs is at its heart an epistemic one. There is nothing inherently morally or politically objectionable about believing that the earth is flat or that vaccines are dangerous. If the world were a different place, vaccines would be dangerous, and it'd be morally better for people to believe that they are. I can see two general strategies for raising the epistemic standards of reasonableness so as to exclude conspiracy theories. First, we could focus on the epistemic properties of the beliefs themselves. Second, we could point to the epistemic properties of the way these beliefs are formed. Any belief formed in an unreasonable way inherents its unreasonability from the method that produced it. For that reason, an interest in conspiratorial belief and life-plans based on it are unreasonable. Hence illegitimate, and not deserving of protection.

As for the first proposal, conspiratorial beliefs themselves are epistemically problematic. We know that the earth is round. It's incredibly unlikely that vaccines are dangerous. We could use some familiar epistemic standard to draw the line. If it is known that p, denying p is unreasonable. Or, if it is at least $x\%$ likely that p, denying it is unreasonable. This raises an immediate problem. Known to whom? Likely as judged by whom? You and I can agree that the chance that the earth is flat is near 0. But flat earthers obviously think its much more likely. They don't know the earth is round. We could slot in expert opinion, or the opinion of a legitimately empowered legislature or judiciary, or the state itself. There are specific problems with these proposals. To give a quick one: Why do astronomers count as experts, but astrologers don't?[12]

[12]Thanks to Laura Callahan, Megan Feeney, Ernie Sosa, and Chris Willard-Kyle for discussion on this point.

But let's set these worries aside. Assume that we can identify a good metric for who should know or deem p sufficiently likely, or what it takes for the state itself to have knowledge or beliefs. There is a deeper problem for neutrality here. On this standard, many religious and moral commitments will count as unreasonable and, therefore, unprotected. I think it will be hard to find moral views that are as unlikely as flat earth, but we need to look at the entire range of issues that the state decides on.

The state can legitimately decide on whether climate change and astrology are real. It decides that vaccines are safe and effective. In regulating the economy, the state presumes that more competition leads to a more efficient outcome. Maybe you disagree with some of the state's decisions on these matters. What matters is that the state should have the power to decide these issues one way or the other, even in the face of controversy. If you pursue this line, then you are saying that the state should be as neutral about morality and religion as it is about economics, medicine, or astronomy. This isn't very neutral at all.

To fix on a concrete example, the state should be allowed to decide that homeopathy and accupuncture are ineffective. These beliefs are important to people, and they build their life-plans around them. And yet there are moral and political views that are in a worse epistemic position than homeopathy. Obviously, any example will be controversial. I will offer my own examples, and you are free to find one friendly to you.

Chastity is a fine choice, but there is no intrinsic value in it. Homosexuality is morally permissible. It is more likely that homeopathy is effective than it is that masturbation is wrong. If there is a heaven, then most people get to go there: Hell is too horrific to be commonly used by a benevolent God. We, and whatever experts exist, know these things to be true. Any ethical system or religion that denies these is, by the standard we're considering, unreasonable.

Supposing the forward-looking reasons were on its side, the state could legitimately

make decisions that reflect these judgments. On this understanding of reasonableness, the state's acting in this way would be unobjectionable. Since it is unreasonable to believe that many people go to Hell, your interest in practicing a non-universalist religion is illegitimate. This should be a bad implication for a neutralist.

The second route points to the way the beliefs were formed. Conspriacy theorists weigh the evidence very differently from the rest of us. We disagree about which sources provide good evidence, and which theories provide the best explanation of the evidence. Reasonable people weigh the evidence in a certain way that (given the evidence we actually have) precludes belief in a flat earth and other conspiracies.

It is up to the neutralist to decide which methods/ways of weighing the evidence are in and which are out. The most obvious standard is that the method must be good enough. That is to say, a commitment warrants the protections of autonomy or respect only if it was arrived at by a method that is sufficiently reliable. Whatever the psychology of conspiracy theorists is exactly, they don't form their beliefs in a way that tracks the truth.

In the U.S., the Daubert standard (§702 of the Federal Rules of Evidence) encodes something like this. It requires that expert testimony be the product of reliable methods, and that the method is reliably applied to the facts of the case under consideration.

As before, it will be hard to argue that this is too high a standard without giving a controversial example. This is only compounded by the fact that it's hard to characterize the methods used in religious and moral investigation, and then rank them by reliability. I offer my own judgments about unreliable methods, and you are invited to substitute your own. When it comes to religious belief, we have good reason to think that direct revelation is not a reliable way of coming to learn the truth about God. Relatedly, neither is an attempt to infer ethical conclusions from religious texts. Committing the naturalistic fallacy is, of course, an unreliable way of getting at the moral truth. Though natural law theorists would deny that they commit the

naturalistic fallacy or that their conclusions are problematically religion-based, their arguments are close enough that I'm happy to call their methods unreliable.

Suppose the state officially denies any faith which depends on revelation, or any moral claims that are held only for one of the reasons above. If I am right about the reliability of these methods, then this action is unobjectionable on the view under considerations. These commitments are unreasonable, because they were formed via methods which don't track the truth of the matter. People's interests in living by these commitments is illegitimate, and all reasonable citizens would reject these commitments. Speaking for myself, I am friendly to policies like this, but we have gone a long way from what we wanted out of neutrality.

It is technically possible for the neutralist to accept all the arguments of this section, and still maintain their neutralism. Even with the much higher standards we considered here, denying Kantianism is still reasonable, and the interest in living a non-Kantian life is therefore legitimate. So the examples of bad policies that we started with are still bad. Still, the neutralist must say that it is unobjectionable for the state to deny many religious and moral views. The version of neutrality we end up with has much in common with Wall (2010)'s. He holds that the state may promote particular conceptions of the good. However, value pluralism means that multiple ways of life are, objectively, equally worthwhile (or incomparably good). The state may not decide between equally worthwhile conceptions of the good. But if one ideal is objectively better than another, the state may promote it. As Wall notes, his principle is very different from the neutrality principles of more orthodox liberals. As we saw, this "neutrality" isn't very neutral.

This concludes my discussion about reasonableness. Reasonableness provided hope for the neutralist. If conspiracy theories are unreasonable, then the concommitant interests are illegitimate. This could allow the neutralist to condemn the bad policies while upholding the good ones. I argued that the neutralist cannot make good on this line of thought. If the standards for reasonableness are low, then the

good policies are objectionable. If the standards are high, then the bad policies (or ones very much like them) are unobjectionable.

3.7 Conclusion: Non-Neutral Liberalism

Based on the arguments above, I conclude that we should reject neutralism. In §2, I challenged neutralists to explain what is in-principle objectionable about the bad policies without also condeming the good policies. After seeing the shortcomings of the harm principle and liberal neutrality in §3, I suggested that the neutralist identify a legitimate interest that is set back by the bad policies. If the good policies do not do the same, then the neutralist can answer the challenge. §4 considered two interests that are often implicated in discussions of neutrality. I argue that, while citizens' interest in autonomy and respect are set back by the bad policies, they are also set back by the good policies. Therefore, the neutralist must argue that these interests set back by the bad policies are illegitimate.

§5 raises two possibilities for delegitimizing the relevant interests; they are based on false or unimportant commitments. I explained why both these tactics fail. Finally, in §6, we saw that an appeal to the reasonableness of a commitment could explain why the good policies on set back illegitimate interests. I posed a dilemma for this view. If the standards for reasonableness are low, too many interests are legitimate. If the standards are high, too many are illegitimate. Either way, the neutralist cannot separate the bad from the good.

The neutralist cannot make good on their claim that there is an in-principle difference between the bad and the good policies. If we are going to be liberals, we should not be neutralists. The difference between the two slates of policies is forward-looking. When it comes to issues of morality and religion, the state has a terrible track record of making the right decision. Even if the state happened upon the right religion, we can expect civil unrest or even violence to arise in a theocracy. Similar

explanations show what's wrong with the bad policies. The good policies look great from a forward-looking point of view, so they are justified. If circumstances were different though, we might have to change our minds.

This brings us back to an issue first raised in §1. Can we be liberals if we only accept a forward-looking difference between the good and bad policies? Yes, there are versions of liberalism that reject neutralism. Here, I want to briefly make some comments about non-neutral liberalisms.

To start, giving up on an in-principle defense of neutrality doesn't mean you have to give up on all non-consequentialist commitments in politics. For example, if the forward-looking reasons were on our side, I believe the state would be justified in officially recognizing the truth of atheism. This might mean teaching classes in public school designed to convince children that there are no gods. But, even if criminalizing religion had forward-looking benefits, it might still be forbidden to do so. There might be a non-consequentialist prohibition on imprisoning somoene for actions that do not wrong another. Practicing a false religion does not wrong anyone else. We must also take into account anti-paternalist considerations.

So giving up on neutralism does mean, under certain circumstances, giving up on a complete separation of church and state. It does not mean giving up on all religious freedoms. Moreover, it is not as if the state is recognizing these freedoms just for forward-looking reasons. There can still be non-consequentialist constraints on state power.

On reflection, neutralism is a strong claim. It says that there are certain controversies where the state may not take sides at all. Non-neutralism lets us keep our options open. Depending on the merits of the case under consideration, it may be best for the state to pick sides or refrain from deciding. Non-neutralism, by itself, doesn't tell us what side the state should take. Even if you and I agree that the state should be able to make a decision in favor of *this* side rather than *that* one, we can still disagree about what the state can do with that decision.

For example, being of a more utilitarian mindset, I think that the state should do whatever will maximize utility. Someone like Raz (1986), on the other hand, will say that the state still must respect and promote autonomy. Raz, though, thinks that part of promoting autonomy means giving people access to genuinely valuable options, and this means deciding which options are valuable in the first place. We agree with each other that, in principle, the state can decide any moral matter that comes before it. We disagree about what the state should ultimately do.

What does this mean for policy? Not a whole lot, unfortunately. In deciding between policies, rejecting neutralism means rejecting arguments that start with "It's not the state's place to decide whether...". There may be a few areas where this is offered as a main argument against policy; we've discussed religious and moral education in school. To that, I would add sex education. Whatever parents might believe, it is the state's place to decide the facts and values surrounding intimate relationships, and what children should be taught about them. Unrelatedly, I would argue that the state should take a more active role in determining which charitable causes are actually worthwhile. It should offer incentives to people based on that decision.

For most policies, though, there are going to be lots of considerations to take into account aside from whether it is in principle the state's business to decide in the first place. Neutralism was supposed to be central to liberalism, but it's unnecessary. There are good forward-looking reasons to keep the state out of churches and bedrooms.

Chapter 4
Praiseworthiness (and Knowledge) from Falsehood

Here is a proposed necessary condition on praiseworthiness:

> **Good Reasons (Informal)**: If S's ϕing is praiseworthy, then S ϕed for good reasons.

Good Reasons is meant to account for cases where a person does the right thing for selfish reasons. Rescuing a drowning child in order to get a reward is not praiseworthy. In this paper, I will argue that (a prominent reading of) Good Reasons is false. Praiseworthy actions need not be motivated by good reasons. In order to perform a praiseworthy action, your reasons need to be "close enough" to good ones.

§1 is introductory. We will tighten up the informal statement of Good Reasons into something more precise. I also define some terms, state some more principles, and discuss how **Good Reasons** fits in with the existing moral and epistemological literature. Good Reasons is related to the epistemic principle Counter-Closure.[1] §2 presents structuarally analogous counterexamples to both Good Reasons and Counter-Closure. Authors have tried various moves to save these principles in light of the examples, which are presented in §3. In §4, I argue that these defenses fail. We must reject Counter-Closure and Good Reasons. §5 briefly sketches what a view without these principles could look like.

---

[1]See, variously, Arnold (2013), Ball & Blome-Tillmann (2014), Buford & Cloos (forthcoming), Coffman (2008), Fitelson (forthcoming), Littlejohn (2016), Montminy (2014), Luzzi (2014), Schnee (2015), Warfield (2005)

4.1 Introduction

This section is introductory. I give a formal statement and explanation of Good Reasons. I'll define some terms that I'll use, and present some other principles that are relevant to this issue.

4.1.1 Good Reasons

Good Reasons is imprecise. If we are going to argue about whether it's true, we'll need a better idea of what praiseworthiness is and what it means to ϕ for good (or bad) reasons.

I could not give a once-and-for-all definition of praiseworthiness. Between general principles and cases, I think we can get a good enough idea of what praiseworthy action is. Praiseworthy actions merit or deserve praise. We often have instrumental reason to praise someone insincerely, in order to encourage others to emulate them. Although their action was *worth praising*, it wasn't *praiseworthy*. There is supposed to be some non-instrumental reason to praise the praiseworthy.

Praiseworthy actions are supposed to have some extra value over right actions. To do what is right is one thing. But to act in a way that merits praise is more noble, or better. It reflects well on the agent who performed the action. A pattern of praiseworthy actions is part of the life of virtue; a tendency to do what is praiseworthy may be part of a good life. Performing praiseworthy actions might mean that you are deserving of material rewards, parallel to how blameworthy actions merit punishments.[2]

Here is an example of a praiseworthy and a not praiseworthy action.

---

[2]Markovits (2010) uses "morally worthy" instead of "praiseworthy". As she rightly points out, something can be praiseworthy for its non-moral qualities—a movie can be praiseworthy for its witty dialogue—and we mean to be talking about moral praise in particular. The point is well-taken. Throughout this paper, "praiseworthy" means morally praiseworthy.

> **Donation**: Petros has the opportunity to donate to charity. After researching his options, he finds that he could do the most good by donating to the Groat's Foundation. He is personally moved by the plight of Groat's sufferers. He'll have to give up on some personal luxuries, but he knows it is worth it to help these strangers.
>
> Abel has the opportunity to donate some money to charity. His peers are highly morally-motivated, and he wants to impress them. So he does some research and determines that a donation to the Groat's Foundation is the most effective place to put his money.
>
> Petros and Abel both go on to donate to the Groat's Foundation.

I claim that Petros is praiseworthy for his action of donating to the Groat's foundation. Although Abel performed the same action, he is not praiseworthy for it. Petros deserves a pat on the back for caring about people in need and helping them. Abel is trying to garner social capital. There's something icky about his attitude towards charity.

Moreover, the difference between Petros and Abel seems to be exactly what Good Reasons is trying to capture. Petros made a donation for good reasons—concern for others. Abel made his donation for bad reasons—concern for his own social standing. So this is a convient place to elaborate on that phrase.

Unpacking the idea of "ϕing for a good reason" requires us to explain what it means to ϕ for a reason, and the difference between a good and a bad reason. To cut to the chase, here is the official statement of Good Reasons that we will use in this paper.

> **Good Reasons (Formal)**: If S's ϕing is praiseworthy, then S's motivating reason for ϕing is a reason morally justifying ϕing.

S's motivating reason is supposed to capture what it means to act for a reason. Requiring that that reason morally justify ϕing is how we regiment the claim that it

is a good reason. I will explain each of these two more specific phrases, and then why I've chosen to formalize Good Reasons in this way.

First, a morally justifying reason. At its most basic, this is a proposition that sufficiently tells in favor of some course of action. "Tells in favor" can be explained in many ways, and I want to be neutral about the different theories. We could take favoring to be a brute relation between a proposition and an action. Or it could be that the proposition explains why the action is justified, or makes it justified. Lastly, it could be that a proposition tells in favor of an action when it is evidence that the action is justified. My arguments do not rely on deciding this issue.

As for sufficiency, when ϕ is morally justified, it should be at least permissible. It need not be obligatory: it is possible to be praiseorthy for performing a supererogatory action, even though it is not obligatory. I think it is possible to be praiseworthy for performing merely obligatory actions. A grocer who gives the correct change to their customers because they don't want to cheat their customers does what is merely obligatory. It is not supererogatory to avoid cheating people, yet they act with moral worth. On the other hand, Kant's grocer, who is honest only to stay in business, does what is obligatory but without any moral worth. At any rate, the important take away is that you cannot do what is wrong in a praiseworthy way.

