Living with ourselves, together

Alexander Edward Horne

Trinity College

September 2022

This thesis is submitted for the degree of Doctor of Philosophy

- This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the preface and specified in the text.
- It is not substantially the same as any work that has already been submitted before for any degree or other qualification except as declared in the preface and specified in the text.
- It does not exceed the prescribed word limit for the Philosophy Degree Committee.

I would like to thank Professor Paulina Sliwa and Professor Richard Holton for supervising this dissertation.

Summary

This PhD dissertation comprises five chapters on a variety of intersecting topics within moral psychology, metaethics, and epistemology. Broadly speaking, the first half of the dissertation focuses on individual agency, normativity and the self; while the second half explores aspects of our normative lives as social agents. More specifically, the first half contains papers on underdetermination by value, self-improvement, and the problem of self-creation; while the second half contains papers on epistemic angst as a social problem and the normativity of social norms. A theme of the first half is that the internal dynamics of individual agency share more in common with inter-agential social dynamics than is usually thought. A theme of the second half is that the reasons emerging from inter-agential social dynamics are more pervasive and powerful than many working in metanormative theory have allowed. Together, the two halves of the dissertation lay some of the groundwork for a novel account of human agency, together with an improvement on our understanding of the nature of robust normativity.

In Chapter 1, I articulate and attempt to solve the problem of rational underdetermination as it confronts idealizing subjectivists. In Chapter 2, I introduce a novel thought experiment designed to sound some skeptical notes about self-improvement and interrogate their significance for our understanding of the relationship between agency, identity and self-improvement. In Chapter 3, I criticise a recent, prominent solution to the old problem of self-creation and propose an alternative I label 'indirect evaluative voluntarism'.

In Chapter 4, I pivot to problems generated by our interactions with other agents. I attempt to establish a claim of *a posteriori* necessity regarding social norms' reason-giving power that follows from the best account of what they and we are like. My argumentative strategy for establishing that conclusion is to show that the relevant instrumental normativity is simply contingent on a human agent's having any desires whatsoever, on the model of a universal hypothetical imperative. In Chapter 5, I articulate a distinctively social, epistemic form of angst and use it to explain some communities' distrust of experts and one another; identify three structural problems manifesting epistemic angst; sketch a partial solution to it; and explain what remains to be done to solve it. Chapter 1 now appears in *Synthese* as 'Too many cooks'.

CHAPTER 1: TOO MANY COOKS

1. Introduction

This morning, I have options.[1] Salient among them: work on this paper, read a novel, go for a run or call my folks. All seem permissible, none seem obligatory. It is tempting to think there is no fact of the matter about what I should do this morning, all things considered. Each option, let's suppose, is adequately supported by reasons of roughly similar weight (Portmore, 2011; 2013).[2] But this indeterminacy is untroubling. In part, because it is utterly unsurprising and in part, because I have time to perform all four options today. Also comforting is that the order in which I perform them hardly seems to matter. But, of course, there are many other valuable things I could be doing with my time: volunteering at a soup kitchen, writing that novel, coaching a little-league team, making my family breakfast, marking, reading a colleague's paper, donating money to charity, spending money getting to the Galapágos Islands, saving money for the future, giving-up philosophy to become a doctor and so on. There are many valuable things one can do in any given moment and in life generally, but one cannot do them all. So, at a whole-life level of analysis too, it is tempting to think there is no unique, best option and so, no fact of the matter about what I should do, all things considered. There will be many comparably valuable options and there will incommensurable options and there will be incomparable options (Chang, 2002). Moreover, since every life contains some 'recharging time' and since the distribution of recharge time can vary greatly over equally valuable lives, it turns out that on any given morning I have just as much reason to sit on my couch drinking beer and watching *Sharknado* as I do volunteering at Animal Rescue (Gert, 2018; Portmore, 2013).

