It Would be Bad if Compatibilism Were True; Therefore, It Isn't

Forthcoming in *Philosophical Issues* (special issue: Free will), eds. Michael McKenna and Carolina Sartorio

I want to suggest that it would be bad if compatibilism were true, and that this gives us good reason to think that it isn't. This is, you might think, an outlandish argument, and the considerable burden of this paper is to convince you otherwise. There are two key elements at stake in this argument. The first is that it would be – in a distinctive sense to be explained – *bad* if compatibilism were true. The second is that the fact that it would be – in this sense! – bad if true gives us reason to think that it isn't. It may be bad that there is no afterlife. But that, in itself, hardly gives us reason to think that there *is* an afterlife. That is true, but – as others before me have suggested – when the object of the relevant badness is *morality itself*, the inference seems secure.

But I'm getting ahead of myself. The structure of the paper is the following. I first explain how I came to the judgment that it would be, in the relevant sense, *bad* if compatibilism were true. The story here involves reflection on longstanding worries about compatibilism and the possibility of *global manipulation*. The thought is that compatibilism ultimately presents us with a picture on which, in principle, powerful manipulators can effectively *guarantee* that finite moral agents should become blameworthy. To my mind, this isn't just false – though I think that it is – it is also such that it would be bad (unfortunate, undesirable…) if it were true. I then turn to the second element of the relevant argument, and make a crucial comparison between the argument I develop here, and Geoffrey Sayre-McCord's so-called "moral argument against moral dilemmas". My dialectical aim is (more or less) to show that my argument is (in principle) no more ridiculous than the argument offered by Sayre-McCord. More generally, my aim is to show that there are striking parallels between the relevant argument against the possibility of moral dilemmas, and the argument I develop against compatibilism. Many already feel the force of the intuition that morality *shouldn't* allow for genuine moral dilemmas – that it would be *bad* if moral dilemmas were really possible – and many (perhaps implicitly) already accept that this seems to suggest that there *can't* be genuine moral dilemmas. I cannot hope to fully defend this argument here – but I do hope to show that there exists a parallel argument against compatibilism.

Bad – but in what in sense and why?

Let me begin, oddly, with some biography about how this (admittedly strange) argument occurred to me in the first place. It all started *via* reflection on so-called *manipulation arguments* for incompatibilism about moral responsibility and determinism. I cannot hope to summarize the extensive discussion about manipulation arguments in this paper; the following (rough and incomplete) sketch must suffice. The idea behind such arguments is *prima facie* simple. If you show me standards for moral responsibility that are compatible with determinism, then I will show you a powerful neuroscientist, AI-overlord, god (or demon!) who can effectively micromanage or otherwise pre-plan some given agent's every move, whilst nevertheless leaving that agent free and responsible by those compatibilist standards. In other words, the picture we get from compatibilists is one according to which a fully responsible, free agent can nevertheless be one whose every move is pre-determined by the relevant "manipulators". Now, the thought behind the relevant incompatibilist strategy is the following. Intuitively, someone whose every thought and deed was "pre-programmed" in this kind of way is neither free nor responsible – that is, is someone we couldn't possibly fairly blame. We can't blame Judas once it becomes clear that his actions were part of the relevant divine plan – even if Judas (*inter alia*) did what he did "because he wanted to", and otherwise met the various compatibilist conditions.

Well, maybe, or maybe not.[1] But back to biography. My own first-order intuitions on these matters were squarely incompatibilist.[2] When I reflected on the relevant cases, it seemed to me as if the given agents *were not* morally responsible for what they do. But as my preoccupation with manipulation arguments grew, at some point a different thought crept in – a thought perhaps *related* to that first thought, but one that is nevertheless phenomenologically distinct. The thought was this. I don't want morality to be like that! I don't want it to be the case that – say – a powerful demon could, in principle, simply *create* the conditions sufficient for someone to be blameworthy; that is, I don't want it to be the case that such a demon could – in principle! – create carbon-copy agent after carbon-copy agent, all of whom were guaranteed

[1] An (I think eminently plausible) assumption I am making here is that compatibilists will have to give a so-called "hard-line" reply to at least *some* manipulation scenarios; cf. McKenna 2008 and Pereboom 2014: Ch. 4.
[2] Todd 2011, 2012, 2013, 2017.

by the conditions laid down by the demon to become blameworthy.[3] I don't want things to be like that! By which I don't mean – or don't just mean – that I don't want there to be such a demon (although I don't); I don't want it to be that if there *were* such a demon, that that demon would have this kind of moral power, the power to effectively settle the question of whether a given agent should become blameworthy. I don't want this to be possible. How horrible that morality should enable such dreadful moral power![4] Instead, I want it to be (and I apologize that I cannot defend this metaphor further…) that there is some kind of "moral inner citadel" such that one becomes blameworthy only if that citadel is overthrown – but that it can only be overthrown from the inside. More specifically: I don't want it to be that the demon could guarantee that the relevant agents become blameworthy *without taking any risk that they should instead become praiseworthy*. But that is exactly what the compatibilist tells us: such a demon needn't take any such risks at all. As I said: I don't want morality to be like that! There is nothing within the purview of morality itself that can prevent there being a demon who wants to create someone who eventually becomes blameworthy. But I want morality to at least force that demon to *take a risk*.[5]

But now the crucial next step. I didn't want morality to be like that. But there was something further. I felt that I was *right* for not wanting morality to be like that – that I was *right* for wanting it to be false that responsibility was consistent with this kind of manipulation (and therefore also determinism). And then I thought the following. How could it be that I am *right* to want compatibilism to be false – and yet it is also true? Wouldn't that be extraordinarily strange? Of course, there are any number of propositions that I rightly wish were false, but are unfortunately true; I rightly wish that there was world peace – but there's no world peace. But the present case seemed different. The difference is elusive, but the first pass thought appears simple. There is a difference between *morality* and *the world*. If I am right to not want compatibilism to be true, and yet it *is* true, then doesn't this amount to the

[3] For instance, I don't want it to be that the agents in the relevant "pods" described in the science-fiction thought experiment of Todd 2019 are genuinely blameworthy; I don't want the relevant "universe-architect" to be able to generate more and more blameworthiness *ad nauseam*, without any corresponding risk that there should instead become praiseworthiness.