According to orthodoxy, a justifying reason is not just any proposition. It must be a true proposition, a fact. I think this is the right view of reasons and it is orthodox,[3] I will continue to talk as if the justifying reasons mentioned in Good Reasons are facts, though this is worth keeping in mind. The counterexample to Good Reasons trades on the claim that justifying reasons are facts. As we'll see, one way to preserve Good Reasons is to allow for falsehoods to be justifying reasons. That possibility will be discussed later.

Moving on, motivating reasons are the reasons for which the agent acts as they

[3]Dancy (2000), Schroeder (2007), Alvarez (forthcoming). My main target, Markovits, also accepts it.

do. They can come very far apart from the moral reasons that bear on their decision. If I commit a murder in order to get an inheritance, my motivating reason is "Killing this person will make me a lot of money". This consideration explains and makes sense of my action, though it is clearly not a good moral reason to kill someone. Motivating reasons can be completely non-moral as well; when I go to the fridge, my motivating reason is "There is a drink in the fridge". But this is not a morally momentous decision. I am neither morally nor immorally motivated in acting as I do.

An agent need not occurrently and explicitly hold their motivating reasons in mind when they act. When I want to turn a car to the left, I turn the steering wheel in that direction. My motivating reason for turning the steering wheel this way is that it will turn the car that way. But I don't need to explicitly think to myself "Turning the steering wheel left turns the car left". This fact motivates me automatically and subconsciously.

Throughout, we will assume that a motivating reason is a proposition of some kind.[4] They can be thought of as premises in practical reasoning: the wealth-motivated murderer reasons "Killing this person will make me a lot of money. I want a lot of money. So I will kill this person." But this reading isn't essential to my argument.

Notice one more thing, implicit in Good Reasons. It assumes that S has a single motivating reason for ϕing. It refers to their motivating reason, and not one of their motivating reasons. We will interpret Good Reasons so that it only applies when the agent has a single motivating reason for acting as they do. We set mixed motive cases aside. Presumably, when agents have multiple motivating reasons, praiseworthiness requires that at least one of these reasons is morally justifying.

Unfortunately, we cannot yet specify when an agent acts for a single motivating

[4]If, instead, they are beliefs, then Good Reasons would need to be restated. It would require that the content of the motivating reason coincides with a moral reason. The No Falsehoods principles below would have to be revised accordingly (the content of the motivating reason must be true). All of my arguments can be translated into that language, and I will not pursue the issue further here.

reason. This is part of what is at stake in the debate over Good Reasons, and to try to give an account here would unfairly bias us towards or against Good Reasons. Positing multiple motivating reasons can defuse the counterexamples, but comes with its own costs (§4).

I've chosen to formalize Good Reasons in a particular way. We started with a simple idea, of ϕing for good reasons. We ended up with something much more complicated. I am going to be arguing against the official statement of Good Reasons. Moreover, because I think I have formalized the informal principle correctly, I take myself to be arguing against the informal version of Good Reasons as well.

My formalization is not a straw target, chosen because it is easy to argue against. It is implied by Markovits (2010)'s coincident reasons thesis:

> S's ϕing is praiseworthy iff S's motivating reason for ϕing is a reason morally justifying ϕing.

As we can see, Good Reasons is just one half of the CRT.[5] Throughout this paper, we will have occasion to consider whether I have formalized Good Reasons fairly. I will argue that I have.

4.1.2 No Falsehoods

From the above discussion, we can see that Good Reasons implies

> **Moral No Falsehoods**: If (i) S's ϕing is praiseworthy, and (ii) S's motivating reason for ϕing is p, then p

Good Reasons and the orthodox picture of justifying reasons mean that praiseworthiness is incompatible with acting on the basis of falsehoods. Agents who perform praiseworthy actions must be motivated by morally justifying reasons. And morally

[5]See also Arpaly (2002), Arpaly & Schroeder (2014) for a related view.

justifying reasons must be facts. Therefore, agents who perform praiseworthy actions must be motivated by facts. Compare Moral No Falsehoods to

> **Epistemic No Falsehoods**: If (i) S knows that q by way of inference, and (ii) the basis of S's inference to q is that p, then p.

We can give a similar argument for Epistemic No Falsehoods. In order to gain knowledge of q by way of inference, S must base their inference on a bit of evidence. And if a proposition is part of an agent's evidence, it must be a fact. Therefore, agents who come to know inferentially must base their inferences on facts. The two No Falsehoods principles are analogous in that similar arguments motivate each.

When I say that S infers q on the basis of p, I mean for this to be underspecified. I do not have an account of which mental processes are inferential, and what are the bases of an inference. Our examples will involve deductive inference, and so can skirt the issue of which processes are inferential; deduction is inferential. In that case, we can take the basis of an inference to be the premises of the deduction.

As in the moral case, I do not think that all inferences need to be made explicitly and consciously, or that their premises need to be occurrently believed by the agent. Presumably, we often make inferences automatically and subconsciously. I will not try to settle this issue here. I think that Schroeder (2007) gives a good model for how to think about the premises an agent uses in an inference: they are the agent's motivating reasons for believing as they do. However, none of my arguments rely on Schroeder's account.

Of course, it is much less obvious that evidence must consist of facts than it is that a morally justifying reason is a fact. This weakens the argument for Epistemic No Falsehoods. We can say a few things in its favor, though. First, it is implied by

> **Counter-Closure**: If (i) S knows that q by way of inference, and (ii) the basis of S's inference to q is that p, then S knows p.

This seems to just push things back a step, since we need an argument for Counter-Closure. But Counter-Closure, in turn, is motivated by the claim that inferential knowledge must be based on evidence, together with E=K, which holds that an agent's evidence consists of all and only the propositions that they know. E=K is considered to be a central part of the knowledge-first program in epistemology. Much of the interest in Epistemic No Falsehoods and Counter-Closure arises because they are thought to be central to knowledge-first epistemology.[6]

Aside from its theoretical role, Epistemic No Falsehoods can help to explain cases like

> **Field**: Brittney looks into a field, and sees what appears to be a sheep. She is a very reliable detector of sheep, and the object is remarkably ovine. She comes to believe that the object is a sheep, and on that basis infers that there is animal in the field.
>
> In fact, the object is a rock. Luckily, there is a dog sitting behind the rock.

Brittney's belief is well-justified. She has every reason to believe that the object she is looking at is a sheep, and so that there is an animal in the field. Moreover, it is true. A dog is an animal. But she is Gettiered and doesn't know that there is an animal in the field. Epistemic No Falsehoods gives a tidy explanation for this fact. She infers that there is an animal in the field (q) on the basis of her belief that there is a sheep in the field (p). Epistemic No Falsehoods implies that if she knows q, then p is true. But p is false. Therefore, she does not know q.[7] Counter-Closure gives a parallel explanation of why Brittney lacks knowledge.

---

[6]Littlejohn (2016) is rare in arguing that those who accept E=K need not accept Counter-Closure.

[7]This example brings to mind the "No False Lemmas" proposal to solve the Gettier problem, and No False Lemmas is very similar to Epistemic No Falsehoods. The difference is that Epistemic No Falsehoods is not intended to be part of an analysis or sufficient condition for knowledge; it is merely a necessary condition and may not be essential to knowledge. I use the name "Falsehood" rather than "False Lemma" because it is not obvious that we should understand false beliefs in practical reasoning as lemmas.

Additionally, as with the moral principles, Epistemic No Falsehoods applies only to cases where the belief that q is based *solely* on the belief that p. In a modified version of **Field**, Brittney sees both the rock and the dog. She now has two routes to the belief that there is an animal in the field. She can reason "That object is a sheep, so there is an animal in the field" and "That object is a dog, so there is an animal in the field". In this case, it is more plausible that Brittney knows that there is an animal in the field. This is consistent with Epistemic No Falsehoods. Brittney does not infer the presence of an animal solely on the basis of a sheep's presence, but also a dog's presence.

So, the two No Falsehoods principles can be motivated in similar ways. One follows from the claim that praiseworthy action is motivated by moral reasons, and that moral reasons are facts. The other follows from the claim that inferential knowledge is based on evidence, and that evidence consists of facts. They are both tied to important theories in ethics and epistemology; the CRT on the one hand, Counter-Closure on the other. There is a further way in which they are analogous. There are structurally analogous counterexamples to both, where agents are moved by falsehoods that are "close enough" to truths.

This may not point to any deep analogy between the two, but it does justify treating them in the same paper. I believe it also motivates the search for a unified solution to this problem; parallel accounts for both knowledge and praiseworthy action. But that is an issue for a different paper.[8] For convenience, I will sometimes discuss only one of the No Falsehoods principles or talk about only one type of case. This is only to avoid reptition; unless otherwise noted, I intend the arguments I make to apply to both the moral and the epistemic domains equally.

Considering Epistemic No Falsehoods gives us an in to another view of praiseworthiness. Paulina Sliwa (2016) claims that ϕing is praiseworthy only if the agent's

[8]Mantel (2013) presents a partial account of praiseworthy action closely modelled on one for knowledge. I do not think she solves the problem of this paper, though.

motivating reason is "ϕing is morally right". If Sliwa accepts Epistemic No Falsehoods, she will be vulnerable to the counterexample I present below. This will be discussed in more detail later.

If we reject the Moral No Falsehoods, we have two choices. Since Good Reasons and the conception of reasons as facts imply Moral No Falsehood, we must reject one or the other principle. Since I am holding "reasons are facts" fixed, this means rejecting Moral No Falsehoods. We face the same choice when it comes to Epistemic No Falsehoods. If you reject Epistemic No Falsehoods, then you must either reject the view of evidence as facts, or the requirement that inferential knowledge be based on evidence.

Epistemologists typically take it that, if we reject Epistemic No Falsehoods and Counter-Closure, we should believe that evidence may consist of falsehoods.[9] In other cases, it is less clear which of these two claims they deny.[10] Explicit denial that inferential knowledge must be based on evidence is comparatively rarer (see Littlejohn 2016). But I will not try to settle the issue. We are, in the first instance, concerned with arguing against the No Falsehoods principles and Good Reasons.

4.2 The Counterexamples

In the last section, we saw two plausible principles that bar falsehoods from playing a certain role in an agent's beliefs and actions. According to Epistemic No Falsehoods, knowledge cannot be gained by inference from a falsehood. According to Moral No Falsehoods, praiseworthy action cannot be motivated by falsehood. But there are powerful counterexamples to these principles, as this section shows.

Contrary to the No Falsehoods principle, falsehoods can play a motivating role consistently with agent's possessing knowledge or performing praiseworthy action.

[9]See, for instance, Fitelson (forthcoming) and Arnold (2013).

[10]I think Klein (2008) could be read in either way.

Consider first

> **Field 2**: Caliban is an expert agricultural scientist. As part of a study, he must count the sheep in a field. There are a lot of sheep, and they are huddled together, but Caliban has an excellent track record at counting sheep. He counts 52 sheep, and on that basis comes to believe that there are less than 60 sheep in the field.
>
> In fact, there are 53 sheep. A little baby sheep was standing behind a much larger sheep.

We also have

> **Donation 2**: Dor has the opportunity to make a large donation to a charity. She wants to make sure her donation goes to the most effective charity out there. She spends hours researching her various options, and comes to believe that a donation to the Groat's Foundation will save 52 lives. This is more than any competing charity could save with her money. In order to save 52 lives, Dor makes a donation to the Groat's Foundation.
>
> In fact, the information she receieved was slightly off. The Groat's Foundation will actually save 53 lives.

Cases like **Field 2** are familiar from the literature, and so is the problem they pose for Epistemic No Falsehoods. Intuitively, Caliban knows that there are less than 60 sheep in the field (q). However, he has inferred this belief from a falsehood. There aren't 52 sheep in the field (p), there are 53. Epistemic No Falsehoods incorrectly predicts that Caliban does not know that there are less than 60 sheep, because the basis of his inference is false.

Donation 2 poses a similar problem for Moral No Falsehoods. Intuitively, Dor's action is praiseworthy. Her donation (ϕ), and the moral motivation that it expresses, is entirely commendable. But "The donation will save 52 lives" (p), her motivating

reason, is false. Her donation will save 53 lives. Since she is motivated by a falsehood, Moral No Falsehoods incorrectly predicts that her action is not praiseworthy. We can also see this false prediction directly from Good Reasons: since p is false, it is not a morally justifying reason. So Dor's motivating reason is not a morally justifying reason, and her action is not praiseworthy according to the CRT.

Examples like **Donation 2** have mostly escaped attention in the literature.[11] The CRT is, I think, our best available theory of praiseworthy action. We either need a replacement account, or need to find some way to defend it from **Donation 2**.

Donation 2 also serves as a counterexample to the conjunction of Sliwa's view of praiseworthiness and Epistemic No Falsehoods. We can imagine Dor inferring that donating to the Groat's Foundation is right because it will save 52 lives. By Epistemic No Falsehoods, she does not know that her action is right. So, on Sliwa's view, her action is not praiseworthy. Even though Sliwa does not accept the CRT, she still needs to take a stand on the issues here. She thinks knowledge of an action's rightness is necessary for it to be praiseworthy. She faces the same choice that epistemologists do: either accept that knowledge need not be based on evidence, or that evidence need not consist of facts.

It would be premature to look for replacements for Good Reasons and Counter-Closure until we are sure that the No Falsehoods principles are well and truly counter-exampled. In the rest of this paper, we will see the options that theorists have for defeinding them. I argue that these fail. We must reject Epistemic and Moral No Falsehoods. For now, we can start to get a feel for what a defense would look like.

A first attempt would be to deny the intuition that Caliban has knowledge, and that Dor's action is praiseworthy. Schnee (2015) pursues a strategy like this. His argument for this has two prongs. Firstly, he argues that only an account that accepts Epistemic No Falsehoods can avoid particular Gettier examples. Secondly, there are

[11] There are short passages in Stratton-Lake (2000) and Markovits (2010), but that's it as far as I've been able to tell.

cases where agents intuitively don't possess knowledge, but (he claims) are relevantly similar to **Field 2**. For the most part, though, authors in the literature agree that Caliban has knowledge, even if they go on to defend Epistemic No Falsehoods.

I side with the literature: Caliban knows that there are less than 60 sheep in the field. His belief is well-justified, and not true just by luck. I do not have the space to fully address Schnee's arguments here. Responding to the first prong would require offering a positive theory of inferential knowledge, and showing that it avoids Gettier cases. Responding to the second would involve going through his various cases, and either showing why his intuitions are wrong, or what the relevant difference is. At least, we should note that whatever plausibility Schnee's move has for **Field 2**, it looks pretty bad for **Donation 2**. Dor does the right thing, and even if her reasons aren't exactly right, they are right enough. It would be bizarrely harsh to judge that her action was morally worthless just because she made a slight calculation error.

A better response to these examples should target the claim that they are inconsistent with the No Falsehoods principles. I claimed that Caliban bases his inference on "There are 52 sheep in the field", and Dor's motivating reason is "The donation will save 52 lives". But I did not offer any argument for these claims. It is open to defenders of No Falsehoods to argue that these claims are false: the basis or the motivating reason is some other, true proposition. Or they could claim that these are not the *sole* motivators. In addition to the false basis Caliban uses, there is another true basis that is not immediately apparent. In either case, the theorist is positing *proxies*.[12] They claim that we were mistaken in our initial assessment of the cases. Those aren't the relevant propositions to slot into the No Falsehoods principles, rather these other proxy propositions are the relevant ones.