For those made of stern stuff, this indeterminacy might be liberating. Myself, I tend towards the existential dread end of the scale. But that's not the point. The point is that underdetermination threatens our ability to make non-arbitrary decisions about what to do. Flipping a coin to decide between equally appealing boxes of cereal seems sensible (Holton, 2006; Ullmann-Margalit & Morgenbesser, 1977). Flipping a coin to decide between different

[1] For comments on earlier drafts of this paper I am indebted to Richard Holton, Yarran Hominh, Nikolas Kirby, Claire Kirwin, Nikhil Krishnan, Paulina Sliwa and *Synthese's* anonymous referees, together with its editors.

[2] In fact, Joseph Raz gives significant weight to what he labels 'the basic belief', namely, that 'most of the time people have a variety of options such that it would accord with reason for them to choose any one of them and it would not be against reason to avoid any of them' (Raz, 1999, p. 100). Cf. (Raz, 2012).

lives seems insouciant. If, moreover, what I should do today is a function of what I should do this week and what I should do this week, of this month and this month, of this year and so on, then indeterminacy is rife. It is not just that the weight of our reasons underdetermines which of a manageable number of options we should choose. It is that:

> At any point in time in your life, many different choice situations are open to you, each of which you have sufficient reason to be in relative to agential values. If there is no more reason for you to be in one rather than another, the trajectory of your life, made up of the choices you make in arbitrarily selected choice situations, is itself profoundly arbitrary… But this makes a travesty of the human condition. (Chang, 2021, p. 105)

Call this 'the rational underdetermination problem'. The existing literature on the rational underdetermination problem often construes it as one resulting from the ubiquity of objective values. It is therefore sometimes argued that subjectivists need not be troubled by the underdetermination problem. But on closer examination, it turns out, they should. Or so I will argue. The task of the first half of this paper is explaining why. The task of the second half is finding a subjectivist solution the rational underdetermination problem. The basic argument is as follows: if subjectivists do not idealize, then their view is implausible.[3] So, subjectivists should idealize. But idealization is notoriously tricky (Enoch, 2005; Dorsey, 2017; Sobel, 2016).[4] Crucially for the argument of this paper, it is often difficult to generate a single, unified evaluative perspective from which to assess an agent's options (Rosati, 1995). Our actual evaluative perspective underdetermines possible idealizations of it. Worse still, inter-perspectival comparisons are mostly blocked by incommensurability. So, I will suggest, idealization leads to indeterminacy and, since there is no further evaluative perspective one can occupy from which the deliverances of these other perspectives can be assessed, we are stuck. Each of us has many possible ideal selves, they all disagree about what to do, and too many cooks spoil the broth. In response: these idealized options should be assessed from the only

[3] Not least, one might worry, because the resultant overgeneration of reasons entails indeterminacy regarding what to do

[4] For example, does idealization involve the actual agent in counterfactual circumstances of full information; an idealized deliberating agent in the actual circumstances of the present agent; or an idealized advisor considering the options of the actual agent? If idealization involves ideal deliberation, what exactly does that involve? Coherence constraints will obviously have a role to play, but does idealization also involve the revision of an agent's current beliefs and attitudes in light of new beliefs and attitudes generated by imaginative acquaintance plus closure?

perspective we can, in fact, occupy, namely, that of our non-ideal, actual selves. Deciding what to do therefore becomes, in part, an exercise in deciding who to be. But one might now worry this just moves the arbitrariness bump in the rug. Privileging the perspective of our actual self seems contrary to the rationale for idealizing in the first place. I consider two solutions to the problem, one democratic, the other modelled on trusteeship. In the end, I argue, our actual self has complete freedom to choose the ideal self it grants rational authority. In the final part of the paper, I present my positive proposal as a solution to the underdetermination problem confronting the idealizing subjectivist and then argue that, so understood, this account vindicates a tidied-up version of how some reflective people already do deliberate in their everyday lives. This, in turn, suggests that a decision-procedure closely connected to the account is both possible (because actual) and attractive.