[4] A potentially similar intuition arises with respect to moral blackmail (cf. McConnell 1981): how horrible that morality should enable (certain kinds of) moral blackmail! Unfortunately I lack the space to draw out this comparison.

[5] Cf. the "equal moral opportunity" view defended by Swenson 2022; cf. also Diamintis 2023. The thought here is the inverse of a point familiar from the project of theodicy: if God wants to create agents who become praiseworthy, the moral facts imply that even God will have to risk that they instead become blameworthy.

paradoxical situation in which – and how else to say it? – *morality itself* is …, well, immoral? But how could it be that morality itself is immoral?[6]

But let me back up and clarify. My feeling was that it would be *bad* if powerful manipulators could (in the relevant way) effectively settle that someone should become blameworthy. It is critical to appreciate the (alleged) grounds of this badness. When I say that it would be bad if manipulators had this kind of power, this badness is *not* grounded in the fear that there may actually be any such dreadful manipulators. I am relatively certain that there are no (and it is unlikely that there shall become) malevolent neuroscientists who might engage in this kind of manipulation; I am confident that there are no parallel AI-overlords, or gods, or demons. I am *not* worried that if compatibilism is true, then *our* blameworthiness is now within the (ultimate, long-term) control of a global manipulator. Thus, when I say that it would be "bad" if compatibilism were true, I certainly do not mean "bad" in any mundane consequentialist sense concerned solely with negative outcomes in the actual world. I mean something frankly much more exotic, and much more difficult to understand. I mean that, quite independently of whether in fact there exists (or are likely to exist) the relevant manipulators, it is just intrinsically a bad thing – a kind of design-flaw in morality itself – that it should be possible for such manipulators to be able to determine someone to become blameworthy. This is a bad-making feature of a moral system, independently of whether the powers granted by this system are ever in fact utilized. The goal is to give sense to this inchoate thought.

Moral dilemmas

Perhaps you are puzzled, and if so, I don't blame you. But perhaps we can get traction on the argument here by considering a parallel argument in a similar domain. Geoffrey Sayre-McCord has recently articulated what he calls a moral argument against moral dilemmas – just as I am offering what amounts to a moral argument against compatibilism.[7] I am in wholehearted

[6] Of course, I don't mean that, on this view, "morality" has performed an immoral action; that plainly makes no sense (cf. van Someren Greve 2014 (b): 916). Hence the discussion in this paper.

[7] An alternative possible precedent may be Thomas Nagel's "curious" argument that we have *rights*:

> We can distinguish the desirability of not being tortured from the desirability of its being impermissible to torture us; we can distinguish the desirability of not being murdered from the desirability of our murder's being impermissible. These are distinct subjects, and they have distinct

agreement with the spirit of the argument, but I am less sanguine about Sayre-McCord's diagnosis of the moral picture that supports it. But first my own account of the basic idea.

The basic idea is simple. Many philosophers instinctively recoil against the idea that there could be *genuine moral dilemmas*, which we shall define as circumstances in which an agent, through no fault of her own, is blameworthy no matter what she does.[8] Now, Sayre-McCord suggests (plausibly) that the argument against the possibility of dilemmas is neither logical nor structural: it is instead moral. It would be, in a certain distinctive way, *bad* if moral dilemmas were possible. Here is Sayre-McCord:

> The argument is prompted by a gut reaction. And the challenge I am trying to meet is that of turning the gut reaction into an argument. So let me begin by highlighting the gut reaction. It is that, when it comes (for instance) to Sophie's situation, the problem with thinking she faced a genuine moral dilemma is that it miscasts a victim as a wrong-doer, and that it does so by imposing standards and demands on her that are

---

values. To be tortured would be terrible; but to be tortured and also to be someone it was not wrong to torture would be even worse.

This is a curious type of argument, for it has the form that P is true because it would be better if it were true. That is not in general a cogent form of argument: One cannot use it to prove that there is an afterlife, for example. However, it may have a place in ethical theory, where its conclusion is not factual but moral. It may be suitable to argue that one morality is more likely to be true than another, because the former makes for a better world than the latter – not instrumentally, but intrinsically. This would require that we be able to conceive and compare alternative moral worlds, to determine which of them is actual. I will not attempt a full defense of the idea here. (1995: 111)

Nagel's argument: it would be bad if we didn't have rights, because then we would be *less valuable* – we would be people it wasn't even always wrong to torture! This may be so, but note: my argument is not motivated by the contention that we are *less valuable* on compatibilism than we are on incompatibilism. The badness of compatibilism, on my account, is not grounded in any negative *axiological* consequences vis-à-vis moral agents – say, that it makes such agents "responsibility-preserving manipulatable", which is worse than being "unmanipulatable" in this sense. (For criticism of Nagel's argument, see McNaughton and Rawling 1998.) As we will see below, however, disentangling my argument from questions of "value" is a fraught affair; ultimately I suggest that the argument I develop relies on the key contention that it worse to be blameworthy, in some sense, than it is good to be praiseworthy. But note: this fact about value does not function in the same way as the relevant fact about value functions for Nagel. To employ our thought experiment involving the "construction" of the relevant moral domains: it is as if Nagel is thinking that if we could endow agents with rights, we will thereby *make* them more valuable. My appeal to value is different: because it *is* worse to be blameworthy than it is good to be praiseworthy, this value helps to *give us reason* to construct the standards of blameworthiness so that they are not consistent with manipulation/determinism. More generally: by respecting the value of agent's blameworthiness being under their (sole) control, we needn't be making those agents *more valuable*.