This is the best hope for defending the No Falsehoods principles, and their stronger versions as embodied in Counter-Closure and Good Reasons. In the next section, we turn to the proposals that authors have made in the literature in their defense. I will

[12]I take this terminology from Luzzi (2014)

argue in §4 that these defenses are more trouble than they're worth.

4.3 Proxies

In this section, I will present the moves that authors have made in order to defend the No Falsehoods principles from counterexamples like **Field 2** and **Donation 2**. In the next section, I argue that they fail; either they do not account for similar cases, or they raise further problems. As discussed in the previous section, the best hope for the No Falsehoods principles is to posit proxy propositions. A theorist could say that in **Field 2**, for instance, "There are 52 sheep in the field" (p) is not the basis on which Caliban infers "There are less than 60 sheep in the field" (q). Caliban's basis consists of some other proposition p^*. The proxy p^* could be meant to replace p entirely as a basis. In this case, p^* had better be true, or else it won't help avoid the counterexample to Epistemic No Falsehoods.

An alternative position is for theorists to claim that the proxy goes alongside p as a basis, to supplement it. In that case, there is no sole basis to the inference. There are the two bases p and p^*. This means the antecedent of Epistemic No Falsehoods is not satisfied, and so it implies nothing about the case. In principle, it is consistent with Epistemic No Falsehoods that both p and p^* are false.

I will follow the general tenor of the literature, mostly considering the proxy p^* as a replacement for the apparent orginal basis/motivating reason p. In the next section, I will argue that it doesn't help matters to posit proxies as supplements, whether they are true or false.

Defenses of the No Falsehoods principles are distinguished by the relationship that they say holds between p and p^*. I'll discuss four initially plausible proxy views, and argue that they fail. The first view says that the motivating reason is actually "approximately p". The second view tries to generalize beyond "approximation" facts, and lets in other nearby claims. Notably, my arguments against these two

proposals only shows that they are bad defenses of Epistemic No Falsehoods. It is only when we look at the last two proposals that the general problems for proxy views, epistemic or moral, become apparent. The third view says that p^* is whatever background evidence the agent has to believe that p in the first place. The last view says that the proxy is "I [the agent] believe that p".

A last preliminary remark, if proxy views are to save the No Falsehoods principles, they must claim that in the bad case agents really are motivated by the appropriate proxy p^*, or that it really is the basis of their inference.[13] It is not enough to point out that agent knows p^*, or would be motivated by p^* if they weren't motivated by p, or wouldn't believe as they do if they didn't believe that p^*. If the agent's sole motivating reason is p, all these other considerations are irrelevant. The No Falsehoods principles imply that the agents don't know or that their action is not praiseworthy.

Now, we might want to weaken the No Falsehoods principles to say that agents who gain inferential knowledge must be *disposed* to use facts as their bases, or that they *would* use a justifying reason as their motivating reason if certain other conditions obtain. I think some such condition along these lines is probably right. But that would involve denying the No Falsehoods principles and Good Reasons/Counter-Closure, and replacing them with something else. These principles are, I think, close to important truths. If we reinterpret Good Reasons along these lines, we can save the informal thought from the counterexamples.

4.3.1 Approximation

On this first view, suggested by Ball & Blome-Tillmann (2014),[14] Caliban's basis for inferring that there are less than 60 sheep is that there are *approximately* 52 sheep, rather than that there are 52. Similarly, we can extend their line of reasoning to

[13]Ball & Blome-Tillmann (2014, 3-4) and Coffman (2008) make the same point.

[14]Their considered view is not tied to approximation in this way, but it is a natural response to the counterexamples.

say that Dor's motivating reason for donating is that approximately 52 lives will be saved.

The approximation view is attractive.[15] In our cases, "approximately p" is a claim that agents know and is plausibly a good reason to believe/act as they do. However, such views falter when agents are trying to infer proxies themselves. Consider the case below:[16]

> **Marbles**: A horde of marbles rolls past Ged, who diligently counts them off. He arrives at 52, and comes to believe that's how many marbles there are. Having learned earlier that day about the relationship between approximation and precision, he infers "There are approximately 52" marbles. As it happens, there were 53 marbles.

Intuitively, Ged can come to know that there are approximately 52 marbles in this way. However, the basis for Ged's inference that there are approximately 52 marbles is that there are exactly 52. This will not do for defenders of Epistemic No Falsehoods, since "there are 52" is false. Claiming that Ged's basis for inferring that there are approximately 52 marbles is "There are approximately 52 marbles" is a very tight, vicious circle.

4.3.2 Other "Downstream" Proxies

Approximation clearly doesn't work in some cases of knowledge from falsehood, but it may in others. Ball & Blome-Tillmann (2014) and Montminy (2014) offer a general account of (apparent) knowledge from falsehood

> In apparent cases of [knowing q on the basis of a falsehood p], there are two true propositions t and p^* such that:

---

[15] At least for cases where p is a numerical claim. "Approximately p" may not always make sense.

[16] The argument targetting approximation views in this way is taken from Luzzi (2014).

1. t evidentially supports both p and p^* for S;
2. p^* is entailed by p;
3. S knows both t and p^*;
4. S's belief that q is properly based on her knowledge that p^*.[17]

Here, p^* is the proxy, and t is the evidence on the basis of which S believes p. In Caliban's case, t would be his count of the sheep. Ball and Blome-Tillmann's proposal identifies the proxy as something causally and evidentially "downstream" from t, in the sense that the agent's basis for believing p^* is t, and t serves as evidence for p^*. For Caliban, p^* can be identified with "there are approximately 52 sheep".

However, it seems we can extend Luzzi (2014)'s argument to show that no such "downstream" proxy can do the work needed. Since p entails p^*, it looks like we can always cook up a case like **Marbles**, where the agent apparently infers p^* on the basis of a false p. These downstream proxies replace one case of knowledge from falsehood with another. Agents in **Marbles**-like cases will need another proxy p^{**} to believe the proxy p^*.

For example, in **Marbles**, I claimed that Ged inferred "There are approximately 52 marbles" (p^*) from "There are exactly 52 marbles" (p). The proxy theorist could claim the real basis of Ged's inference is "There are between 49 and 55 marbles" (p^{**}). In response, I modify the case slightly. In the new case, Ged (apparently) infers p^* from p^{**}. But, in turn, p^{**} is (apparently) inferred from the falsehood p. So we need yet another proxy for this modified **Marbles**-case.

Now, Ball and Blome-Tillmann do not give a recipe for determining what the proxy is, so it is possible that they could come up with a view that is immune to **Marbles**-like cases. Or they could give up on having a systematic view, and say that the appropriate proxy is different in different cases: it is the "approximate fact" in

[17]Ball & Blome-Tillmann (2014, 5). I have renamed variables to fit my usage. Montminy accepts a slightly different view, but the differences do not matter for our purposes.

our original cases, but it is some quite different fact p^* in **Marbles**, and some yet further fact p^{**} in the **Marbles**-like cases we cook up for p^*, etc.

This is not a regress! In each case, there is a single proxy that gets an agent to believe the proposition in question. But this loss of systematicity is unfortunate. It would be better if we could have a more systematic view. In all such cases, the proxy is going to have to be appropriately related to the background evidence (t) that S has for believing p. **Marbles**-like cases seem to show it can't be anything other than the background evidence itself. Let's consider such a view.

4.3.3 Background Evidence

The third strategy we'll look at identifies the agent's background evidence for believing p as the proxy. It is the background evidence that forms the basis of the inference to q, or is the motivating reason that the agent ϕs.[18] So in **Donation 2**, Dor has some evidence for her belief that the Groat's Foundation will save 52 lives. This is the research that she did into the charity. The suggestion is that her motivating reason for making the donation is the facts that she discovered over the course of his research. Similarly, Caliban's motivating reason for believing that there are 52 sheep is something like "I (seem to?) have counted 52 sheep". On this proposal, his basis for inferring there are less than 60 sheep is also "I have counted 52".

I have two problems for this view, that also affect the fourth proxy view. In fact, I will argue that they are problems for all proxy views. Rather than repeat myself, let's get the fourth view on the table.

---

[18]Montminy (2014, fn. 13) may be sympathetic to a view of this type. Markovits (2010, 26–28) suggests a response along "background evidence" lines, but most of her discussion focusses on the "I believe" view of the next section.

4.3.4 Internalization

The fourth strategy "internalizes" the apparent motivating reason to get a proxy that is true, and so consistent with the No Falsehoods principles. In discussing a case of Stratton-Lake (2000)'s, Markovits (2010, 25–30) suggests that, in cases like **Donation 2**, Dor's motivating reason is not the false "My donation would save 52 lives". Instead, it is "I believe that my donation would save 52 lives".

Similarly, we can say that in **Field 2**, the basis for Caliban's inference is "I believe there are 52 sheep in the field". On that basis, he infers that there are less than 60 sheep. These internalized proxies are true! Dor does believe that her donation would save 52 lives. If this is her motivating reason, then her action's being praiseworthy is consistent with Moral No Falsehoods.

Now that we have the views on the table, we can argue against them.

4.4 Two General Problems for Proxies

Here, I present two problems that affect all proxy views. Neither is iron-clad, but each poses a difficult challenge to defenders of the No Falsehoods principles. The first argument is a dilemma: proxy views are either disjunctive, or they imply that our actions and beliefs are radically and routinely over-determined. The second argument is that proxy views undermine the original motivation for the No Falsehoods principles.

4.4.1 A Dilemma: Disjunctivism or Over-Determination

Here, I will present my first argument against proxy views. They are committed to either an objectionable disjunctivism or an objectionable over-determination claim.

To illustrate the problem, it will be helpful to have some concrete proxy motivating reasons. In line with the background evidence proposal, let's say that Caliban's basis for inferring that there are less than 60 sheep in the field is whatever background

evidence (t) he has, on the basis of which he believes that there are 52 sheep (p). From that, he infers that there are less than 60 sheep. In line with the internalization move, Dor's motivating reason is "I believe my donation would save 52 lives". This is what actually motivates her donation (ϕing).

In assessing these views, it will be helpful to compare how they treat the "bad cases" **Field 2** and **Donation 2** with some "good cases":

> **Field 3**: Enbar is an expert agricultural scientist, in a situation very similar to Caliban's. She is just as competent a counter. In her case, there is no little sheep behind a big sheep. She correctly counts 52 sheep and concludes that there are less than 60 sheep in the field.
>
> **Donation 3**: Faisal has an opportunity to donate that is very similar to Dor's. Through no special effort of his own, he gets information that is slightly more accurate than that available to Dor. He correctly concludes that his donation would save 52 lives, and on that basis decides to donate to the Groat's Foundation.

Caliban and Enbar are very similar, as are Dor and Faisal. Caliban and Enbar both believe that there are 52 sheep in the field, and conclude from that there are less than 60 sheep. However, Enbar's belief about the exact number of sheep is true. Her knowledge of the upper bound on sheep is not threatened by Epistemic No Falsehoods. We can say the same thing abut Dor and Faisal. They both believe that their donation will save 52 lives. But it is consistent with Moral No Falsehoods that "My donation will save 52 lives" is Faisal's motivating reason.

Enbar has just the same background evidence as Caliban does. She believes (and knows) t, as he does. And t plays the same role psychologically for each of them. Each believes p on the basis of t, and goes on to infer q. If either were to get solid evidence against t (e.g., some convincing evidence that they have miscounted), they would drop their belief in p and their belief in q. We can say the same thing about

Dor and Faisal. Both know that they believe their donation would save 52 lives. If they got some convincing evidence that their donation would not save 52 lives, the proposition about their beliefs would become false. Presumably, they'd reconsider their action in light of this new evidence.

These observations point to a general fact about proxies. Whatever proxy we identify must be usable as a motivating reason or basis of inference in bad cases like Caliban's and Dor's. However, when an agent is in a bad case, they cannot know they are not in the good case. Whatever proxy we posit for bad cases will also be available for agents in good cases. If t is the basis for Caliban's inference, why is it not also the basis for Enbar's? What makes it the case that Faisal is motivated by "My donation will save 52 lives", rather than "I believe my donation will save 52 lives."?

If the defender of No Falsehoods insists Enbar infers "There are less than 60 sheep" on the basis of "There are 52 sheep", rather than t, they are led to an implausible disjunctivism. This is the first horn of the dilemma. Agents' motivating reasons and bases for inference will differ between good and bad cases in a way that is not plausibly reflected in their psychologies. As we saw, Dor and Faisal on the one hand and Caliban and Enbar on the other are disposed to act in just the same ways. They'll cite the same propositions in defense of their views and respond to counter-evidence in just the same way. To put the point colorfully, Dor does not know she is in a bad case. But her motivating reason "knows" that it can't be the false p, so instead it "becomes" the true proxy t or "I believe that p".

So, if agents have different motivating reasons/bases of inference in good and bad cases, we are led into disjunctivism. Alternatively, the proxy theorist could deny disjunctivism. Since the proxy plays the same role in agents' psychologies in both good and bad cases, the proxy serves as a motivating reason in both cases. On this horn of the dilemma, agents' actions and beliefs will be routinely overdetermined. Faisal does not have the single motivating reason "My donation would save 52 lives". He is also motivated by "I believe my donation would save 52 lives". Notice that either

reason, on its own, would be sufficiently motivating. When agents find themselves in good cases, there are going to be many more motivating reasons than we initially thought.[19]

We can make this overdetermination look more implausible by imagining more complicated cases. Suppose that an agent recognizes two considerations, p_1 and p_2. The agent has been very conscientious in coming to believe that both are true. In their estimation, neither is a good enough reason to ϕ on its own. Taken together, though, the agent thinks there is a strong reason to ϕ. Initially, we might be tempted to say that their motivating reason is $p_1 \& p_2$. However, to avoid counterexamples like the one above, we have to recognize that "I believe that $p_1 \& p_2$" is their motivating reason, as is p_1&I believe that p_2", and "I believe that p_1&I believe that p_2". Where we started out with one motivating reason, we end up with four distinct motivating reasons any one of which would be sufficient to explain the agent's ϕing.

Additionally, if we go in for overdetermination, then the No Falsehoods and Good Reasons principles are toothless. As discussed above, they only apply when an agent has a single motivating reason/basis of inference. If you are impressed enough by these principles that you buy into massive overdetermination in order to save them, you probably want a principle that is going to allow you to distinguish between cases. This is not to say there aren't nearby principles that do the job, but Good Reasons and No Falsehoods probably aren't them.[20]

My claim is not that actions can never be overdetermined by motivating reasons.

---

[19]Montminy (2014, 471–472) accepts that Caliban's belief is overdetermined in the bad case. Presumably, he would accept that it is overdetermined in the good case as well. (Ball & Blome-Tillmann 2014, 3–4)'s test for identifying motivating reasons seems to imply that there is overdetermination in the good case.

[20]I think it would be natural to require that at least one of the motivating reasons is a justifying reason or fact. I think this view leaves an important question unanswered. In a case like **Donation 2**, the first motivating reason that comes to mind for Dor are "Her donation will save 52 lives", which is not a justifying reason. Why do we immediately describe her action in a way that looks inconsistent with Moral No Falsehoods and Good Reasons? Overdeterminers owe us a story about why examples like these have been so gripping to philosophers.

My concern is that, on this view, actions are nearly always overdetermined. Whenever an agent is moved by p, they will also be moved by their belief that p, or the background evidence they used to arrive at p, or whatever the appropriate proxy would be. This concludes my first general objection to proxy views. They face a dilemma: either disjunctivism or massive overdetermination. We may find ways to live with either horn of this dilemma. But there is a second general problem with proxy views; they undermine the motivation for the No Falsehoods principles.

4.4.2 Self-Undermining

In this section, I explain my second general objection to proxy views. I argue that a proxy view undermines the original motivation behind the No Falsehoods principles as well as Good Reasons. This doesn't mean that the principles are false, but it should give us pause when saving the letter of a view requires us to sacrifice its spirit.