2. Underdetermination by reasons

Not everyone is as squeamish about rational underdetermination as I am. Some even think it is a pro of objectivist theories that they generate significant normative wiggle-room. Joshua Gert provides an excellent summary of how this issue plays out between objectivism and subjectivism:

> It is somewhat surprising that objectivism should go hand-in-hand with the idea that reasons—and the values that give rise to them—do not determine how we ought, rationally, to act. After all, objectivity seems intuitively to be associated with determinacy. Perhaps equally surprisingly, the converse of the link between objectivity and rational underdetermination also seems to hold. That is, if one takes reasons to be subjectively determined by one's contingent desires, it then turns out to be natural—not to say obligatory—to say that we have *fewer* rational options than the objectivist countenances. Why? Because it is more plausible that one's desires are characterizable in terms of a single, commensurable sort of strength than it is that objective values are. And if one's desires are characterizable in this way, and if one's reasons are a function of these desires, then it is a natural consequence that one's reasons also will have unique and commensurable strength values. To the degree, then, that one thinks that we do in fact have a great deal of rational latitude in our everyday choices, one has reason to reject subjective, desire-based theories of reasons, and to accept a more objective, desire-independent view. (Gert, 2018, pp. 459-460)

Of course, some subjectivists are likely to draw a different conclusion, namely, that it is a significant advantage of their theory that there are often determinate answers to questions about what one should do, all things considered, and that these answers are not alienating or arbitrary from the perspective of the agent subject to them. But, I will argue in the next section, the appearance of increased commensurability to which Gert alludes is misleading. As a result, subjectivists must also confront the rational underdetermination problem. Either that, or revel in rational latitude comparable to that enjoyed by their objectivist colleagues down the hall.

3. Subjectivism: why idealize?[5]

Idealization is central to two related but distinct classes of subjectivist theory, each with multiple variants. The first are full information accounts of well-being. The second are full-information accounts of reasons for action (Sobel, 2016, p. 115).[6] In each case, the basic rationale for idealization is that our uninformed desires can lead us astray (Heathwood, 2005; Hubin, 1996; Murphy, 1999). In the context of idealizing subjectivist accounts of reasons for action, Bernard Williams' celebrated example is that of my mistakenly believing a glass contains gin when, in fact, it contains petrol. I desire to drink it, but it seems wrong to say I therefore have a reason to drink it. If I knew it contained petrol, I would no longer desire to drink it. It is the desires of my better-informed self that determine my reasons for action (Williams, 1981). A similar rationale motivates idealization in full information accounts of well-being. Perhaps what I most desire right now is to be a famous athlete: the excitement, the money, the adoration! But, unbeknownst to me, I would in fact find the life of endless training required boring and shallow. Worse still, I would find the public scrutiny of my every moment and the pressure to perform on game-day incredibly anxiety-inducing. It is not what I think I desire, but what I would desire were I properly acquainted with it, that determines the good for me. In cases like the desire to be a famous athlete, I am mistaken about what it would actually be like to *be* a famous athlete so what I now desire is a chimera. I really do desire the life of an athlete *as I imagine it*, but not the life of an athlete as it really would be.

In both cases just discussed, idealization plays what might loosely be described as a 'negative' or 'defeater' role. I desire something because I am mistaken about its true nature. Were I properly acquainted with the true nature of the object of my desire, I would no longer

[5] See (Enoch, 2005).

[6] Cf. (Brandt, 1979); (Darwall, 1983); (Dorsey, 2017); (Gibbard, 1990); Lewis, D. (1989); and (Railton, 1986a).