[8] The focus on *blameworthiness* is important here (cf. Sayre-McCord's understanding provided below); some might maintain that, in "moral dilemmas", one does something *wrong* no matter what one does – but one isn't *blameworthy*. I have little moral objection to the existence of "dilemmas" in this weak sense.

> impossible to satisfy. Slightly shifting the reaction to another context, consider a God that, first arranges things so that his creation cannot meet all his commandments, and then, second, condemns them for their failure to live up to those commandments. A God who ensures that you are damned if you do, and damned if you don't, is a God against whom one has a legitimate complaint. Or consider a legal system structured so that no matter what one does, one violates the law and is, as a result, subject to punishment. Here too one would have, I suggest, a legitimate moral complaint.

The appeals to God and the legal system are both instructive; more on this below. Sayre-McCord continues:

> In all these cases, I find myself thinking that there is something wrong with the system of norms, or the God, or the laws, in question. As I will put it, there is something "unfair" about them, and it is traceable to their demanding the impossible. Each of these situations is one in which the impossible is demanded, and they are morally objectionable on those grounds. I don't mean to put too much weight on the term "unfair" here. It has connotations, and perhaps implications, that are not important to the argument. The point is simply that there is something wrong, something morally objectionable, about demanding the impossible. And the gut reaction, which I hope to turn in to an argument, is that the problem with thinking there are moral dilemmas is that thinking there are, is thinking that morality is, itself, unfair (or otherwise morally objectionable).

This is a gut reaction with which I am entirely sympathetic: (1) a morality that permits moral dilemmas is itself morally objectionable; and (2) it is impossible that morality should be objectionable in this way.[9]

[9] We noted above the comparison to Nagel on rights. There are other discussions and arguments very much in the vicinity. What McConnell (1985) calls "meta-prescriptions" are in a sense norms for normative theories. Daniel Statman (1995: 45 - 6) gives a highly similar argument for the principle that Ought Implies Can: on his account, OIC amounts to a second-order principle that seeks to limit first-order theories regarding the range of their commands, more or less on grounds that it is *bad* that someone should be obligated to do what he cannot do. (For a reply to a similar argument from Copp 2003, see Van Someren Greve 2014 (b)). Enoch (2009) briefly defends the cogency of some arguments of the form, "It would be good that *p*, therefore *p* (for a moral *p*)". Van Someren Greve (2011) replies that Enoch's argument would allow us to conclude that this is the best of all possible worlds; Bruno (2022) – whom I follow in this respect

In defense of (1), we can perhaps rest content with the intuition of unfairness. To this plausible thought, however, I would like to add another. It is a bad thing that moral dilemmas should be possible, because then morality will have a structure according to which a suitably positioned demon could (echoing our previous discussion) simply *generate* blameworthiness *ad nauseam*: this would be a picture according to which my blameworthiness is in principle fully within the control of *someone else*, someone in whose power it is to see to it that I am in a moral dilemma. Yet worse: the possibility of dilemmas will grant to any such demon the power to turn any finite agent, not only into sinner, but a *monstrous* sinner, simply by ensuring that this agent is placed in moral dilemma after moral dilemma until the end of time. On the given picture, such an agent is fated to become more and more blameworthy, through no antecedent fault of her own. This situation strikes me as intolerable, and only an inexcusable design-flaw in the moral order could explain it.[10] Morality should not place one's *basic moral status* – one's status as blameworthy or praiseworthy – within this kind of direct control of someone else.

The legal analogy is instructive. Suppose I am clerk at the Supreme Court. And the following curious situation arises. I notice that if a given case is decided this way rather than that, then it will be entirely possible for certain blameless agents to be caught up in legal dilemmas – situations in which, no matter what they do, they do something illegal, and are thus subject to the distinctive penalties of breaking the law. For instance, suppose I point out that if the Court upholds a given statute, then a given resident of California will be required by the Federal Government to report for military service, but prohibited by the state of California from reporting for that service – and that both laws will be applicable to one and the same agent at one and the same time.[11] It is highly plausible that this situation would amount to a kind of design-flaw in the legal system – a kind of flaw which we may tolerate only if there were some good reason why we must do so. In short, the lawmakers plainly have reason – defeasible

---

– plausibly replies that we must restrict the relevant inference to deontic (not axiological) *ps*. Hendricks 2021 considers – in a very different sense than the one at issue here – whether it would be better if the pro-choice position were correct, or instead the pro-life position, argues that it would be better if the former were true, but concludes that this is unfortunate, because it is likely that it isn't. Blanchard 2020 considers whether *meta-ethical* theories might be subject to normative criticism.

[10] To repeat: it is not in the purview of "morality" to ensure that no demon has the power to place an agent in putative moral dilemma after putative moral dilemma until the end of time: morality has no choice about *that*. (To prevent analogous forms of abuse is the job of the police, not morality.) All morality has a choice about, as it were, is whether these putative moral dilemmas are *genuine* moral dilemmas; it can decide that, yes, blameworthiness attaches to this agent, or it can say that no, it doesn't.