Consider Good Reasons. It requires a perfect coincidence between an agent's motivating reasons and the reasons justifying their action. This is motivated by the observation that, when agents' motivating reasons come far apart from the justifying reasons, their actions are not praiseworthy. In **Donation**, Abel is not praiseworthy because his motvating reason is not at all morally significant. He has done a terrible job of determining which considerations are relevant to his decision, and Good Reasons rightly condemns him for it.

I think it is important not just that agents take into account all and only relevant factors when acting or believing. They also need to take these considerations into account in the right way. Even when agents' motivating reasons do line up with the justifying reasons, they can still fail to act in a praiseworthy way or gain knowledge. Such examples are familiar from epistemology, since they illustrate the difference between doxastic and propositional justification. p might support an inference to the best explanation to q, and so someone who knows p could come to know q by inferring that it is the best explanation. However, seeing that q entails p, an agent might affirm

the consequent and come to believe q. Even though p is great evidence for q, agents who fail to take into account the way in which p supports q do not come to know q.

We see something similar in the moral case. Consider

> **Donation 4**: Hilde finds an extra penny on the ground, and is presented with a donation box for the Groat's Foundation. She remembers reading earlier that day that such a donation would save a million lives.
>
> Hlide carefully considers whether a penny is worth a million lives. She ultimately decides that it is, but just barely. If the penny would save only 999,999, then she would keep it for herself.

I claim that Hilde's action is not praiseworthy. Her action doesn't merit praise or esteem, or have any of the other characteristic signs of praiseworthy action.[21] Donating is the right thing, and her reason for doing it (that it would save a million lives) is a good reason to donate.

But Hilde has a very strong obligation to give up her penny in **Donation 4**. She treats the million lives as if they provide a very weak, easily defeasible reason. Failing to be motivated by the right reason *in the right way* stands in the way of praiseworthiness. So, in epistemology and in ethics, it is not good enough that you're moved by good reasons or good evidence. In order to gain knowledge or perform praiseworthy action, you must be moved to act or believe in the right way. In epistemology, "the right way" comes to something in the realm of doxastic justification. I am claiming that there is something similar when it comes to praise.

What does this have to do with the No Falsehoods principles? Since Caliban knows and Dor's action is praiseworthy, they must have been moved to do as they did in the right way. But, if the "background evidence" explanation of Caliban's

[21] Markovits (2010, 213–5) argues that counterfactuals about an agent cannot determine whether their action was praiseworthy. I think her arguments to this effect fail, but we do not need to settle the point here. I am directly relying on the intuition that Hilde's action is not praiseworthy. I'm using the counterfactual just to illustrate that Hilde has inappropriate moral concern.

knowledge is correct, then his motivating reason for believing that (q) there are less than 60 sheep in the field is (t) the background evidence he had to believe that (p) there were 52. To be sure, t is very strong evidence for q. In general, it will be better evidence for q than it is for p. However, t will typically be ampliative evidence for q, the way it was for p. At least in cases like **Field 2**, agents can't properly deduce either p or q from t.

But Caliban doesn't know he's in the bad case. He looks to all the world like he's performing a deduction from p to q. The proxy theorist claims he is actually inferring from t to q. This inference has been performed in the right way, but there is no right way to *deduce* q from t. Evidently, even though Caliban performs a deduction where he shouldn't have, he performed his inference in the right way. This means that, on the background evidence view, perfect coincidence between the way t supports q and the way Caliban gets from t to q is not required for knowing by inference.

We get a similar result on the internalization view and when considering praiseworthiness. Very often, when p implies q, "I believe that p" will not imply q. So agents deduce when they shouldn't. When p is a reason for Dor to ϕ, evidence for p or Dor's belief that p may also be reasons for her to ϕ. In general, though, they will be weaker reasons. If we use a proxy view to save Good Reasons or Moral No Falsehoods, praiseworthy action must be compatible with an agent's overestimating the strength of their reasons.

On Good Reasons, making a mistake of fact is disqualifying for praiseworthiness, but making a mistake of value is compatible with it. Counter-Closure requires that agents get the evidence right, but allows that they go on to misuse it. It is odd for a view to treat these two features so differently. The reasons that motivate you and the way you are motivated by them go hand-in-hand. A view which handles them the same way would be better.

Apparently, agents must do a good enough job of being moved by the relevant considerations. Although Caliban and Dor don't get things exactly right, they do a

good enough job of being motivated in the right way. Our consequent-affirmer and Hilde do a poor job of being properly motivated, even though they are motivated by good reasons. We don't require perfect coincidence between the way an agent is motivated and the way their action or belief is supported. Yet proxy views require perfect coincidence between motivating reason and justifying reason. In order to account for **Donation 2** and **Field 2**, proxy views make commitments that undermine the motivation for the perfect coincidence enshrined in the No Falsehoods princiles.

To summarize my second general argument against proxy views: Proxy views must allow agents some leeway in moving from the proxy to their belief or action. Caliban performs a deduction, where the proxy doesn't support one. So they cannot say that any sort of mistake deprives an agent of knowledge or praiseworthiness. At the same time, a related mistake—inferring from falsehood or being motivated by a non-reason—is incompatible with knowledge or praiseworthiness. A better view would treat them similarly.

In this section, I argued that we should reject the No Falsehoods principles, and so Counter-Closure and Good Reasons. In order to defend the principles from the counterexamples, these views must posit proxies (§3): un-apparent motivating reasons/bases of inference that coincide with justifying reasons/evidence. In general, all proxy views will face two problems. First, a dilemma. They could be disjunctive, and treat the proxies as motivating in bad cases but not good cases. Or they could go in for massive overdetermination of thought and action, allowing that both the proxy and the apparent motivator are at work in good cases. Second, proxy views partially undermine themselves. They require agents to perform perfectly with regards to picking out reasons/evidence, but must allow for mistakes in using the reasons/evidence. A better view allows agents to make small mistakes of either kind.

4.5 Living with Falsehood

This section is conclsuory. First, I will recount the argument of this paper. Next, I give the lay of the land, assuming my arguments against the No Falsehoods principles succeed. To end, I sketch the main features of a view that rejects the No Falsehoods principles.

Good Reasons says that praiseworthy action is motivated by good reasons. But an orthodox view of reasons holds that they are facts. So, in §1, I argued that Good Reasons implies Moral No Falsehoods: praiseworthy action is motivated by facts. A similar line of argument supports Epistemic No Falsehoods: inferential knowledge is based on facts. In §2, I presented the counterexamples **Field 2** and **Donation 2** to the No Falsehoods principles. They show that knowledge and praiseworthiness can come from falsehood. Given that I accept the orthodox view of reasons, I take this to be an argument against Good Reasons. Given that I accept that evidence consists of facts, I also take this to be an argument that inferential knowledge need not be based on evidence.

§3 presented four strategies for defending the No Falsehoods principles from the counterexamples. These strategies all posit proxies: non-obvious motivating reasons/bases of inference that align with justifying reasons/evidence. The last two, which appeal to an agent's background evidence or internalize the motivating reason, are the front-runners. I gave two general arguments against proxy views in §4. The first argument poses a dilemma. Proxy views are either disjunctive, or committed to massive overdetermination. The second argument is that proxy views are self-undermining. They require that agents use the perfectly right reasons, but must allow that they use those reasons in imperfect ways.

I conclude that the No Falsehoods principles are false. Since I accept the orthodox account of reasons, I reject Good Reasons. Since I accept that evidence consists of facts, I reject the requirement that inferential knowledge be based on evidence. Dor's

donation is praiseworthy, even though her reason for donating wasn't a good reason to donate. Caliban knows there are less than 60 sheep in the field, even though he didn't use good evidence to reach that belief.

Alternatively, we could deny that evidence or reasons must consist of facts. The reason for Dor to donate is that she would save 52 lives by doing so, even though she wouldn't save 52 lives. Caliban's evidence for believing that there are less than 60 sheep is that there are 52 sheep, even though there aren't 52 sheep. This would allow us to save Good Reasons and the claim that knowledge is based on evidence.

There is a predominant view among epistemologists who allow that falsehoods can be part of an agent's evidence. Though not any falsehood that the agent believes can serve as evidence, they say that falsehoods which are "close enough" to the truth can be. Authors disagree on what it takes to be "close enough", but we can see how the general move is supposed to work.[22] So, a person who arbitrarily believes a falsehood cannot use it as a premise in inference to gain knowledge.

But consider **Field 2**. The basis of Caliban's inference is (p) "There are 52 sheep in the field". Though this is false, it is not far off from the truth (p^*) "There are 53 sheep in the field." Moreover, Caliban has good reason to believe p, and both p and p^* support an inference to q. On this view, then, p is part of Caliban's evidence, even though it is false. On the other hand, Brittney's belief that "There is a sheep in the field" is too far from the truth to serve as the basis of a knowledge-producing inference. We can use similar reasoning in **Donation 2**. Dor's motivating reason is that her donation would save 52 lives. Though this is false, there is a truth in the vicinity: her donation would save 53 lives. Since her motivating reason is close enough to a fact, and either way would support her action, "the donation would save 52 lives" is a reason justifying her action.

I recognize that, in order to reject Good Reasons, I would have to argue against this unorthodox view on which reasons can be falsehoods. I do not have the space

[22]For instance Warfield (2005) or Klein (2008)

for that here. Instead, I want to explain what an alternative view can look like. It is inspired by virtue epistemology.[23]

The thought is that praiseworthy action and knowledge are matters of exhibiting the appropriate virtue. Praiseworthy action is right action that manifests moral virtue. Knowledge is correct belief that manifests epistemic virtue. Virtuous actors are pretty good at tracking reasons. For the most part, they know a reason when they see one. This sensitivity to moral reasons seems to be foundational to moral virtue. Similarly, virtuous believers are pretty good at tracking evidence. An important epistemic virtue is basing your belief on good evidence, and reasoning from that evidence correctly.

However, even the most virtuous agents aren't perfect. They can still make mistakes. Virtues are dispositions of thought and character, and most dispositions aren't sure-fire. Dor manifests the virtues of charity and kindness in her action. Her motivations were pure, and even if she didn't hit the nail on the head, they reflect well on her character. Even though she wasn't moved by a good reason, her moral virtue led her to be motivated by something that was pretty close to a good reason.

Similarly for Caliban. His reasoning was, by and large, good. He exhibited an appropriate—epistemically virtuous—sensitivity to the evidence. His epistemic virtue led him to the truth of the matter on whether there were less than 60 sheep. That is how his correct belief manifested excellence, and why it counts as knowledge.

I believe this view provides a better explanation for why we care about the role of facts in epistemology and ethics than the view that allows for falsehoods to be reasons. That is to say, competitor views require that justifying reasons coincide with motivating reasons, but allows that justifying reasons can be false. Thus, they have to explain which falsehoods can count as reasons. What makes this falsehood "close enough" to a truth that it can be reason, where this other falsehood is "too far away"? If falsehoods can be reasons, why do they need to be "close" to the truth

[23]See especially Sosa (2007) and Mantel (2013)

in the first place?

My view avoids these tricky questions. Justifying reasons are facts, but motivating reasons can be falsehoods. I must explain when a motivating reason is "close enough" to a justifying reason. I say that, as long as you are virtuous, even if you are not perfect, you can be praiseworthy or gain knowledge. Since we already need to distinguish between virtue and perfection, I have apparently collapsed a new problem into one we already needed an answer to.

Clearly, much remains to be said on either side. We should reject the No Falsehoods principles. We have the choice of whether to keep Good Reasons and accept that reasons can be falsehoods, or we can accept that reasons are fact and ditch Good Reasons.

Chapter 5
Moral Swamping

It matters what we do. But it also matters how we do it. Whether you succeed or fail, act rightly or wrongly, it's important whether you made that decision for yourself, freely, on your own terms.

In this paper, I'm going to argue that the above truism is not as straightforward as it seems. There is a real problem here in explaining why it is true: given the properties that make an action free, why should it matter whether you act freely? I'll press a version of the swamping problem against some standard moral claims. The swamping problem[1] is a challenge to explain the value of epistemic justification in light of the descriptive features of a belief that make it justified.

My goal in this paper is to show that there is a swamping problem for voluntary, free, or autonomous action.[2] I will not try to solve this moral swamping problem, nor will I argue that it is unsolvable. I will identify three families of reactions to the swamping problem, and discuss some of their apparent merits and drawbacks.

Before we start, here is a quick presentation of the problem to come. The capacities that make for autonomy are at least instrumentally valuable: they help people make good decisions. When these abilities are exercised, the resulting action is free. For this reason, a freely-made decision is more likely to be a good one. Whether it ultimately turns out to be good or bad, why should we care that it was antecedently more likely to be good? If we shouldn't, then it is hard to maintain the view that free action is

[1] Zagzebski (1996) introduced the problem. Pritchard (2010) is a good, contemporary survey.

[2] I'll use "voluntary", "free", and "autonomous" as synonymous. In §1, I'll explain my usage.

specially valuable.

Here is the plan. In §1, I'll elaborate on how autonomy seems to matter. In the terminology of this paper, freedom apparently has final value. Having established the appearance of final value, §2 turns to some of freedom's interesting descriptive features. I argue that, on the two most popular theories, an action or decision's being free is partly a matter of how the agent came to it.

With the relevant features made salient, §3 presents the moral swamping problem itself. In order to be free, an action must have a certain pedigree. However, this pedigree does not typically make for final value. In light of that, where do free actions get their extra final value?

§4 considers three families of responses to the swamping problem that both leave something to be desired. The first is deflationary. It may mean giving up on the claim that freedom is finally valuable. The second is primitivist. There is no explanation for why we should treat free actions differently from unfree actions. It is a brute fact that if you chose freely, your action is morally different from an unfree one.

The third family tries to derive an explanation of the value of freedom from the value of something else. The most prominent candidate view is that free actions have the special value that they do because of their relationship to the agent (or maybe the agent's will). This would provide a satisfying solution to the swamping problem. But spelling out the details of this account is harder than it first appears. §5 concludes with a gesture at two other swamping problems, and a very quick discussion of what this means for the familiar epistemic swamping problem.

5.1 How Freedom Matters

In this section, we'll get acquainted with autonomy, and how it is supposed to matter in ethics. I will establish a presumptive case for the claim that freedom is finally valuable.

5.1.1 Ground Clearing and Terminology

Ethicists distinguish between the finally and non-finally valuable.[3] I will slightly repurpose this terminology. Traditionally, if something is finally valuable, it is valuable for its own sake. Pleasure, as an intrinsic good, is finally valuable. Pain is finally disvaluable; worth avoiding for its own sake. When something is non-finally valuable, it is valuable for the sake of something else. Lollipops, for instance, are instrumentally valuable. They are valuable only for the sake of the pleasure that they produce. Torture devices are non-finally disvaluable; bad only as a means to causing suffering.

Something can be finally valuable without being intrinsically valuable. Suppose you think that we should preserve the original copy of the Declaration of Independence. This isn't for the sake of some further good, like directing us to Washington's hidden treasure. A molecule-for-molecule duplicate is just as good a treasure map as the original is. We care about the original document for its own sake. However, the original is not more intrinsically valuable than a duplicate. They have the same intrinsic properties. What makes the original more valuable (for its own sake) are its extrinsic properties: that *this* piece of parchment was the original, that it was signed by Jefferson and pals, etc. If this is right, then an object can have final, non-intrinsic value.

Additionally, something can be non-finally valuable without being instrumentally valuable. Suppose you take a test for bone-itis, and it comes back negative. It indicates you do not have bone-itis. That's a good thing! Your test results have *some* kind of value. However, negative test results are not good for their own sake. They are good only as a sign of a further good, health. So they are non-finally valuable. Although they are non-finally valuable, they are not instrumentally valuable. Negative test results do not cause your health to improve. Instead (we can say), they have indicative value. Rather than being instruments of value, they are indicators of it.