desire it.[7] Here, the non-idealized agent's misapprehension is highlighted by the idealized agent's choice and defeated by it. That is, the ideal agent's preference acts as a *disabler*. In this guise, idealization does indeed look like it might lead to greater determinacy. Fanciful, poorly thought-out options are trimmed from my deliberative purview: I no longer have reasons to be a dinosaur or an 18th century pirate, astronaut, firefighter or wizard, though all of these have been aspirations of mine at some stage of my life. All that remains, let's suppose, are more "grown-up" desires for health, happiness, intellectual stimulation, enriching experiences, meaningful relationships, perhaps doing a bit of good and so on. Four observations are in order. First, at this level of desiderative generality and even allowing for my personal quirks, dispositions, tastes and the like, there are a bewildering number of utterly different lives I might lead that would contain satisfaction of these various desires in different ratios but with roughly the same degree of overall preference satisfaction. Second, how the details are to be filled in is precisely where one might have hoped deliberation on what to do might help, but it is at precisely this point that overall rough parity between total welfare across the candidate lives deprives one of principled means of distinguishing between them and flipping a coin seems utterly unsatisfactory for so momentous a decision. Third observation: at this point, idealization plays what I will label its 'positive' or option-enhancing role, and so at this point increases, rather than trims, indeterminacy about what to do. That is, what would otherwise be unreasonable options are enabled by the choice of an idealized agent that makes them reasonable. Here, the ideal agent plays the role of *enabler*. Suppose that I have never tried a kiwi-fruit, am unaware they exist and so presently have no desire to try kiwi-fruit, but that, were I to try one, I would love it and come to desire it as much as any fruit with which I am presently acquainted. Most proponents of idealizing subjectivism maintain, very plausibly, that I therefore have some reason to try kiwi-fruit and that trying kiwi-fruit would be good for me. It would be unduly impoverishing to limit our good to options of which we happen to be aware. But now the extent to which subjectivism provides a bulwark against indeterminacy is seriously in doubt. If I now have reason to perform each option that I would desire to perform were I aware of how the experience would affect me, and the aim is then to generate some unique combination of options that would determine what I should do with my life, the chances of success – even of generating a manageable range – are slim indeed. The fourth observation is that, insofar as the rational underdetermination problem is concerned, it now looks like

[7] Of course, this is perfectly consistent with my continuing to desire a glass of gin or the life-of-an-athlete-as-I-imagined-it.

subjectivists and objectivists are presented with comparable deliberative scenarios. Perhaps I presently desire to steep myself in a world of historicism, whimsy, and camaraderie, but am unaware that joining my local Morris dancing troupe is an optimal way of achieving this, though I am aware of the troupe's existence. Or perhaps I do not have any present desires for historicism, whimsy, and camaraderie or Morris dancing but, were I to take-up Morris dancing I would love it for the historicism, whimsy, and camaraderie it provides, filling a hole I did not know was there. Once we are contemplating what we *would* desire, were we acquainted with it, there is a significant degree of overlap between the options subjectivists and objectivists will find they have. After all, there is a significant degree of overlap between what objectivists and subjectivists find *valuable*.

4. Idealization and underdetermination

If the argument so far is on track, then *pace* Gert's suggestion, there is reason to think there may be little difference between the degree of indeterminacy facing subjectivists and objectivists on the whole. In fact, we would expect this to vary with the preference structures and desiderative profiles of different agents. In this section, I argue that subjectivists are actually confronted with *more* indeterminacy than objectivists. The basic idea is that subjectivist idealization is not subject to deontic constraints that rule out otherwise internally coherent but intuitively unappealing combinations of desires. This problem is then compounded by incommensurability between the perspectives generated by idealization.

To illustrate the increased indeterminacy facing idealizing subjectivists post-idealization, consider the case of an agent we'll call 'Greenie'. Greenie desires health, happiness, intellectual stimulation, enriching experiences, meaningful relationships and so on, like many agents (Railton, 1986b). Additionally, Greenie has a very strong desire to protect the environment. Given her desiderative profile, both objectivists and subjectivists will, I assume, agree that various lives could be equally good for Greenie: perhaps there is nothing to choose between Greenie's becoming a conservationist living in Rangat, a marine biologist living in Port-aux-Français, a climate scientist living in New York, and a green-tech investment banker living in Brussels. But we might expect objectivists to baulk at other possibilities equally a function of her desiderative profile, say, becoming a violent eco-terrorist or leader of a euphoric enviro-anti-natalist-death-cult. Presumably, this is because there are objective reasons not to risk or intend to damage others' property, kill people, encourage others to kill themselves in the name of environmentally-motivated population control and so on. This is a variation on the theme of what Sharon Street calls 'ideally coherent eccentrics' (Street, 2009). Since, we may

suppose, it is possible that Greenie would be happy as an eco-terrorist, would find it a meaningful existence full of fulfilling relationships with other eco-terrorists, enjoy the intellectual stimulation of planning eco-terrorist operations, would find carrying out these operations full of enriching experiences and be filled with an overwhelming sense of purpose and so on, it actually looks like the idealizing subjectivist has more options before her than the objectivist. Sure, we might think, Greenie is not right now contemplating blowing up coal-fired power-stations, but that is only because she has not realised that this too is well-supported by her reasons, given her desires. Blowing up coal-fired power stations is, one might argue, like kiwi-fruit or Morris-dancing. Don't knock it, 'til you've tried it (in the fully-informed imagination of your ideal self).