[11] I owe this example (lightly modified) to Jackson 2009, which itself criticizes Sayre-McCord's moral argument against dilemmas in highly interesting ways.

reason, perhaps, but nevertheless reason – to make the laws *consistent*, and to minimize the existence of any legal dilemmas.[12]

Sayre-McCord's ultimate account of these facts is roughly the following; here I must be brief. For any dilemma-allowing theory T1, there is an otherwise similar theory T2 that doesn't allow for dilemmas. Since it should be agreed by all that dilemmas are "unfair", this unfairness must somehow be recognized by T1 itself – in which case it turns out that T2 is the "better" theory even by the lights of T1. The idea here is interesting, but the problem is the suggestion that the possibility of dilemmas is bad *by the lights of T1*. The possibility of dilemmas may be bad by the lights of *proponents* of T1, but these things are importantly different.[13] What Sayre-McCord can get is that for any dilemma-allowing theory T1, there is another non-dilemma-allowing theory T2 that is otherwise similar but more fair. What he *can't* get is that this second theory is more fair *by the standards of T1*. But then we are stuck with our key question: assuming that T2 is more "fair" in this sense than T1 – and other things are equal – why are we entitled to include that T2 is the true moral theory, and not T1?

The Perfection of Morality

But let's back up. Our interest here isn't solely a moral argument against moral dilemmas; our interest is more generally in the very style of argument at issue. Ultimately, I suggest, what we want is an explanation of what we might call *The Perfection of Morality*.

> *The Perfection of Morality* [first pass]: If [perhaps *per impossible*] the commands/policies of morality were the commands/policies of a particular agent, it cannot turn out that that

[12] Again, a further reason for this limitation: to prevent its being that case that a wrongdoer could *see to it* that a person does something illegal (through no previous fault of that agent); consider a legal system which would allow agent A to directly bring about a legal requirement that agent B should be sent to prison. There is one worry about *fairness*, and another about *abuse*.

[13] Preston-Roedder (2014) contends that arguments of the form "it would be good that *p*, therefore *p* (for a certain kind of moral *p*)" (again, see Enoch 2009) have a place in moral theory – but that this form of argument has not been adequately explained and defended. With that much I agree. Preston-Roedder ultimately ends up giving a *self-defeat* account; in the context of Act Utilitarianism (see below), the idea is that AU is "self-defeating" in the following sense: it assigns to us the aim of making the world "better", but since the truth of AU would in itself make the world *worse* (by, e.g., prohibiting moral freedom), it is therefore self-defeating. Problem: AU assigns to us the aim of making the world contain more happiness (say) – not more moral freedom. AU may disallow moral freedom, but this isn't a failure by the lights of AU itself.

agent would have to be (even in part) ignorant, bad, or irrational for having made these commands or selected these policies.

As stated, the thesis is hopelessly vague, but its intent is at least fairly clear: if someone "selected" the truth of compatibilism over the truth of incompatibilism, the thought is that they would be subject to criticism: don't do that! What could justify your decision to select a standard of responsibility that will render it within the power of a demon to guarantee that someone else shall be blameworthy? Similarly: what could justify your decision to select an overall package of moral principles which would make possible the dreadful result that someone could be forced to become blameworthy no matter what she does?

Now our key question: what could explain the perfection of morality, thus construed? Why exactly is this a useful heuristic? I can think of three options, and maybe a fourth, none of which I can explore adequately.

Axiarchism is an extreme minority answer to the vexed question, *Why is there something rather than nothing*? The Axiarchist answers: because it is good that there should have been something. More particularly: the moral requiredness of *there being something* brings it about that there *is* something.[14] Axiarchism is *prima facie* incredible: how could it be that the goodness of there being something, by itself, makes it the case that there *is* something – that is, something concrete, something like the concrete universe? The Good lacks the right kind of power to *cause* the existence of the concrete universe. More particularly, we need something with *causal power* to intervene here – perhaps something that *sees* that it is good that there should be a universe, and then for that reason *makes* a universe. (Viz., God.)

So Axiarchism seems incredible. But *Ethical* Axiarchism is certainly less so:

Ethical Axiarchism: The fact that it would be good that it is fair/right that *p* directly brings it about that is fair/right that *p*.[15]

[14] Axiarchism is most prominently defended by John Leslie (2001), and prominently (albeit sympathetically) criticized by Parfit, in his essay, "Why Anything? Why This?" (reprinted in his 2011: 622 - 48). For discussion, see Russell (2014).

[15] A *comparative* principle may be more suitable here: the fact that it would be *better* that it is fair/right that *p* than that *q* directly brings it about that it isn't fair/right that *q*.

The Ethical Axiarchist limits the power of the Good to the power to bring about the truth of certain *moral* facts – facts like that it isn't possible to be in a moral dilemma, or isn't possible to causally determine that someone else should eventually be blameworthy. (More generally: to constrain the selection of the fundamental principles concerning fairness/rightness.[16]) Ethical Axiarchism thus amounts to a certain primitivist answer to our explanatory question. If our question is, Why is it that when it would be good that it is fair that *p*, it follows that *is* fair that *p*? – the relevant reply is nothing further than: because the fact that would be good that it is fair that *p* directly brings it about that it *is* fair that *p*. The question is whether this is a mystery, and whether this is a mystery with which we can learn to live.