---

[3]Korsgaard (1983). Kagan (1998) recognizes the distinction, but decides to repurpose the word "intrinsic" to name it.

So the intrinsic/instrumental distinction gives us a good first pass at the final/non-final distinction, but they don't line up perfectly. Our interest in final value particularly lies in the impact it has on certain "minimally different" pairs of cases. Some examples will make this clearer.

> World x is just like world y in almost all respects. There is one difference. In world x, a lizard gets to bask in the sun for 5 more minutes of pleasure than it would in world y.
>
> World v is just like world w in almost all respects. There is one difference. World v has an extra lollipop floating in deep space. But nobody ever gets to enjoy it, and the same amount of pleasure is felt across both worlds.

When I consider these cases, I judge that world x is better than world y (even if it's only by a little bit). And I judge that world v is exactly as good as world w. This is evidence that pleasure has some final value, but that lollipops don't. The thought is that, if something is finally valuable, then its alone presence will increase the total value of the world. If it is not finally valuable, then merely toggling its presence should not have an impact on the world's total value. Pleasure's presence or absence makes a difference to total value, so pleasures are finally valuable. Adding an extra lollipop to the world, and *nothing* else, doesn't make a difference to total value. Lollipops are not finally valuable.

Throughout this paper, when I say that "freedom is finally valuable" or the like, I officially mean that an action's being free sometimes makes a difference between minimally different objects of evaluation. It seems that whether an action is free can, by itself, make a difference to how we evaluate other things (like whether it's permissible to paternalistically interfere with an agent). I will remain neutral on whether freedom matters for its own sake or for the sake of something else; whether it is"finally valuable" in the traditional sense of the phrase.

For example, suppose that Abyzou and Baal perform similar actions. They both

ϕ, facing the same stakes and with the same results. The only difference between their ϕing is their praiseworthiness. Abyzou ϕed out of concern for the people whose welfare was at stake, so their action was praiseworthy. Baal ϕed in hopes of a reward, so their ϕing is not praiseworthy.

If you think that Abyzou's action is better than Baal's, then praiseworthiness can, by itself, make a difference to the value of an action.[4] If it's more fitting to have pro-attitudes like gratitude towards Abyzou than to Baal, then a bare difference in praiseworthiness makes a difference to which attitudes are appropriate. If giving a reward to Abyzou is better, preferable, or more right than giving it to Baal, praiseworthiness makes a difference to which rewards are right.

On any of these views, then, a bare difference in whether an action is praiseworthy has an impact on the moral landscape. Praiseworthiness can, by itself, make a difference in moral matters. For this reason, I say praiseworthiness is finally valuable. Similarly, I will argue in this section that freedom appears to be finally valuable. As we will see, one way to respond to the swamping problem I will later pose is to abandon the claim that freedom is finally valuable. The cases I present here can only establish a presumption in favor of its final value. I'll return to this possibility at the end of this section.

I have departed from standard terminology, using "finally valuable" to mean "sometimes makes a difference to minimally different pairs" rather than "valuable for its own sake". I do this because I want my arguments to speak to the concerns of non-consequentialists. Non-consequentialists should be able to recognize that some factors make a difference to minimally different pairs without appealing to axiology. For example, compare two cases where Czernobog and Dagon each ϕ, and all the consequences of ϕing are the same. Czernobog promised not to ϕ, but Dagon didn't make any promise.

---

[4] I use the causal/explanatory language of "difference making" for ease of expression. Officially, I only mean to be pointing to the fact that the value of an action (the propriety of praising it, the rightness of rewarding it, etc.) covaries with whether or not it is praiseworthy.

If Czernobog's ϕing is wrong but Dagon's is permissible, this indicates that breaking a promise can by itself make an action wrong. Promise-breaking sometimes makes a difference to minimally different pairs. But this need not be because breaking a promise has any special disvalue; a non-consequentialist could say that keeping and breaking promises is not a matter of value. The point is well taken, and that is why I use "final value" as a technical shorthand for "makes a moral difference between minimally different pairs". Non-consequentialists should be able to accept my terminology and arguments without giving up on their anti-axiological scruples.[5]

This brings us to another non-standard usage of terminology. The words "free", "voluntary", and "autonomous" have been used in a myriad of ways, sometimes synonymously and sometimes not. My stance on their relationship is ecumenical. Ordinary language doesn't distinguish clearly between these words (or concepts, or properties, or whatever). Freedom, autonomy, and volition have been asked to do a lot of work in ethics, politics, and philosophy more broadly. At the outset of inquiry, we should not assume that any pair of them is coextensive. At the same time, we should be open to the idea that some disagreement about whether an action is free or autonomous is merely verbal.

For example, I could claim that it is wrong to paternalistically interefere with autonomous (not necessarily voluntary!) actions, while you claim that voluntary (not necessarily autonomous!) actions are protected from such interference. This could turn out to be a real disagreement, or merely a difference in terminology. That will depend on the details of our views.

I care about the final value these properties are supposed to have and, ultimately, how they are supposed to get this added value. I care much less about the details of an account. For this reason, I use the words "free", "voluntary", and "autonomous"

---

[5]Additionally, if there are organic unities, the relationship between what is valuable for its own sake and what makes a difference in pairs of otherwise-similar cases becomes murky. This is a generalization of the worry that performance on "isolation tests" is not a good measure of intrinsic value (Kagan 1988)

interchangeably. For example, if you are an anti-paternalist, then there's a bit of moral work that you need a property to do. Some property of actions makes the difference between "is ok to paternalistically interfere with" and "is not ok to paternalistically interfere with".

I claim that there is *some* sensible way of using the word "autonomous" to name that property. Similarly, there is *some* sensible way of using "free" or "voluntary" to draw the line between which actions are vulnerable to paternalistic intervention and which aren't.

Moreover, we should not assume that all of freedom's jobs are done by a single property. For example, a life filled with free actions is supposed to be better, in some respect, than a life of coerced or compelled actions. We shouldn't assume that the kind of freedom relevant to anti-paternalism is the same as that relevant to quality of life. It is possible that they are the same, but showing this would require argument. At the end of the day, we'll probably need to say something like "Free_1 actions make a life better in some respect. It is wrong to paternalistically interfere with free_2 actions. A freely_3-made promise is binding" and so on. Again, I refer to all of these indiscriminately as "free", "autonomous", or "voluntary".

5.1.2 Some Jobs for Freedom

That's enough by way of terminology. The rest of this section discusses the kinds of minimally different pairs where autonomy seems to make a difference. That is to say, I'll argue that freedom is apparently finally valuable. We'll see some jobs freedom is supposed to do. In the next section, we'll start to worry about how it's supposed to get those jobs done. I have hinted at a couple of places where freedom's final value shines through. Here is a case to make those clearer.

> Geryon and Haures are each having a hard time with their respective family's farms. The crops are failing, and each has considered selling the

> farm and moving to the city for uncertain prospects. Upset, each goes to their local bar. At this point, their stories diverge.
>
> As Geryon approaches the bar, a representative of Big Agra comes with an offer. Knowing Geryon's fallen on hard times, Big Agra wants to buy the land. This would be just enough for Geryon to establish themselves in a nearby city. Geryon carefully considers the risks and rewards. This includes the loss of an agrarian way of life and the family farm. Geryon cares a lot about these things, but ultimately decides to sell.
>
> After the representative finishes up with Geryon, they give Haures the exact same offer with the exact same benefits and risks. By this point in the night, Haures has gotten blind drunk and quite maudlin. Though just as much is at stake for Haures as for Geryon, Haures isn't in a cognitive or emotional position to weigh options carefully against each other. Haures accepts the offer and promptly loses consciousness.

I take it that Geryon has made their own decision about whether to sell the farm. They considered the issue carefully, and their decision properly reflects their values. On the other hand, Haures's decision is suspect. That Haures was so drunk and distraught is reason to believe that something's gone wrong with their decision-making process.

Let's consider some of the normative differences between Geryon's and Haures's decisions. Suppose that their decisions turn out well. Each goes on to live a fulfilling urbanized life. Geryon is in a position to feel proud of the decision they've made. They can look back on that night in the bar as a major turning point, and see themselves as author of their own life. Haures, on the other hand, cannot take credit for improving their life. If anything, Haures should be glad for their good fortune! Despite their incapacity, they luckily made a great decision.

These judgments reflect the final value of autonomy in several ways.

1. If a decision to act is autonomous, it has extra value.

   As things turned out, Geryon made a good decision. I feel some resistance to saying that Haures even made a decision in the first place. Insofar as Haures decided at all, they didn't do a good job at deciding.

   Alternatively, suppose that their decisions are going to turn out poorly. Agrarian tides will turn, and both Geryon and Haures would be better off keeping their farms. Geryon's decision still seems to be better, in some way. Geryon chose poorly (as things turned out), but at least Geryon chose for themself. Though Geryon's decision was bad, they made it on their own terms. Haures's decision lacks that value.

2. Freedom licences different self-regarding attitudes.

   By this I mean that that how an agent should feel about themselves and their actions depends on whether they've acted freely.[6] Geryon is in a position to feel proud. Haures should not.

3. Similarly, third parties should respond to an action differently based on whether it was autonomous.

   Geryon's decision is admirable. They made a difficult, life-changing decision, on their own terms, in their own way. As third parties, that's something we can respect and admire. On the other hand, Haures's decision is not a proper object of veneration. Haures got lucky, but we shouldn't admire or want to emulate them.

4. Freely-made commitments are binding.

   Suppose that, the next day, Geryon and Haures regret their decisions. They wish to unsign the contracts. Haures has a case that they should not be bound

[6] Alternatively, the reasons agents have to adopt those attitudes, or the propriety of those attitudes, or the attitudes' value depends on whether the agent has acted freely.

by the terms of the contract. A commitment made while blind drunk does not bind the person who committed. If it became a serious dispute, I would feel uncomfortable enfocring the contract, as written, against Haures.

On the other hand, if Geryon regrets their decision, they can't just go back on it. We might feel sympathy for Geryon; it was a tough decision to make. Maybe they should have a chance to re-negotiate the terms of the contract. But if push comes to shove, I feel much more at-ease enforcing Geryon's contract.

5. Freely-made decisions are protected from paternalistic interference.

   Suppose that, having done some philosophy, I decide that the agrarian life is better for the farmers than the urban one. Though agrarian pleasures are less intense than big city experiences, they are pleasures of a higher, more valuable kind. For the sake of their own wellbeing, I want Geryon and Haures to stay on their farms.

   As Geryon is about to sign on the dotted line, I swoop in and steal the contract right out from under them. I've forcibly delayed Geryon's decision overnight, until I've decided that they've considered the issue carefully enough. Later on, I swoop in and steal the contract meant for Haures, delaying their decision overnight, until they've sobered up.

   I'm not sure whether I've treated either of Geryon or Haures appropriately on balance. However, I am confident that my treatment of Geryon is worse, more wrong, less respectful.

   Geryon's made a decision for themselves. Even if I think it is the wrong decision, it isn't my place to meddle in their affairs. Haures wasn't in a state to decide for themselves. It's more acceptable to keep a drunk person from making an important decision until they're sober, especially if they've decided incorrectly.

This is a small sample of the cases that illustrate autonomy's final value. Autonomous actions are importantly different from non-autonomous ones. We could

sit here all day listing other jobs that philosophers have found for freedom. Before moving on, I want to mention three other prominent ones.

Consent is at the core of medical and sexual ethics. Consent makes the difference between a routine root canal and a gruesome assault. But consent must be freely given in order to make a difference. People who are incapacitated can say yes to surgery, but they can't offer valid consent. In order to make a moral difference, consent must be given voluntarily.

Many of the points I made above about private morality carry over to politics. You and I shouldn't paternalistically interfere with autonomous-but-bad decisions, and neither should the state. The state should not bind people by contracts or promises that were not made voluntarily. Unsurprisingly, freedom has many jobs in politics.[7]

Last, and related to consent and promise-making, we have more general powers to change the rights and obligations of others. To give an example, suppose that Ipos breaks Juiblex's window. This normally creates an obligation in Ipos to compensate Juiblex for the cost of the window. However, Juiblex has the power to forgive Ipos and thereby erase that obligation. This raises two questions. First, is Ipos still obligated to compensate if they broke the window involuntarily? Second, if Juiblex forgves Ipos's debt unfreely (e.g. while intoxicated to the point of incapacitation), does that erase the debt? If the answer to either of these questions is no, then we've found another place for freedom to make a difference.

This concludes my argument for freedom's apparent final value. It is possible that this appearance is illusory. In order to test for freedom's final value, we need to make sure that the cases of comparison differ only in whether the action was performed freely. A person who denies freedom's final value can try to find some other factor that differs between the cases. They could say:

---

[7]See especially Christman (2005, 2008), Shiffrin (2000)

> For complicated empirical reasons, it would decrease aggregate utility to paternalistically interfere with Geryon's action, but not with Haures's. We should treat these cases differently for familiar utilitarian reasons, and not special freedom-based ones. When we really ensure that there is no difference to aggregate utility, we should treat these free and unfree actions in the same way.
>
> The jobs we've discussed above are all filled by other properties. That an action is free is usually a good sign that it has one of these other properties. But freedom can never make a moral difference by itself.[8]

In this way, a person can argue that freedom only appears to be finally valuable. This is a familiar dialectic when we are trying to determine whether something is finally valuable. We'll have occasion to reconsider this tactic in §4.

In this section, I've argued that autonomy apparently has final value. Other moral matters are supposed to turn on whether an action was free. I named eight roles that autonomy plays in the moral world. In the next section, we'll look at the descriptive properties of free actions, which enable them to fill their roles.

5.2 Pedigree

In the previous section, we discussed how freedom make a moral difference. For instance, an autonomous action has added value, and it is wrong to paternalistically interfere with it. In this section, we'll look at the properties that make an action free, as opposed to unfree. That is to say, there must be some descriptive difference between autonomous and non-autonomous actions that gives rise to their difference in final value.

I will argue that, on two prominent views of autonomy, an action's pedigree makes

[8]More carefully, there is some set of other properties that screen off the correlation between freedom and the value of an action, or the propriety of interfering with it, etc.

a difference to whether it is voluntary. Free actions have a special causal history. Normative competence views say that free actions come from decision makers who are able to choose well. Bipartite "authenticity and capacity" views hold that free actions come from decision makers who are able to choose well by their own lights. In the next section, I will argue that this kind of history doesn't typically make for the final value discussed in §1.

In this section, I will write as if the views under discussion seek to give necessary and sufficient conditions on free action. Realistically, we should understand authors as only offering necessary conditions on autonomy. For example, a Raz (1986)-inspired view would also hold that in order to choose freely, agents must have a diverse and valuable enough set of options to decide between. For the most part, I will set aside these complications, as I do not believe they are relevant to my main point. We will return to the other conditions on autonomy in §3.3.

5.2.1 Normative Competence

There are many competing accounts of what makes an action autonomous. I want to start with an often-overlooked family of theories, which say that autonomy is a matter of normative competence. These deserve special attention because they will be the most obviously vulnerable to the problem posed in the next section. I'll leverage the swamping problem for normative competence views to pose a broader problem for a more popular family of views.

Normative competence views of autonomy get their contemporary, Anglophone start in Wolf (1990). She argues that an agent is morally responsible for an action only when the agent has the skills and abilities necessary to make the right decision. As I see it, Wolf's key insight is that moral responsibility requires the ability to understand the difference between right and wrong, make a decision on the basis of their understanding, and then act on it. They must be normatively competent.

More recently, feminist theorists of autonomy have picked up on normative competence as a theory of autonomy. For example, Stoljar (2000) argues that some women who take contraceptive risks do not do so freely. They subscribe to false and harmful ideas about gender and womanhood. This interferes with their ability to understand and be motivated by the genuinely good reasons to use contraception during sex. She argues that this renders their contraceptive decisions non-autonomous.