At this point, you may be thinking that no one self – even one with full information and idealized deliberative and imaginative capacities – could simultaneously entertain the prospect of being a mild-mannered, family orientated marine-biologist and a cut-throat eco-terrorist while holding onto a sense of how to compare them. I suspect there is some truth to this conjecture, and further reflection on it suggests a further potential indeterminacy problem for the idealizing subjectivist, by way of a dilemma. On the one hand, let's suppose that there is just one ideal self that can indeed contemplate and find equally attractive becoming one of many different selves as a result of pursuing an optimal, idealized option. In that case, if the actual deliberating agent should do what its ideal counterpart would advise, say, then the actual deliberating agent has equally weighted reasons to become any one of its ideal counterparts: biologist, banker, conservationist, terrorist, climate scientist, cultist and no way to choose between them. On the other hand, suppose that it is not possible to integrate these various possible idealizations into a single perspective. In this case, the actual deliberating agent has multiple ideal selves each of which has a different idea of what would be best.[8] Some of the moderate ideal selves may well be able to understand the appeal of other moderate ideal selves: the marine biologist and the conservationist will, we might think, probably get along. But other selves will find one another utterly mystifying and repulsive, in part precisely due to their shared starting point. The eco-terrorist will be disgusted by the self-serving comfort and meagre contribution of the conservationist, while the conservationist will, one assumes, be horrified by

[8] Cf. (Walden, 2013)'s reflections on open-ended constructivism with (Sampson, 2021), who takes something like this argument to present a direct challenge to reasons internalism. But given that rational underdetermination is a pervasive problem for externalists about reasons as well, it is difficult to see why this should be a *special* problem for internalists.

the violence and zealotry of the eco-terrorist. The actual deliberating agent, contemplating which course of action to pursue, is at this point confronted not just with equally weighted options and no principled decision procedure, but with incommensurable conceptions of the agent's own good, despite each conception resulting from idealization of her actual evaluative perspective. In short, the actual agent's desiderative profile either underdetermines the options available to its ideal self or underdetermines the agent's ideal selves. Either way, indeterminacy looms.

In fact, the situation facing the idealizing subjectivist may be trickier still. Throughout the discussion of Greenie, I have granted, *arguendo*, a tacit preservative assumption. This preservative assumption is that an idealization of an agent's actual evaluative perspective will be maximally preservative of her existing desires and values, integrating *all* of them, to the extent that it is possible to do so, into a unified evaluative perspective. So, if an agent desires and values health, happiness, intellectual stimulation, enriching experiences, meaningful relationships and protecting the environment, then each of her ideal selves will also desire and value these things and the lives they choose will reflect this. Different equally rational selves – and, therefore, options – will do better and worse vis-à-vis different desires, we might think, but what matters, let's suppose, is overall satisfaction. Perhaps the green-tech investor has much higher disposable income facilitating better health treatment and many more enriching experiences than the conservationist, but the conservationist enjoys much richer personal relationships and so on. But the conservative assumption looks like it requires independent justification. This is not the Millian point that it is better to be less than satisfied Socrates than a very satisfied pig, rather, it is not a *given* that the very same rationale that permits differential levels of satisfaction of each desire, provided overall satisfaction is equal, might not permit compensation for the loss of a desire at the ideal level (or its dissatisfaction at the non-ideal level), with a massive increase in the satisfaction of the others desires and values an agent has (Mill, 2009). Perhaps becoming a hugely successful green-tech investor will instead have deleterious consequences for the investor's health in the long run, but this is more than compensated for by corresponding gains in the richness of the experiences and success in protecting the environment thereby achieved. But it seems odd to think this not a rational option for Greenie, because Greenie also values her health. Given Greenie's other values, this seems a permissible trade-off. But at this point, the indeterminacy worry presses further still. After all, if candidate options need not satisfy each and every desire and value and agent has – and it is implausible to think they must, given how many desires each of us has, and the difficulty of holding them all together coherently, let alone realizing them in our actual lives – then there