But perhaps these reflections naturally prompt a different thought. Perhaps the Perfection of Morality is grounded in the Perfection of That Which Grounds Morality – viz., the perfection of an Anselmian God, the being than which none greater can be conceived. Perhaps on the relevant kind of Anselmianism, the moral is just one further aspect of – or is otherwise grounded in – the nature of a being than which none greater can be conceived. And observe: in the Anselmian case, there is plainly no cogent distinction (to use Sayre-McCord's phrase) between a theory of a better God and a better theory of God: if a given theory *T* paints a picture of a better God than *T**, then this *just is* decisive reason to think that *T* is a better theory of God than *T**: given the Anselmian understanding, it is straightforwardly conceptually incoherent to suppose that God is a given way, although it would have been better that God should have been some alternative way. God *just is* that being such that it is better that the being be this way rather than any alternative possible way a being could have been. If the existence of such a being grounds the shape of the normative domain, then perhaps the existence of such a being straightforwardly grounds the fact that that shape could have been no better – viz., contains no "design-flaws" of the sort we have been discussing.

The story here might be developed in different ways, some in the fashion of the "voluntarist" or "divine command" traditions in theological ethics, and some in, well, the non-voluntarist traditions. If we take a flat-footed account on which God – who is, again, perfectly good – simply *selects* the true fundamental moral principles, then our job appears simple: we simply must show that God would have had antecedent, undefeated reason to select principles

[16] Again, here and elsewhere we must restrict the relevant *ps* to the *basic principles* of fairness/rightness. It would certainly be good if the division of housework amongst Jack and Jane were fair; this, however, gives us no reason to think that it *is* fair. (Again, see Van Someren Greve 2014 (b).)

which prohibit dilemmas, or prohibit the possibility that someone's blameworthiness should be within the control of someone else (in the relevant way), and so on: this will force God to select principles of fairness on which incompatibilsm is true. However, it seems to me that there are irresolvable problems with this naïve account, none of which I can presently develop. Consider, then, a different account: the relevant principles are ultimately grounded in divine dispositions – perhaps divine dispositions to blame.[17] God, because perfect, is "always already" such that he is not disposed to blame agents who have been caught up in a dilemma – or whose actions were determined (irrespective of whether there should actually exist any such agents). God is like this (and necessarily so) because it is better – not instrumentally, but intrinsically – that God should be like this than otherwise. I will not attempt a full defense of this idea here.

Both options – especially the latter – involve heavyweight commitments. It is worth asking whether we might get by with something lighter, and so let me develop the following thought at somewhat greater length. One suggestion is that the Perfection of Morality falls out of a natural, contractarian/constructivist approach to morality – with a certain modal spin. Imagine that we are legislating the conditions of blameworthiness. Well, since moral truths are necessary truths – which hold for all possible agents in all possible circumstances – we are not merely legislating what will *actually hold* concerning these standards, but what will hold for any agent in any possible world. We thus must give all agents in all possible worlds a voice in our deliberations. To give sense to this thought, we can imagine deliberating behind the modal veil of ignorance.[18] We see that there are possible worlds in which powerful manipulators determine the relevant behavior, possible worlds that are deterministic without such manipulators, possible worlds in which manipulators attempt to get the relevant agents to become blameworthy but leave those agents indeterministically free (and thereby risk that they instead resist their efforts), and more else besides. Well, we don't know which world we'll

[17] The position then becomes a theological counterpart to the "Strawsonian" thesis that the standards of responsibility are determined by the dispositions to blame of actual normal adult human beings. Notably, this position would escape certain worries about variability (cf. Todd 2016) that have plagued the Strawsonian program.

[18] The argument considered here is thus very different in kind from the contractualist argument for compatibilism advanced by Lenman 2006; unfortunately I lack space to make a comparison between my argument and Lenman's. Another way to put the key thought: since the fundamental moral principles shall hold irrespective of what the empirical world happens to be like, the fundamental moral principles must be selected in ignorance of what the empirical world happens to be like.

inhabit; we don't know which of these creatures in which possible world is *us*. How then to choose?

My claim is that there is strong, undefeated reason to adopt a principle that limits moral responsibility to cases in which one has free will as imagined by the libertarian.[19] For the sake of illustration, consider the familiar metaphor of one's "moral ledger". Now, grant that, behind the modal veil of ignorance, one is concerned for one's moral ledger. One wants to adopt a principle that makes it possible to get positive marks on one's ledger[20] – but one naturally wants to minimize the risk of negative marks as well. One must select a coherent principle giving the standards of responsibility for all worlds. And one thinks like this. If I select a compatibilist standard, then if I end up in a world with benevolent forces that determine me to the good, then, yes, I will be praiseworthy. But then if I end up in a world with malevolent forces that determine me to the bad, then I will be blameworthy. There is no more chance of the former than the latter. But now considerations of risk weigh heavily: I am more concerned to avoid becoming blameworthy than I am concerned to secure becoming praiseworthy! On the other hand, if I choose an incompatibilist standard, then in none of the deterministic worlds am I praiseworthy – but of course, in none of them am I blameworthy. In particular, by choosing an incompatibilist standard, I obviate the risk that I shall end up as one of the agents in a possible world who is fated to become horribly blameworthy by malelovent (or perhaps simply indifferent) forces.[21] That is considerable solace. Yes: if I choose an incompatibilist standard, it will still be true that I could end up in a world in which a demon tries to make me blameworthy, but does not *determine* this outcome in advance. (Again, there is nothing morality can do about *that*.) But in that case, there will at least be an objective chance – one within my power to actualize – that I shall instead become praiseworthy. Between the two standards – one compatibilist, and the other incompatibilist – the incompatibilist standard wins.