At a high level, normative competence theories hold that agents act freely only when they have the competence to take into account the reasons bearing on their decision. They must be competent to evaluate their decision, in light of what the real norms (whatever those are) demand. The basic idea behind normative competence views of autonomy is the same as that for moral responsibility. In order for a person to truly make their own, free decision, they've got to be able to understand what's truly at stake in making their decision, and how good their options really are.

We can get stronger and weaker versions of a normative competence theory by tweaking various elements. Is it enough for agents to be competent, or must they exercise that competence? Must the agent competently account for *all* of the objective values at stake, or is it enough for them to consider how a choice will impact a narrower range of values?[9] How competent do they have to be; how much room for error in evaluation is there?

There's lots of good work to be done in figuring out the details of a normative competence view. Notice that they will all share a feature. On all normative competence views, free actions will come from a source that is at least able to choose in line with the reasons that there are. It is easy to see why possessing normative competence is instrumentally valuable. All else equal, a normatively competent person is more likely to do the right thing than a normatively incompetent person.

So on a normative competence view, free actions are the products of instrumentally

[9]For example, a view could say that agents need only be able to evaluate how much is objectively at stake for the agent and their loved ones. It should be compatible with choosing freely that you don't give appropriate moral consideration to the interests of strangers.

valuable sources. This means that free actions will typically have indicative value, like good test results. A free action is (whatever else it is) the action of a normatively competent person. A normatively competent person is more likely to do good things than a normatively incompetent person. So, when we see that Geryon's action was free, this is evidence that their other actions are going to be good. On the other hand, when we see that Haures's (or Stoljar's agents', who take contraceptive risks) action was not free, that's evidence that they've made the wrong decision, and may do so on other occasions.

Whether or not free actions are finally valuable, then, they have indicative value. Coming from a place of normative competence, free actions are the products of instrumentally valuable sources. Therefore, ϕing freely is evidence that you've ϕed rightly, and that your actions on other occasions will be right.

5.2.2 Authenticity and Capacity

Normative competence views of freedom are by no means orthodoxy. Most accounts of freedom tie it more to an agent's judgment and desires rather than the normative facts. To put the concern colorfully, freedom is a matter of choosing by *your* lights, not the objectively true lights. Normative competence views supposedly confuse autonomy—self-governance—with orthonomy—correct governance.

As Christman (2008) has pointed out, most views of autonomy are bipartite. On these views, autonomy is a matter of exercising the capacity to pursue one's authentic goals, desires, or preferences. The meat of a view is in giving accounts of authenticity and of capacity. Authenticity is a feature of preferences, values, desires, or something similar; the psychological factors that play an appropriate role in decision-making and motivation. An authentic desire is one that is *really*, *truly* the agent's. When agents are brainwashed or in the grips of compulsion, their authentic desires are replaced with or masked by some inauthentic ones.

Capacity is a feature of agents. Agents should have the information and decision-making ability necessary to pursue their authentic desires. Only when agents exercise decision-making capacity in pursuit of an authentic desire do they act autonomously.

For example, on Christman (1991)'s view authentic desires are ones that the agent is not alienated from. Were the agent to critically reflect on these desires and their origins, the agent would not have a certain kind of negative reaction. On a Frankfurt (1971)-style view, an authentic desire to ϕ is not accompanied by the second-order desire to not want to ϕ. Decision-making capacity includes things like whether the agent understands the choice before them, and whether they can weigh their options, in light of their desires, in a sufficiently rational way (Grisso & Appelbaum 1998).

We can see bipartite and normative competence as related in the following way. Both require that autonomous agents are capable of living up to certain norms. Normative competence say that the relevant norms are (appropriately related to) the objectively correct norms dictating right behavior. Bipartite views hold that the relevant norms are the ones that this individual agent would authentically endorse and prefer to abide by.

It seems like bipartite views are committed to the claim that it matters whether agent's authentic desires are satisfied. Otherwise, it would be a mystery why autonomy, the capacity to pursue one's authentic desires, would be important.[10] Granting that an agent's authentic desires matter,[11] we see that these views have another thing in common with normative competence views. According to bipartite views, free actions are the products of instrumentally valuable sources.

An agent acts autonomously when they exercise their autonomy. This consists in

---

[10]This is also borne out by our reasoning about cases: On reflection, a life where someone's authentic desires are satisfied is importantly different from a life where they are brainwashed by a cult, and their new brainwashed desires are satisfied. See for instance Nussbaum (2000, ch. 2).

[11]I am not sure we should grant this. Valdman (2010) also worries that authentic desires are not importantly different from inauthentic ones, but he notes it is a minority position.

exercising a capacity of the agent's to (confidently, rationally, with sufficient information, etc.) pursue their authentic desires. The capacity is valuable as an instrument to getting one's authentic preferences satisfied. All else equal, a person with decision-making capacity is in a better position to live the life that they truly want than a person who is incapacitated.

In this way, as with normative competence views, free actions have indicative value. When an action is freely performed, that is evidence that it will satisfy an agent's authentic preferences. Geryon, who competently weighed their options in light of their values, is more likely to have chosen well for themselves than Haures is. Free actions get this indicative value by virtue of their pedigree, as the products of an agent's decision-making capacity.

In this section, we saw how free actions have indicative value. Whether you accept a normative competence or a bipartite view, an action's being free is evidence that it is valuable in other respects. The swamping problem in the next section is a challenge to explain the final value of autonomy. I will argue that neither normative competence nor authenticity and capacity provide an obvious explantion of freedom's final value.

5.3 Moral Swamping

In this section, I will press a version of the swamping problem against free action.

First, I will explain the[12] swamping problem. Next, I will argue that normative competence views of autonomy are vulnerable to swamping. I start with normative competence views because I think the problem is more apparent for them. That done, I will leverage my argument into a swamping problem for the bipartite family of views.

To summarize my argument in this section: In §1, I argued that free actions are

[12]At this point, many authors have written on swamping. I don't want to claim that there is only one problem worth dealing with in this neighborhood. Perhaps it would be better to say "a swamping problem".

apparently finally valuable. It matters whether an action is free or unfree. In §2, I argued that these apparently finally valuable actions all have a special history. Free actions are free—hence finally valuable—because of their pedigree. They are the products of instrumentally valuable sources. However, I will soon argue that things with this kind of pedigree do not usually have final value. Free action's distinguishing feature is also a sign that it isn't finally valuable. The swamping problem challenges us to resolve this tension.

Zagzebski (1996) first posed the swamping problem against reliabilist theories of knowledge. Knowledge is supposed to better in some way than mere true belief. Where does this difference in value come from? Compare an agent who knows that p to an agent who merely believes that p. On a reliabilist theory, the fundamental difference is that when an agent knows that p, their belief that p was produced by a process that reliably produces true beliefs.

A process that reliably produces true beliefs is clearly instrumentally valuable. It helps to produce true beliefs, which are valuable. But, Zagzebski says, this is not enough to secure the value that is characteristic to knowledge. To establish this point, she draws an analogy with coffee. Compare two cups of coffee which are, molecule for molecule, identical. They both have a beautiful color and flavor, not too acidic or bitter. They are intrinsically indistinguishable. However, one cup was produced by a reliable coffee machine that was operating normally. The other cup was produced by an unreliable machine that usually produces horrible cups of mud; it got lucky this time.

Zagzebski claims, and others have agreed, that one cup is just as good a cup of coffee as the other. You have no reason to prefer the one cup over the other. Whatever value is supposed to derive from the fact that one cup was produced by a reliable machine is swamped by the value it gets from its flavor, aroma, color, etc. By analogy, then, the fact that one belief was produced reliably (and so amounts to knowledge) does not give you a reason to prefer it over the unreliabe belief. Whatever value is

supposed to derive from the belief's causal history is swamped by the fact that it's true. Zagzebski concludes that reliabilism is unable to explain the special value of knowledge, since it makes knowledge out to be too similar to a reliably produced cup of coffee.

Given our discussion in the previous section, this conclusion looks hasty. After all, being the product of an instrumentally valuable source, knowledge should have indicative value. If you know that p, then your belief was formed in a way that reliably produces true beliefs. Other of your beliefs which were produced in the same way are more likely to be true. The person who believes p without knowing that p is less likely to believe the truth on other matters, since they apparently reason in unreliable ways.

We should not see Zagzebski as denying that knowledge has indicative value. Rather, she doubts that knowledge is of merely indicative value. For example, suppose Kroni and Lempo each truly believe that p. Kroni's belief was produced reliably and amounts to knowledge, but Lempo made a lucky guess. If we know their entire track records for true and false beliefs, their belief that p ceases to have indicative value.[13] And yet Kroni's belief still seems to be better, preferable, more admirable than Lepo's. Justified belief seems to have some added *final* value that reliabilists cannot account for.

Since Zagzebski, others have pressed this argument against veritism more broadly.[14] Veritism, also called truth monism, is the view that true belief is the sole fundamental value in epistemology. According to veritists, all other epistemic values must be explained in terms of the value of true belief. The argument goes like this: if all that ultimately matters in epistemology is true belief, then justification can matter only as a means of getting to the truth. The sort of belief-forming methods that produce

[13]Compare: If the Oracle tells you that you will never have boneitis, there is no value in taking further tests for it.

[14]Pritchard (2010)

justifed belief are thus instrumentally valuable. Therefore (according to this argument), veritists can say that justified beliefs are at best the products of instrumentally valuable sources.

That doesn't make for extra value in cups of coffee, and it shouldn't make for extra final value in beliefs. To state the swamping problem more succinctly,

> Justified beliefs and well-made cups of coffee are analogous in that both are the products of instrumentally valuable sources. This pedigree doesn't make for additional final value in coffee. Where does the additional final value of justification come from?

At this point, we can see the swamping problem for free actions start to take shape. Like cups of coffee, these actions are the products of instrumentally valuable sources. Where do they get their extra final value from? If you are convinced up to this point, feel free to skip to §4. I think there is value in laying the swamping problems out more explicitly, to make the problems more gripping. That's our next task, starting with normative competence.

5.3.1 Swamping Normative Competence

Recall that, on normative competence views, a free decision is one that is made by a normatively competent agent. Such agents are able to understand and weigh the reasons bearing on their decision. They are competent to make the decision that they ought to.

However, there is a gap between competence and performance. A normatively competent decision can still be, ultimately, wrong. Recall Geryon, who freely decided to sell the family farm and move to the city. This may not be the objectively right course of action. Suppose that the way of life that Geryon could have at the family farm is, objectively speaking, better for them than the way of life they'd have in the big city.

In this case, Geryon made the wrong decision. All things considered, they shouldn't have sold the farm and it would be better, for their own sake, to stay. But they still decided competently, with the best information available and to the best of their abilities. That is why we say that the decision was made freely. Making mistakes is an important part of acting autonomously.

However, if freedom is just the ability to decide correctly, why should we care about the freedom to make mistakes? An autonomous but incorrect decision is analogous to a bad cup of coffee made by a reliable machine. We can generally count on the normatively competent to do what's right, and we can generally count on reliable machines to make good coffee. When a reliable machine malfunctions, and burns a cup of coffee, we don't insist that the coffee is valuable anyway. It's just as bad as any other burnt coffee. Yet we insist on respecting poor decisions that people make, as long as they have done so freely.

On the other hand, suppose that it is objectively better for Geryon and Haures that each moves out to the city. Geryon made their decision competently, and Haures made theirs incompetently. But they made the same decision, so they both chose the objectively best option. Given that the decisions live up to the objective norms perfectly, why should it matter that Geryon's was made competently? Antecedently, agents like Geryon are more likely to choose correctly than agents like Haures. Once the decisions are made, though, they are both correct; once the coffee is brewed, both cups are equally tasty.

On normative competence views, free actions are the products of normative competence. This is a good explanation of their indicative value, but it seems lacking as an explanation of their final value. For all we've said so far, normatively competent decisions are like cups of coffee made by a reliable machine. If normative competence views are to explain the value of freedom, it looks like they must say that normative competence is not *just* an instrument to good decision-making.

5.3.2 Swamping Bipartite Autonomy

Bipartite views of autonomy are also vulnerable to swamping. On these views, autonomous decision makers exercise capacities to pursue their authentic desires.

So, when agents choose freely, they are more likely to decide in a way that will satisfy their authentic desires. They are in a better position to secure the kind of life that they truly want. However, just because they're well-positioned to succeed by their own lights, that doesn't mean that they will. On the other side, just because a person does not make an autonomous decision, that doesn't mean that they won't end up satisfying their authentic desires.

This time, let's focus on Haures. Haures decided to sell the family farm while drunk to the point of incapacitation. Their intoxication means that they weren't in a position to make a well-reasoned, voluntary choice. But it is consistent with this that deciding to sell the farm and move to the city does best satisfy Haures's authentic desires. If we knew what Haures truly wanted from life, we'd recommend that they sell the farm.

If freedom is the ability to decide in a way likely to satisfy your authentic preferences, why should we care that Haures decided unfreely? Haures is like the poorly-functioning coffee maker in that we can't rely on them to produce good coffee or good decisions. But lucky things can happen. The coffee maker can produce a tasty cup of coffee, and Haures can fortunately hit on the decision that's best by their own lights. For all we've said so far, we should treat Haures's fortunately good decision like the fortunately good cup of coffee. Though it doesn't have a good pedigree, it is just as good as its kin.

Agents who decide like Geryon did are antecedently more likely to get the kind of life they truly want. In this case, though, both Geryon and Haures equally satisfy their authentic preferences. So it is not apparent why Haures's decision is less finally

valuable. On bipartite views of autonomy, free actions are the products of instrumentally valuable sources. That explains why they have indicative value. It does not seem to be a source of final value.

5.3.3 Necessity and Sufficiency

Until this point, I have discussed whether normative competence or authenticity and capacity can explain the final value of autonomy. Realistically, these are necessary conditions on an action's being free. For instance, a free action is not only competent, but chosen from a sufficiently good menu of options. Someone who believes in freedom's final value can try to appeal to the other features of an action that make it free. They could say that, on its own, normative competence is not a source of final value. However, normative competence and a sufficient menu of options together make an action both free and finally valuable.

I do not think such a strategy will work for two reasons. First, even if normative competence and bipartite views only offer necessary conditions, they are pointing to some aspect of freedom that seems to be finally valuable. For example, there seems to be something finally valuable about a normatively competent decision, even if the agent didn't have a sufficient range of options (and so didn't act freely). An explanation for how free actions are finally valuable is not yet an explanation for how normatively competent decisions are finally valuable. And that should be explained as well.[15]

Second, I doubt that the other conditions on autonomy do help to explain its final value. For example, we can argue that Raz's "diversity and value of options" requirement iis only of instrumental value. It's good to have more and better options because the agent is more likely to hit on a better decision. Raz isn't the only game

[15]To draw an analogy with epistemology, suppose that we have an explanation for why knowledge is more valuable than true belief, that crucially relies on the anti-Gettier condition. This wouldn't yet explain why justified true belief is more valuable than unjustified true belief. I'd like an explanation of that fact as well.

in town, of course.

Proponents of so-called weakly substantive accounts of autonomy argue that certain self-directed attitudes are necessary for autonomy. Agents must feel a sense of self-respect or self-trust in order to choose freely.[16] These attitudes may add instrumental or indicative value to a decision, but they don't seem to add final value. Similarly, relational conditions like having sufficient power to make one's will effective[17] are clearly instrumentally valuable. But it is not obvious why they would also be finally valuable.

For these two reasons, I think it is safe to focus on normative competence and authenticity and capacity as our accounts of autonomy. If they have the resources to solve the swamping problem, we don't need to appeal to other conditions. If they don't it's not clear how adding other conditions will help.