are many, many more options on the table than we might first have thought, while holding onto the tacit assumption that we preserve as many existing desires or values as possible. For example, if Greenie desires to both protect the environment *and* alleviate the condition of the working-poor in her hometown *and* to care for her elderly parents, it is going to look much more difficult to integrate satisfaction of every desire into every rationally justifiable option or ideal self. For the idealizing subjectivist, Sartre's case of the torn dutiful-son-resistance-fighter generalises across many more quotidian cases.

To re-emphasise, I am not saying this is a problem for idealizing subjectivism – it may be that this is just the way the cookie crumbles, normatively speaking. But I am saying it is a mistake to think subjectivists are *prima facie* any "better" off when it comes to facing determinacy than objectivists, in fact, things may be significantly "worse". I employ scare-quotes here because, of course, if one thinks this really is how things are or, like Gert, thinks that delivering a highly permissive account of rationality is attractive, then subjectivists are actually *better* off in this regard than objectivists. So, if this is part of the motivation for being an objectivist, it might be worth switching sides.

5. Tie-breakers?

But perhaps I have been been uncharitable – or, given the final point of the previous section, *overly* charitable – in my assessment of what idealization achieves. Specifically, the subjectivist could argue that for an idealization of a given agent's actual evaluative perspective to count as *genuinely* ideal it must meet certain further conditions and that meeting those conditions will very likely exclude all possible idealizations bar one.

In the previous section I prefaced an objection to this strategy. I argued that breaking a deadlock between (equally) ideal selves – or claiming that one was *more* ideal than all others – would require appealing to some external, objective (deontic or otherwise) value and that this was *prima facie* inconsistent with subjectivist theories of well-being and reasons. There are, however, two ways an idealizing subjectivist might try to rebut this *prima facie* inconsistency. The first is to deny that some values – theoretical virtues like elegance, conservatism and simplicity – are external and objective in the relevant sense. These values, the idealizing subjectivist might argue, are certainly external and objective in one sense. But in another sense, they are not external to but rather constitutive of the very normative practices idealizing subjectivist theories are trying to explain. Appeals to such values are necessary in order to even understand what it means for one evaluative standpoint to be ideal or for what counts as idealizing an actual evaluative perspective instead of, for example, arbitrarily substituting some

internally coherent but alien and un-related evaluative perspective for a person's actual evaluative perspective. These values are therefore not external and objective in the sense to which the idealizing subjectivist objects.

But although I think this account of these purportedly external and objective theoretical values might allow idealizing subjectivists to avoid the challenge that their justification for appealing to them is objectionably *ad hoc*, I think this comes at the cost of the relevant values' potential utility in breaking the deadlock between different possible ideal selves. The reason characterising theoretical values in this way will undermine their use in prioritising one ideal self over another is as follows. If the values in question are constitutive of the normative practices of which the idealizing subjectivist is trying to provide an account and so are, in turn, the very norms by which idealization is to be understood, then they cannot be used to decide which ideal self should take priority over the others. By definition, each ideal self meets the norms constituted by these values equally, otherwise the self in question would not be ideal. For the idealizing subjectivist's characterisation of these objectivist-looking values as constitutive of the enterprise to work, it must be the case that, for example, ideal coherence is much more than just mere consistency. Ideal coherence will then be constituted not just by perfect consistency plus full information but also by, for example, maximal elegance.

The idealizing subjectivist might then have hoped to use these values to measure degrees of idealness but, to the extent that the values can be put to that use, they simply rule out counting some purported idealizations as genuinely ideal in the most robust sense. But this does not provide a way of prioritising one ideally coherent self over other equally coherent and ideal selves, which is what they require to avoid the underdetermination problem. Rather, this is just an explanation of the way in which we might rank different possible selves against a metric of how close to ideal they really are, namely, by considering the extent to which they meet the relevant theoretical virtues. In order for this to help the idealizing subjectivist, they must provide a further argument to the conclusion that in most or all cases, there will be one self that is "more ideal" than all the others. The mere fact that idealization is a rich and complicated process is not sufficient to render this claim plausible. If I am right about this, then an appeal to theoretical virtues will not break the deadlock between selves, even if the idealizing subjectivist can consistently appeal to such values in this context.