Parallel remarks might be made about moral dilemmas. Imagine a similar procedure, and imagine someone selecting a set of principles that enables moral dilemmas. And imagine a defense of this selection as follows: this was the only way to secure the possibility of *reverse*

[19] Specifically: the libertarian who endorses the "equal moral opportunity" thesis articulated by Swenson 2022.
[20] It is this, I contend, that tells against adopting a standard on which responsibility is simply *impossible*.
[21] More carefully: I make it the case that any world in which I exist in which I have been pre-determined to perform various immoral actions, I am at least not *blameworthy* in that world. In this world, of course, I do various things that certainly appear to be morally terrible. But my ledger is clean.

moral dilemmas, that is, *immoral dilemmas* – circumstances in which, to no credit of one's own, one is in a situation in which one is praiseworthy no matter what one does.[22] This defense is patently unconvincing. It is a scandal that morality should allow moral dilemmas, but it is not nearly so much of a scandal that it should *fail to enable* immoral dilemmas. (Witness the massive literature on the possibility of moral dilemmas, and witness that, as far as I am aware, there is no literature at all concerned to demonstrate the possibility of *immoral* dilemmas.) That a malevolent demon could generate blameworthiness *ad nauseam* by constructing dilemma after dilemma – that is a horrible result to avoid; that a benevolent god is *not* able to generate praiseworthiness in a parallel way is merely disappointing. Or to put this matter another way: it is not justified to construct a regime that puts blameworthiness within the control of the bad guys, merely to ensure that praiseworthiness is within control of the good guys, *especially* when we are no more likely to be controlled by the good guys than the bad.[23]

Let me clarify the argument here in one respect. In mounting this argument, I do not mean to rely on the judgment that meeting pertinent libertarian conditions gives one *more control* over one's blameworthiness than meeting the relevant compatibilist conditions; this judgment is of course very much contested (cf. "the problem of enhanced control" [Franklin 2011]). I instead rely merely on the (much more plausible) judgment that meeting these conditions does not *detract from* one's control. This claim then supports the key thought: if meeting a libertarian condition is required for blameworthiness, then this fact will give someone else *less control* over one's blameworthiness – *even if* we grant that meeting those

---

[22] I owe these terms to an unpublished paper from Daniel Speak and Manuel Vargas, "Satan's Good Works," in which they consider – amongst other eccentricities – whether God could put Satan in a situation in which Satan is praiseworthy no matter what he does. My first-order intuition: clearly not. My second-order intuition: it is perhaps a bit regrettable that this is morally impossible, but not nearly so regrettable as it would be if it were morally possible that Satan could put someone in a situation (after situation, after situation…) in which she is *blameworthy* no matter what she does.

[23] I mentioned a fourth option. Are we perhaps overthinking this? Perhaps "morality" just *means* "whichever system of principles is best". (If you torture yourself over the question, "How do I know that God is the best a being could be?", many theists will reply: by knowing the meaning of "God" – by knowing that, in the intended sense, "God" just refers to "whichever being is best"; the question whether God exists *just is* the question whether there exists the greatest possible being.) Or perhaps we can pick up on Sayre-McCord's appeal to the set of principles to which we owe "allegiance". Once we realize that theory T1 allows dilemmas, whereas theory T2 doesn't, and things are otherwise equal, then it follows that T1 is *better* than T2, in which case we owe allegiance to T1, not T2. But if "morality" *just is* "the principles to which we owe allegiance", then it will follow from analytic facts alone that T1 is morality, not T2. On this view, the question whether morality exists is simply the question of whether there exists the set of principles that is, in the relevant way, best. But (unlike in the case of God) the existence of the abstract principles – the theories themselves – comes cheap; the question is simply which abstract principles are best. (Let's ignore the possibility of ties at the top.) Could this be all there is to see here?

conditions doesn't give one *more control* over one's *own* blameworthiness. And the key judgment: it is a bad thing that others should have direct control over one's blameworthiness! The suggestion, in short, is that the libertarian conditions fall out of a proper balancing of the following considerations: *give people as much control over their blameworthiness as we can reasonably give, and others the least amount of control over other people's blameworthiness as we can reasonably give.* Compatibilist and libertarian standards may (or may not) be tied on the first front, but they are not tied on the second.

Loose Ends

A full discussion of these issues would have to address the following. We have discussed a moral argument against moral dilemmas, and a moral argument against compatibilism, both grounded in the value of our blameworthiness not being within a certain kind of control of anyone further. But *prima facie* there are other similarly structured arguments discussed (or perhaps merely implicit) in the literature. How does my argument relate to these others? Notably, all have been advanced against various iterations of utilitarianism:

(1) It shouldn't be immoral to promulgate the truth of the true moral theory.[24] [Consider lawmakers who would make it illegal to hear and promulgate the laws.]
(2) The true moral theory shouldn't be *overly demanding.*[25] [Consider laws which are so demanding that compliance becomes irrational even for ordinary citizens in ordinary times.]
(3) The true moral theory should leave room for *moral freedom.*[26] [Consider laws which may be easy to follow, but which nevertheless dictate one's every move.]