In this section, I've argued that being the product of an instrumentally valuable source does not usually make for final value. It does not make cups of coffee more valuable, and epistemologists hve argued that it does not make true beliefs more valuable. Freedom is in the same position. As argued in §2, free actions are the products of instrumentally valuable sources. For this reason, they have indicative value. There remains a puzzle in explaining how they have the final value they apparently do (§1). In the rest of this paper, I will discuss three kinds of responses to the swamping problem. Without deciding between them, I want to point out that each has its advantages and disadvantages.

---

[16]Govier (1993), McLeod (2002)

[17]Oshana (2006)

5.4 Three (Families of) Responses

In the previous section, I posed a swamping problem for autonomy. Free actions are similar to reliably produced cups of coffee in that both are the products of instrumentally valuable sources. This pedigree doesn't typically make for final value. Given that, where does freedom's final value come from? In this section, I will explore three families of responses to the swamping problem. Each has its benefits and drawbacks.

The responses are divided according to how they answer the question "Why is freedom valuable?". At one end, deflationary responses say that free action is valuable for familiar reasons, and that no extra work is required to explain its value. At the other end, primitivist responses deny that freedom's value can be explained. That it is finally valuable is a primitive fact of moral theory. Last, derivational responses try to derive the final value of freedom from the value of something else. Free actions aren't just the products of instrumentally valuable sources; they relate to other valuable things in other ways. Their final value derives from one of these other relations.

I do not claim that these are exhaustive or exclusive categories. For example, I classify a rule-consequentialist view as deflationary, though this may not be fair to rule-consequentialists. They might not fit comfortably in any category. Rather than read this as a once-and-for-all taxonomy, I ask you to read it as field notes on an uncharted swamp. In exploring the conceptual space, this is what I found. If you want to solve the swamping problem, here are some things you should keep in mind or try to avoid.

5.4.1 Deflationary

The first family of responses is deflationary. These views hold that there is not much more to free action than being the product of an instrumentally valuable source. Whatever value autonomy has, we have already hit on the essentials of where that value comes from. I call these views deflationary because they seem to deny that

freedom is finally valuable. Rather than some great and lofty thing, free actions are fundamentally like reliably produced cups of coffee.

A familiar example of a deflationary view is hedonism. The only thing of final value is pleasure, and the only thing of final disvalue is pain. Hedonists can say that other things have non-final value: lollipops are instrumentally valuable, and medical test results have indicative value. Hedonists can agree with my diagnosis in §2. Because free actions are the products of instrumentally valuable sources, they have indicative value.

They can also argue that free actions are instrumentally valuable. Exercising the capacities that constitute autonomy—acting freely—is like exercising a muscle. Not only is free action evidence that your other decisions will be valuable, it helps to promote value in your other decisions by giving you experience.

Hedonists will say that the appearance of final value in §1 is just an appearance. They have a suite of replies to defuse the cases we considered in that section. For example, paternalistic interference with free action is likely to produce frustration, resentment, and other dolorous states. Paternalistic interference with unfree action usually doesn't. That is why we shouldn't (usually) paternalistically intervene on autonomous agents.

These kinds of moves are not exclusive to hedonists. The richer your underlying axiology, the more resources you can appeal to in explaining why freedom merely appears to be finally valuable. You can be a non-consequentialist (because, for example, you believe in the doing/allowing distinction), and still hold that being the product of an instrumentally valuable source accounts for freedom's value. That said, it can be helpful to think of deflationaries as giving consequentialism-adjacent accounts of autonomy.

Because they accept that free actions are fundamentally like reliably produced cups of coffee, most deflationary views will end up denying that free actions are finally valuable. If the only difference between two actions is their pedigree, we

should treat them like a pair of cups of coffee which differ only in pedigree. I can think of one exception to this general claim. Rule-consequentialists seem to accept that the analogy is accurate, but they can uphold the final value of freedom. To elaborate, a rule-consequentialist can say:

> Free action has great indicative value. That an act was performed freely is evidence that it is valuable in other ways. For this reason, it is good as a general policy to treat free and unfree actions differently. Anti-paternalist rules, for example, help to bring about the best consequences. On particular occasions, there may be no benefit to respecting a free action. But adherence to the rule requires that we respect it.

In this way, a rule-consequentialist can claim that free actions really are finally valuable. Even when the only difference between two actions is that one was performed freely, the rule requires that we treat them differently. This example opens up the possibility that there are other deflationary views—which accept the analogy between actions and coffee—but claim that freedom is finally valuable.

The benefits of deflationary views are familiar. If successful, they can explain most cases where we want to treat free actions differently from unfree ones. Moreover, they can do this without adding anything to the moral theory we already agree on. That is to say, we can all agree that free actions have indicative and instrumental value. A successful deflationary view only uses tools that we already had to explain new cases.

The downsides are also familiar. Such views tend to be revisionary: deflationary views explain the apparent final value of freedom by citing other goods that are at stake when freedom is. For instance, from a hedonistic point of view, it would be bad in the long-run to paternalistically intervene with free actions. However, prohibitions on paternalism are supposed to apply even when it requires some sacrifice of the hedonistic good. Deflationary views may be unable to uphold that claim.

Moreover, deflationary views seem to give the wrong explanation for the cases they

do get right. Granting that a hedonist can explain why paternalism is usually wrong, an anti-paternalist could object. We don't reject paternalistic inteference because of what paternalism means for other actions or the aggregate good. Paternalism is objectionable because of what it means for this very agent and their action.

To summarize, deflationary views accept the analogy that motivates the swamping problem. Free actions are similar to reliably-produced cups of coffee. Whatever value autonomy has, it will be explained in those terms. Whether this approach is appealing to you will turn on whether you like the "consequentialist style" of accounting for freedom.

5.4.2 Primitivist

This second response holds that free action is radically different from a reliably-produced cup of coffee. Not only are free actions finally valuable, their final value cannot be further explained.

Explanations must come to an end somewhere. Hedonists accept that pleasure is finally valuable. They can give arguments and explain the evidence for why we should believe that it is valuable. But they probably can't explain what it is about pleasure that makes it valuable. At a certain point, no further explanation is possible. Primitivsts add autonomy to the list of things whose value can't be explained.

I am not sure if this counts as a solution to the swamping problem, or as a refusal to answer it. It has one major factor counting in its favor. Primitivists can always get the cases right. Once we have an account of autonomy that we are happy with, the primitivist can point to that property and say "That's the thing with final value, and here's how much final value it has". Unlike deflationaries, primitivists don't have to pull off fancy moves to explain why we should respect free action in a given case. We simply should.

I suspect many of my readers will be initially attracted to something like a primitivist answer. As we saw in §1, autonomy sure looks like its finally valuable. And

deflationary views are going to have a hard time accounting for all the ways that we are inclined to value freedom. Though nobody should be too quick to add primitives to their theory, freedom (alongside pleasure, and maybe a couple of other things) looks like a pretty good candidate for the fundamentally valuable. If we just need to add one more primitive to our moral theory in order to do all the work in §1, that's a bargain.

I want to caution against buying into primitivism too quickly. I have two reservations about it. For one, it may not be as cheap as it seems. If there's more than one important property that we've called "freedom", we'll need to posit multiple primitives to account for everything we want to. Suppose that there is the kind of autonomy that is relevant to anti-paternalism and the value in a life, and there is a distinct property, also called autonomy, relevant to making a promise or signing a contract. If we take the final value of both of these varieties of autonomy to be primitive, we're adding a lot of primitives to our moral theory.

I don't want to scaremonger about adding primitives to our moral theory. Not everyone puts the same premium on a parsimonious theory with few primitives. If you are willing to add eight or ten or twenty primitive-but-closely-related values to your moral theory, there's not much I can do to convince you not to. My second reservation about taking autonomy as a primitive is that its value seems like it should be explainable.

For example, Wolf (1990, pp. 55–67) argues against "uncaused cause" views of moral responsibility on the grounds that they cannot explain why moral responsibility is a good thing. Why should we value the arbitrariness of an uncaused (or self-caused) action? She argues that only normative competence can explain why moral responsibility matters by tying responsible choices to valuable ones.

Though she was discussing moral responsibility (which may not be freedom, autonomy, volition), and normative competence views face their own problems with explaining freedom's value (as I have argued), this form of argument is worth taking

seriously. It counts against a view of freedom if it is mysterious why we should our actions to be like *that*.

Similarly, O'Neill (2003, §3) discusses views of autonomy that hold that autonomous choices are rational. She notes with approval that rational (as opposed to unconstrained) choices have some kind of connection to morality. Ultimately, though, she rejects these views because they are unable to explain why autonomy has more than instrumental value (p. 6).

These authors reject competitor views on the grounds that they cannot explain certain claims about freedoms' value. This should make us weary of primitivist views, which deny that its value can be explained at all.[18] If our choice is between deflationary and primitivist views, then it looks like we must choose whether to deny freedom's final value, or leave it unexplained. There is something to the idea that *if* autonomy is finally valuable, this fact is explainable. We shouldn't be too eager to go primitivist.

5.4.3 Derivative

This brings us to the last family of responses to the swamping problem. These derivative views attempt to derive the final value of freedom from the value of something else. As we saw, free actions derive their indicative value from their sources. That is to say, because they are the products of instrumentally valuable sources, free actions serve as evidence of other goods. In this way, they get some non-final value. The strategy here is to find some other source that lends some *final* value to free action.

Once we recognize the distinction between final and intrinsic value, we should be open to this possibility. By being appropriately related to momentous events, historic objects derive some final value and become worthy of protection. A similar process could be at work in explaining autonomy's final value.[19]

---

[18]Hurka (1987, p. 364) also cautions against primitivism on the same grounds.

[19]Brogaard (2007) has suggested something similar for the swamping problem in epistemology.

A successful derivative account would provide all the benefits of primtivist and deflationary views, with none of their drawbacks. Like primitivism, derivative views hold that freedom is finally valuable. Even absent other considerations, freedom's final value could tip the scales one way or another. This also means that derivative views have a better chance at avoiding the revisionary pitfalls of deflation. At the same time, derivative views do not require us to posit any new fundamental or unexplained phenomena in our moral theory. Like deflationary views, derivative views are parsimonious.

In addition to this, a derivative view just seems right. Suppose that you like a bipartite view of autonomy; free action is the capable pursuit of an agent's authentic preferences. Presumably, you think that it is good for an agent to live by their authentic preferences, to have the kind of life they truly want to live. And it is something about an agent's authentic preferences that makes free action valuable. We care about free action because it is the capable pursuit of *authentic*, rather than inauthentic, preferences. So it makes sense to try to derive the value of freedom from the value of authenticity.

There may be downsides to a derivative view, but I am not aware of any. If successful, such a view would provide a satisfying explanation for why autonomy matters. This would be a perfect way to solve the swamping problem. Unfortunately, the problem is that so far we have not had any successful derivative views.

We know at the outset that not just any derivation will work. For instance, a theorist could claim that freedom derives its final value from the instrumental value of normative competence. They claim that "being the product of an instrumentally valuable source" is a way to derive final value. But we know, from the swamping problem in §3, that it isn't. So this attempt fails. Other attempts, I worry, face similar arguments. The problem cases may not involve cups of coffee, but the relation

---

Sylvan (2018) does a good job of laying out the derivative program, and attempts to derive the final value of justification and knowledge.

that these accounts point to won't confer value from one relata to another.

I do not have the space to review every attempt to explain the value of freedom and argue that it fails. At any rate, my goal here is to pose a problem and argue that it's worth taking seriously. I do not want to try to argue that no solution could ever work. Instead, I will briefly point out a feature that is common to many attempts, and show why they have not yet solved the swamping problem.

Authors of a Kantian disposition care not just about the actions themselves, but the underlying capacities that produce them. That is to say, part of what distinguishes people as specially valuable ends-in-themselves is that they have certain abilities. They can value things and set maxims for themselves. They can weigh options rationally. They can tell the difference between right and wrong. And they use these abilities to pursue one option rather than another.

It would be natural to suggest that free actions get their final value because of how they are related to the special capacities that produce them. We should respect free actions *because* we respect people's ability to value things and choose rationally. Cholbi (2013) makes this explicit[20]:

> In caring about our rational autonomy, the object of our concern is a capacity... But even when exercised badly, this power is still exercised and is worthy of others' respect... To permit paternalistic interference [with the agent's action] would show little respect for [the capacity] (p. 122–3)

Free action, the exercise of a capacity, is worthy of respect because the capacity itself is. Failing to value the product is akin to failing to value the capacity. Notice that, on these views, the underlying capacities are not just instrumentally valuable. Unlike coffee making machines, the abilities which constitute people as rational and autonomous are specially value regardless of their products. On this view, free actions

[20]See also Korsgaard (1996, p. 120–123) and Hills (2005)

are more than the product of instrumentally valuable sources, they are the products of specially valuable sources. The special value of the source somehow translates into the product, making them finally valuable.

This all sounds promising to me. But the devil is in the details. First, I take it that one of the benefits of deriving final value is that it is more parsimonious than primitivism. Insofar as the special value of the capacities is taken as a primitive, we are therefore compromising on parsimony. To be fair, a hardcore Kantian might try to derive *all* of morality from the special value of rational autonomy. That's a very parsimonious theory! Even more moderate Kantians will probably find more work for the capacity in their moral theory, meaning that they can get more out of their unexplained fact than primitivists.

Still, this is something worth keeping in mind. If the only reason you posit a special value for the capacity to act freely is to explain why free actions matter, you might do better by just taking the value of free actions as a primtive.

My main concern about this view is that it doesn't actually explain why autonomous action is finally valuable. Let's grant that the relevant capacity is finally valuable. This doesn't mean that its product, free action, is also finally valuable. Consider Picasso's painting supplies. Since he was such a great artist, his supplies will have some final artistic or historic value. His supplies are more worthy of our respect and protection than otherwise-identical pots of paint and brushes that didn't belong to anyone important.

Now suppose that I use those supplies to paint a picture. I'm an awful artist. I guarantee you that my painting is no great work of beauty. However, my painting is the product of a finally valuable source (Picasso's supplies). Does my painting have any more final value than a molecule-for-molecule duplicate that wasn't made with Picasso's supplies? If not, then being the product of a finally valuable source does not typically make for final value.

Here is another example:

> Mammon sits down for a big steaming bowl of worms. The worms are very tasty, and give Mammon some amount of pleasure. However, Mammon has a weird digestive system. Gustatory pleasure now always causes gastrointestinal pain later. Mammon suffers some amount of pain.
>
> Naamah eats just as many worms, and derives just as much enjoyment from them. By complete coincidence, Naamah suffers the exact same amount of gastrointestinal pain later on. Naamah's pain is unrelated to the worms they ate earlier.

Compare Mammon's pain to Naamah's. The two experiences are equally painful. For that reason, they've got some measure of final disvalue. However, Mammon's pain is was caused by the pleasure they felt earlier. It is the product of a finally valuable source. Naamah's pain is not the product of a finally valuable source. On net, they feel the same amount of pleasure and pain. Given these facts, is Mammon's pain better, more valuable, more worth wanting than Naamah's?

I think the answer must be no. Whatever final value is supposed to be passed onto the pain is swamped by the painfulness of the experience itself. This case shows that being the product of a finally valuable source does not always make for final value in the product. It's easy to multiply examples beyond the two we have here. This puts pressure on the derivationalist claim that free actions are finally valuable because they are produced by valuable capacities. They need to do more work to explain why free actions aren't like bad paintings made with good materials or pain caused by prior pleasure. That's work that hasn't been done yet.

I end this section with a summary. Deflationary views accept the analogy between free actions and cups of coffee. They are parsimonious, but they don't seem to go far enough in vindicating the final value of autonomy. Primitivist views say that the value of free action cannot be explained. Though these views don't have to be revisionary, they give up on parsimony. As a theorist accepts more varieties of free

action, they end up with a longer list of somehow closely-related-but-entirely-distinct fundamental values.