My sense is that the arguments against dilemmas and against compatibilism discussed above are in some way crucially different than these characteristic arguments against utilitarianism, but I must set this issue aside.[27]

[24] For discussion in the neighborhood, see Van Someren Greve 2014 (a).
[25] E.g, Portmore 2011.
[26] E.g., Slote 1985: Ch. 2, Vallentyne 2006; cf. Preston-Roedder 2014 for discussion.
[27] At least regarding (1) and (2), the ideal/non-ideal theory distinction appears relevant: the utilitarian may agree that *ideally* it shouldn't be immoral to promulgate the truth of utilitarianism; if it is sometimes impermissible to do so, that is a concession to the ignorance and folly of human agents. Similarly, the

Finally, Daniele Bruno has recently defended a highly similar argument concerning the nature of promissory obligation. It is – at least to some – mysterious why we should have the normative power to control our obligations *via* promising. Bruno's suggestive account: we have these normative powers because it is good that we should have them. To what extent is my argument similar? At some level of abstraction, they can appear just the same. Why should it be that moral agents have the unconquerable power to control their own moral blameworthiness? Because it is good that they should have it. I regret that I lack the space for a further discussion of Bruno's arguments.[28]

Atheism and anti-theism, incompatibilism and anti-compatibilism

Kahane (2011) inaugurated a debate concerning the so-called *axiology of theism*. His question: should we want God to exist? Will it be better if God exists than does not? I am tempted to construe my intent in analogous fashion: to inaugurate a debate concerning the *axiology of responsibility*. Should we want compatibilism to be true? Will it be better that compatibilism is true than incompatibilism? But with the following key *proviso*: not "better" instrumentally, but intrinsically. (Not "want" for our own sakes, but want *for all possible worlds*.) When one asks (in the relevant way) whether it will be "better" if God exists than does not, one needn't be asking a question that is sensitive to substantive content about one's actual position in the world; one needn't be asking whether it will be *better for me* that God should exist, or better *for humanity* (although one could do that). One instead may ask a question at a substantially more confusing level of analysis: will it be better *per se* that God should exist rather than not? The question is intelligible, I believe, but is one we struggle to grasp, as the growing literature on this subject (cf. Kraay 2018) reveals.[29]

---

utilitarian may agree that the true moral theory shouldn't be overly demanding *given ideal compliance*. I am less sure that this distinction has anything to do with (3) – but more to the point, I am unsure how the ideal/non-ideal distinction is relevant to the arguments against dilemmas and against compatibilism. I am grateful to David Enoch for useful discussion on these points.

[28] Cf. also Tiefensee 2019: 878, who remarks, "there is an important moral constraint on moral theories, such that a moral theory cannot be true if there is another moral theory which is morally better."

[29] Also highly relevant is Kahane 2012; the question there is whether we can intelligibly compare different metaphysical theories in terms of their value: is the world better, say, on idealism than it is on materialism? The project here assumes that we can – in some analogous fashion – compare the moral consequences of different *moral* theories: is the world better on incompatibilism than on compatibilism? (Cf. the quote from Nagel above.) But again our key point: since one might think that idealism/materialism are necessarily true (if true at all), one is really asking some different question: will the *pluriverse* be better on idealism than it is on materialism? Likewise: we aren't merely asking whether the *actual world* will be better given this theory of

Three further notes on these themes. First, it is perhaps possible that *knowing what we know about the conditions that actually prevail,* we should want compatibilism to be true. (Only then, perhaps, are *we* morally responsible.) But that is irrelevant to the axiological question I mean to raise: for this question must be asked and answered behind the modal veil of ignorance. Behind the modal veil of ignorance, I contend, one should want compatibilism to be false.

Second: the literature on the axiology of theism makes a distinction between atheism – the thesis that God does not exist – and *anti-theism* – the thesis that it would be *bad* if God existed, that we should not *want* God to exist. And just as we can make a distinction between atheism and anti-theism, so we can make a distinction between incompatibilism and anti-compatibilism. And just as it is an interesting and difficult question whether the truth of anti-theism would support atheism, so it is an interesting and difficult question whether the truth of anti-compatibilism would support incompatibilism. The argument of this paper is that anti-compatibilism is true, and that anti-compatibilism supports incompatibilism.

Finally: the comparison here is instructive and illuminating, but we shouldn't put more weight on it than it can bear. There is an enormous difference in kind behind the possible ramifications of the existence of a maximally perfect concrete being with the power to have causal effects, and the possible ramifications of the truth of moral principles which are by everyone's lights causally inert.

Conclusion

There have been several themes in this paper: (1) on compatibilism, the value of the pluriverse – the space of all worlds – is less than it would be on incompatibilism; and this is an unforced error that morality needn't make, and (2) from behind the modal veil of ignorance, we should – because of the relative badness of blameworthiness versus the relative value of praiseworthiness – select an incompatibilist standard of responsibility, and (3) it is intrinsically a bad thing that an agent's blameworthiness should be within someone else's control, perhaps just in itself – but also at least in part because then a controlling agent with enough power

---

responsibility or that, but whether the space of all worlds will be better, whether the geography of the pluriverse – to use Lewis' memorable phrase – will be better. It is hard to feel confident that we know what we are asking when we ask these questions, and hard to answer them even if we feel like we know.

could generate the disvalue of blameworthiness *ad infinitum* and without a corresponding chance of generating praiseworthiness instead. These thoughts seem to me to be different sides of the same (multi-sided, complex) coin, but I admit that the discussion above doesn't come close to fully illuminating these connections.

The argument developed above plainly relies on a range of difficult theses, many of which I have not so much as adequately explained, let alone adequately defended. It would be a mistake to pretend otherwise. Still, the possibility of this style of argument is difficult to unsee once you've seen it, and the argument has the potential to be extremely psychologically powerful for anyone who falls under its spell. Of course, one might seek to explain how the truth of compatibilism does not amount to some kind of "design-flaw" in the moral order – but it is psychologically difficult to think that morality itself is somehow flawed, disappointing, or regrettable. That would be like getting to heaven, beholding the face of God – and *grimacing*. But I do not want to have to grimace when contemplating the nature of the standards of moral responsibility. But that is, I fear, exactly what I must do if those standards are compatibilist. We shouldn't want compatibilism to be true. Thankfully, I contend, that gives us good reason to think that it isn't.[30]

References

Blanchard, Joshua. 2020. "Moral Realism and Philosophical Angst," in *Oxford Studies in Metaethics* (vol. 15), ed. R. Shafer-Landau.

Bruno, Daniele. 2022. "Value-based accounts of normative powers and the wishful thinking objection," *Philosophical Studies* 179: 3211 – 3231.

Copp, David. 2003. '"Ought" Implies "Can", Blameworthiness, and the Principle of Alternate Possibilities'. In McKenna, M. and D. Widerker (eds.), *Moral Responsibility and Alternative Possibilities* (pp. 265-299). Aldershot: Ashgate Press.

Diamintis, Mihailis E. 2023. "The Moral Irrelevance of Constitutive Luck," *Erkenntnis* 88: 1331

[30] There are only so many times one can say in a paper, "I regret that there is no space for further discussion"— and yet, if there were space for further discussion, I would have felt a burden to fill it, in which case this paper would likely have never been written. I have been thinking on and off about these themes for over 10 years, and have many people to thank for helpful discussion, esp. John Fischer, Alex Pruss, Stephen Finlay, Terrance McConnell, Guy Kahane, Dana Nelkin, Andrew Chignell, Kristin Inglis, Andrew Bailey, Kenny Boyce, Alex Arnold, Philip Swenson, Neal Tognazzini, David McNaughton, Mike Ridge, Michael McKenna, Daniele Bruno, and David Enoch. I first presented "The Perfection of Morality" in November 2012 at Oxford's Moral Philosophy Seminar; I thank the audience for useful feedback.

– 1346.

Enoch, David. 2009. "Wouldn't It Be Nice If *p,* Therefore, *p* (for a moral *p),*" *Utilitas* 21: 222 – 224.

Franklin, Christopher Evan. 2011. "The Problem of Enhanced Control," *Australasian Journal of Philosophy* 89: 687 – 706.

Hendricks, Perry. 2021. "The Axiology of Abortion: Should we Hope Pro-Choicers or Pro-Lifers Are Right?" *Ergo* 7: 774 – 788.

Jackson, Vincent P. 2009. "Inescapable Wrongdoing and the Coherence of Morality: An Examination of Moral Dilemmas," MA Thesis, Virginia Tech University.

Kahane, Guy. 2011. "Should we want God to exist?" *Philosophy and Phenomenological Research* 82: 674 – 696.

Kahane, Guy. 2012. "The Value Question in Metaphysics," *Philosophy and Phenomenological Research* 85: 27 – 55.

Kraay, K.J. Ed. (2018) *Does God Matter? Essays on the Axiological Consequences of Theism*. Routledge.

Lenman, James. 2006. "Compatibilism and Contractualism: The Possibility of Moral Responsibility," *Ethics* 117: 7 – 31.

McConnell, Terrance. 1985. "Metaethical Principles, Meta-prescriptions, and Moral Theories," *American Philosophical Quarterly* 22(4): 299-309.

McKenna, Michael. "A Hard-line Reply to Pereboom's Four-Case Manipulation Argument," *Philosophy and Phenomenological Research* 77: 142 – 159.

McNaughton, David and Piers Rawling. 1998. "On Defending Deontology," *Ratio* 11: 37 – 54.

Nagel, Thomas. 1995. "Personal Rights and Public Space" *Philosophy and Public Affairs* 24: 83-107.

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

Portmore, Douglas. 2011. "Consequentialism and Moral Rationalism," *Oxford Studies in Normative Ethics* vol. 1 (ed. Mark Timmons).

Preston-Roedder, Ryan. "A Better World," *Philosophical Studies* 168: 629 – 644.

Roberts, John Russell. 2014. "Axiarchism and Selectors," *Faith and Philosophy* 31: 412 – 421.

Sayre-McCord, Geoffrey. Ms. "A Moral Argument Against Moral Dilemmas." UNC-Chapel Hill.

Statman, Daniel. 1995. *Moral Dilemmas* (Value Inquiry Book Series 32). Amsterdam: Editions

Rodopi.

Tiefensee, Christine. 2019. "Relaxing About Moral Truths," *Ergo* 6: 869 – 890.

Slote, Michael. 1985. *Common-Sense Morality and Consequentialism*. London: Routledge.

Todd, Patrick. 2011. "A New Approach to Manipulation Arguments," *Philosophical Studies* 152: 127 – 133.

Todd, Patrick. 2012. "Manipulation and Moral Standing: An Argument for Incompatibilism," *Philosophers' Imprint* 12: 1 – 18.

Todd, Patrick. 2013. "Defending (a modified version of) the Zygote Argument," *Philosophical Studies* 164: 189 – 203.

Todd, Patrick. 2016. "Strawson, Moral Responsibility, and the 'Order of Explanation': An Intervention," *Ethics* 127 (1): 208-240.

Todd, Patrick. 2017. "Manipulation Arguments and the Freedom to do Otherwise," *Philosophy and Phenomenological Research* 95: 395 – 407.

Todd, Patrick. 2019. "The Replication Argument for Incompatibilism," *Erkenntnis* 84: 1341 – 1359.

Vallentyne, Peter. 2006. "Against Maximizing Act-Consequentialism," in *Contemporary Debates in Moral Theories,* ed. James Dreier. Blackwell.

Van Someren Greve, Rob. 2011. "Wishful Thinking in Moral Theorizing: Comment on Enoch," *Utilitas* 23: 447 – 450.

Van Someren Greve, Rob. 2014. (a) "The Value of Practical Usefulness," *Philosophical Studies* 168: 167 – 177.

Van Someren Greve, Rob. 2014. (b) "'Ought', 'Can', and Fairness," *Ethical Theory and Moral Practice* 17: 913 – 922.