Derivative views try to derive the final value of freedom by some way other than pointing out that they are the products of instrumentally valuable sources. They promise both extensional adequacy and parsimony. A prominent, Kantian view says that free actions are finally valuable because they are the products of finally valuable sources. However, this general principle is vulnerable to counterexample. If a derivative view is to win the day, they will need a better explanation for how final value goes from source to product.

5.5 Conclusion

In this section, I want to offer some concluding remarks about the original swamping problem from epistemology and a different swamping problem in ethics. First, though, let's recap.

In §1, I argued that free action is apparently finally valuable. There are at least eight types of cases where the fact that an action was free can, on its own, makes a moral difference. In §2, we discussed two families of accounts of autonomous action. On both normative competence and bipartite views, free actions are the products of instrumentally valuable sources. §3 pressed the swamping problem. Although their pedigree can explain why free actions have indicative value, it is unsuited to explain their final value.

Last, §4 considered three types of responses to the swamping problem. Deflationary views accept that not much more can be said in favor of free action. They might have to give up on the claim that autonomy is finally valuable. Primitivist views take autonomy's final value to be an unexplained brute moral fact. This sacrifices parsimony, and is otherwise unsatisfying. Derivative views promise the best of everything: genuine final value, without adding primitives, explained in a satisfying way.

Unfortunately, we don't yet have a derivative view that does what it promises.

It would be premature to conclude in favor of one of these responses rather than another. For my money, I am inclined towards the deflationary views. I don't think the derivative views will ultimately work out, and I have a deep-seated aversion to primitivism about autonomy. Hopefully, others will go to work trying to get a good derivative view off the ground.

To finish up, two more points. First, it seems that this pattern of swamping argument can be used in many places. I think there's a case that praiseworthy and blameworthy action are vulnerable to swamping in the same exact way as free action. This spells trouble for the final (dis)value of praiseworthy and blameworhy action. Other moral properties may be similarly at-risk.

Second, we haven't said that much about the original swamping problem. I do think that my discussion here has ramifications for the original debate.

A minor point: Suppose I am right that there is a swamping problem for praiseworthiness. I believe this would also be troublesome for credit-based responses to the epistemic swamping problem (Riggs 2002). That is to say, we can gloss credit-based views as saying that knowledge is more valuable than true belief for the same reason that praiseworhy action is more valuable than right action. That means a good explanation of why praiseworthy action is valuable would also explain why knowledge is valuable. But if I am right that there is a swamping problem for praisworthiness, more work needs to be done.

More importantly, we can learn something by comparing our responses to the two swamping problems. For instance, epistemologists have been reluctant to accept knowledge's value as primitive. There is considerable pressure towards a deflationary or derivative response. If it turns out that we should be primitivists about the final value of free action, primitivsm about knowledge doesn't look as bad. On the other hand, if we can learn to live with a deflationary view of knowledge and justification, I hope we can learn to live with a deflationary view of autonomy.

Bibliography

Bibliography

Ahlstrom-Vij, K. (2013), *Epistemic Paternalism*, Palgrave Macmillan.

Alvarez, M. (forthcoming), 'Reasons for Action, Acting for Reasons, and Rationality', *Synthese* .

Arnold, A. (2013), 'Some Evidence Is False', *Australasian Journal of Philosophy* **91**(1), 165–172.

Arpaly, N. (2002), 'Moral Worth', *Journal of Philosophy* **99**(5), 223–245.

Arpaly, N. & Schroeder, T. (2014), *In Praise of Desire*, Oxford University Press.

Ball, B. & Blome-Tillmann, M. (2014), 'Counter Closure and Knowledge Despite Falsehood', *Philosophical Quarterly* **64**(257), 552–568.

Barry, B. (1995), *Justice as Impartiality*, Oxford University Press.

Bedau, H. & Kelly, E. (2017), Punishment, *in* E. Zalta, ed., 'Stanford Encyclopedia of Philosophy'.

Brogaard, B. (2007), 'Can Virtue Reliabilism Explain the Value of Knowledge?', *Canadian Journal of Philosophy* **36**(3), 335–354.

Buford, C. & Cloos, C. (forthcoming), 'A Dilemma for the Knowledge despite Falsehood Strategy', *Episteme* .

Cahill, M. (2007), 'Attempt, Reckless Homicide, and the Design of the Criminal Law', *University of Colorado Law Review* **78**(3), 879–956.

Chiao, V. (2010), 'Intention and Attempt', *Criminal Law and Philosophy* **4**(1), 37–55.

Cholbi, M. (2013), Kantian Paternalism and Suicide Intervention, *in* C. C. M. Weber, ed., 'Paternalism: Theory and Practice', Cambridge University Press, pp. 115–133.

Chrisman, M. (2012), 'The Normative Evaluation of Belief and the Aspectual Classification of Belief and Knowledge Attributions', *Journal of Philosophy* **109**, 588–612.

Christman, J. (1991), 'Autonomy and Personal History', *Canadian Journal of Philosophy* **21**(1), 1–24.

Christman, J. (2005), Procedural Autonomy and Liberal Legitimacy, *in* J. S. Taylor, ed., 'Personal Autonomy: New Essays on Personal Autonomy and Its Role in Contemporary Moral Philosophy', Cambridge University Press, pp. 277–298.

Christman, J. (2008), Autonomy in Moral and Political Philosophy, *in* E. Zalta, ed., 'Stanford Encyclopedia of Philosophy'.

Coffman, E. J. (2008), 'Warrant Without Truth?', *Synthese* **162**(2), 173–194.

Dancy, J. (2000), *Practical Reality*, Oxford University Press.

Dotson, K. (2011), 'Tracking Epistemic Violence, Tracking Practices of Silencing', *Hypatia* **26**(2), 236–257.

Dotson, K. (2012), 'A Cautionary Tale: On Limiting Epistemic Oppression', *Frontiers* **33**(1), 24–47.

Duff, A. (2001), *Punishment, Communication, and Community*, Oxford University Press.

Duff, A. (2007), *Answering for Crime: Responsibility and Liability in the Criminal Law*, Hart.

Dworkin, R. (1978), Liberalism, *in* S. Hampshire, ed., 'Public and Private Morality', Cambridge University Press, pp. 113–143.

Feinberg, J. (1965), 'The Expressive Theory of Punishment', *Monist* **49**, 397–423.

Feinberg, J. (1984*a*), *Harm to Others*, Oxford University Press.

Feinberg, J. (1984*b*), *Harmless Wrongdoing*, Oxford University Press.

Fitelson, B. (forthcoming), Closure, Counter-Closure, and Inferential Knowledge, *in* P. Klein, C. de Almeida & R. Borges, eds, 'Knowledge Explained: New Essays on the Gettier Problem', Oxford University Press.

Frankfurt, H. (1971), 'Freedom of the Will and the Concept of a Person', *Journal of Philosophy* **68**(1), 5–20.

Fricker, M. (2007), *Epistemic Injustice*, Oxford University Press.

Gardner, J. (2007), *Offences and Defences: Selected Essays in the Philosophy of Criminal Law*, Oxford University Press.

Goldman, A. (1999), *Knowledge in a Social World*, Oxford University Press.

Govier, T. (1993), 'Self-Trust, Autonomy, and Self-Esteem', *Hypatia* **8**(1), 99–120.

Grisso, T. & Appelbaum, P. (1998), *The Assessment of Decision-Making Capacity*, Oxford University Press.

Hampton, J. (1992), 'Correcting Harms versus Righting Wrongs: The Goals of Retribution', *UCLA Law Review* **39**, 1659–1702.

Hills, A. (2005), 'Rational Nature as the Source of Value', *Kantian Review* **10**(1), 60–81.

Hookway, C. (2010), 'Some Varieties of Epistemic Injustice: Reflections on Fricker', *Episteme* **7**, 151–163.

Hurka, T. (1987), 'Why Value Autonomy?', *Social Theory and Practice* **13**(3), 361–382.

Husak, D. (2000), 'Holistic Retributivism', *California Law Review* **88**(3), 991–1000.

Husak, D. (2013), Penal Paternalism, *in* C. Coons & M. Weber, eds, 'Paternalism: Theory and Practice'.

Jønch-Clausen, K. & Kappel, K. (2016), 'Scientific Facts and Methods in Public Reason', *Res Publica* **22**(2), 117–133.

Kagan, S. (1988), 'The Additive Fallacy', *Ethics* **99**(1), 5–31.

Kagan, S. (1998), 'Rethinking Intrinsic Value', *Journal of Ethics* **2**, 299–320.

Kappel, K. (2017), 'Fact-Dependent Policy Disagreements and Political Legitimacy', *Ethical Theory and Moral Practice* **20**(2), 313–331.

Kappel, K. & Jonch-Clausen, K. (2015), 'Social Epistemic Liberalism and the Problem of Deep Epistemic Disagreements', *Ethical Theory and Moral Practice* **18**(2), 371–384.

Klein, P. D. (2008), Useful False Beliefs, *in* Q. Smith, ed., 'Epistemology: New Essays', Oxford University Press, pp. 25–63.

Korsgaard, C. (1983), 'Two Distinctions in Goodness', *Philosopical Review* **92**, 169–195.

Korsgaard, C. (1996), *Sources of Normativity*, Cambridge University Press.

Kymlicka, W. (1989), *Liberalism, Community, and Culture*, Oxford University Press.

Lackey, J. (1999), 'Testimonial Knowledge and Transmission', *Philosophical Quarterly* **49**, 471–490.

Lackey, J. (2007), 'Norms of Assertion', *Nous* **41**, 594–626.

Larmore, C. (1987), *Patterns of Moral Complexity*, Cambridge University Press.

Larmore, C. (1990), 'Political Liberalism', *Political Theory* **18**(3), 339–360.

Littlejohn, C. (2012), *Justification and the Truth-Connection*, Cambridge University Press.

Littlejohn, C. (2016), Learning from Learning from Our Mistakes, *in* P. Schmechtig & M. Grajner, eds, 'Epistemic Reasons, Norms, and Goals', De Gruyter, pp. 51–70.

Luzzi, F. (2014), 'What Does Knowledge-Yielding Deduction Require Of Its Premises?', *Episteme* **11**(3), 261–275.

Mantel, S. (2013), 'Acting for Reasons, Apt Action, and Knowledge', *Synthese* **190**(17), 3865–3888.

Markovits, J. (2010), 'Acting for the Right Reasons', *Philosophical Review* **119**(2), 201–242.

McLeod, C. (2002), *Self-Trust and Reproductive Autonomy*, MIT Press.

Mill, J. S. (1859), *On Liberty*.

Mills, C. (2007), White Ignorance, *in* S. Sullivan & N. Tuana, eds, 'Race and Epistemologies of Ignorance', State University of New York Press, 11-38.

Montminy, M. (2014), 'Knowledge despite Falsehood', *Canadian Journal of Philosophy* **44**(3-4), 463–475.

Morris, H. (1968), 'Persons and Punishment', *The Monist* **52**, 475–501.

Nussbaum, M. (2000), *Women and Human Development*, Cambridge University Press.

Nussbaum, M. (2011), 'Perfectionist Liberalism and Political Liberalism', *Philosophy and Public Affairs* **39**, 3–45.

O'Neill, O. (2003), 'Autonomy: The Emperor's New Clothes', *Aristotelian Society Supplementary Volume* **77**(1), 1–21.

Oshana, M. (2006), *Personal Autonomy in Society*, Ashgate Publishing.

Pohlhaus, G. (2012), 'Relational Knowing and Epistemic Injustice: Toward a Theory of Willful Hermeneutical Ignorance', *Hypatia* **27**(4), 715–735.

Pritchard, D. (2010), What is the Swamping Problem?, *in* A. Reisner & S.-P. Asbjorn, eds, 'Reasons for Belief', Cambridge University Press.

Quong, J. (2010), *Liberalism without Perfection*, Oxford University Press.

Rawls, J. (1971), *A Theory of Justice*, Harvard University Press.

Rawls, J. (1996), *Political Liberalism*, Columbia University Press.

Raz, J. (1982), 'Liberalism, Autonomy, and the Politics of Neutral Concern', *Midwest Studies in Philosophy* **7**(1), 89–120.

Raz, J. (1986), *The Morality of Freedom*, Oxford University Press.

Riggs, W. (2002), 'Reliability and the Value of Knowledge', *Philosophy and Phenomenological Research* **64**(1), 79–96.

Riggs, W. (2008), The Value Turn in Epistemology, *in* V. Hendricks, ed., 'New Waves in Epistemology', Palgrave Macmillan, pp. 300–23.

Saul, J. (2012), *Lying, Misleading, and What Is Said*, Oxford University Press.

Schnee, I. (2015), 'There Is No Knowledge from Falsehood', *Episteme* **12**(1), 53–74.

Schroeder, M. (2007), *Slaves of the Passions*, Oxford University Press.

Sher, G. (1997), *Beyond Neutrality: Perfectionism and Politics*, Cambridge University Press.

Shiffrin, S. V. (2000), 'Paternalism, Unconscionability Doctrine, and Accommodation', *Philosophy and Public Affairs* **29**(3), 205–250.

Sliwa, P. (2016), 'Moral Worth and Moral Knowledge', *Philosophy and Phenomenological Research* **93**(2), 393–418.

Sosa, E. (2007), *A Virtue Epistemology*, Oxford University Press.

Sosa, E. (2009), *Reflective Knowledge: Apt Belief and Reflective Knowledge, Volume II*, Oxford University Press.

Sosa, E. (2011), *Knowing Full Well*, Princeton University Press.

Stoljar, N. (2000), Autonomy and the Feminist Intuition, *in* C. Mackenzie & N. Stoljar, eds, 'Relational Autonomy: Feminist Perspectives on Autonomy, Agency, and the Social Self', Oxford University Press, pp. 94–111.

Stratton-Lake, P. (2000), *Kant, Duty, and Moral Worth*, Routledge.

Sutton, J. (2005), 'Stick to What You Know', *Noûs* **39**(3), 359–396.

Sylvan, K. (2018), 'Veritism Unswamped', *Mind* **127**(506), 381–435.

Valdman, M. (2010), 'Outsourcing SelfGovernment', *Ethics* **120**(4), 761–790.

von Hirsch, A. (1993), *Censure and Sanction*, Clarendon Press.

Walen, A. (2016), Retributive Justice, *in* E. Zalta, ed., 'Stanford Encyclopedia of Philosophy'.

Wall, S. (1998), *Liberalism, Perfectionism, and Restraint*, Cambridge University Press.

Wall, S. (2010), 'Neutralism for Perfectionists: The Case of Restricted State Neutrality', *Ethics* **120**, 232–256.

Warfield, T. A. (2005), 'Knowledge From Falsehood', *Philosophical Perspectives* **19**(1), 405–416.

Westen, P. (2008), 'Individualizing the Reasonable Person in Criminal Law', *Criminal Law and Philosophy* **2**(2), 137–162.

Williamson, T. (2000), *Knowledge and Its Limits*, Oxford University Press.

Wolf, S. (1990), *Freedom Within Reason*, Oxford University Press.

Wood, D. (2010*a*), 'Punishment: Consequentialism', *Philosophy Compass* **5**(6), 455–469.

Wood, D. (2010*b*), 'Punishment: Nonconsequentialism', *Philosophy Compass* **5**(6), 470–482.

Worsnip, A. (2016), 'Moral Reaasons, Epistemic Reasons, and Rationality', *Philosophical Quarterly* **66**(263), 341–361.

Zagzebski, L. (1996), *Virtues of the Mind: An Inquiry Into the Nature of Virtue and the Ethical Foundations of Knowledge*, Cambridge University Press.