6. Conservatism

But there is a second route the idealizing subjectivist might take to rebut the *prima facie* inconsistency between subjectivism and appeal to such apparently external values in order to

break deadlocks between ideal selves. The second route is to argue that the value of *conservatism* is primary and is capable of generating a unique ideal self. This is analogous to its use as a tie-breaker in cases of underdetermination of scientific theory. Some argue that where a choice must be made between two equally coherent theories, we should, or at least do, prefer the one that requires the least alteration of our existing belief set.[9] So, we might claim that one ideally coherent self has greater normative authority over us than any other because it preserves the greatest number of the values, preferences and dispositions that were constitutive of our starting actual evaluative standpoint. In other words, the idealizing subjectivist might claim, we should do what our ideal self would want us to do, where our ideal self is the self that is both ideally coherent and most similar to our actual self.

Before assessing the extent to which this appealing suggestion can be squared with the idealizing subjectivist's other commitments, I note that, like the other theoretical virtues, conservatism is open to the challenge identified above. That is, either conservatism is an external value of the sort to which the idealizing subjectivist cannot legitimately appeal or it is constitutive of the relevant normative practice: one of the values by reference to which idealization should be understood. If the latter, the idealizing subjectivist must argue for the claim that this value will (a) generate a unique ideal self and (b) that this self always trumps other potentially ideal selves that are generated by placing less weight on the value of conservatism and more on, say, elegance or creativity. In any event, there are reasons to doubt that even were conservatism granted primacy as an organising value, it would generate a unique result. For example, how are we to understand which starting preferences, dispositions and values count for the purposes of this measurement? By number, number of entailments, strength, duration or some other metric? As a matter of fact, it seems likely that for many of us there will be multiple equally conservative ways of idealizing our actual evaluative standpoint. If that is true, then conservatism will not do the trick.

I doubt that this problem can be overcome. But even if it could, I believe the conservatism proposal is mistaken, albeit tempting. It is tempting for at least two reasons. The first reason is that giving the value of conservatism priority generates an intuition that there will be more values left at the end of the idealization process and that the ideal self will be correspondingly more substantial. If what I have been saying about idealization is correct, then ideal coherence can come cheaply by simply deleting and adding values from our actual

[9] See (Quine, 1980) and (Quine, 1976). Quine argues that our conservative disposition is a matter of fundamental psychology, though others argue that conservatism is also a theoretical virtue: Cf. (Sklar, 1975).

evaluative standpoint. A concern one might therefore have is that many of our ideal selves will be extremely sparse: single-minded and devoted to a cause. Sure, they are ideally coherent in one sense, but there is not much to them. One might think, then, that the ideal self that conserves the greatest number of our existing values will be a much richer and more complex self and is, for that reason, preferable.

But there are three problems with this thought. The first problem is that it is not clear that conservatism will generate richer or more substantial selves. Idealization means that there are as many ideal selves as there are consistent starting sets of desires and values in a person's actual evaluative perspective (where a set can, of course, be a singleton). But once that starting point is adopted, the ideal self is constructed from not just the starting value(s) but also those values that will follow from and cohere with them. As Smith observes when building on Williams' account of idealization:

> Williams is right, I think, that deliberation can both produce new and destroy old underived desires. But he is wrong that the only, or even the most important, way in which this happens is via the exercise of imagination. For by far the most important way in which we create new and destroy old underived desires when we deliberate is by trying to find out whether our desires are *systematically justifiable*. (Smith, 1994, pp. 158-159)[10]

It is for this reason that idealizing subjectivists can argue that a person's ideal self would advise them to perform some action that is only related to his or her actual evaluative perspective tangentially, in the sense that the relevant desires are consequences of having the other desires they do. This means that a great many of our ideal selves will be just as rich and complicated as the ideal self that follows from idealization plus conservativism.

The second problem is that it now looks like we are trying to justify an appeal to conservatism in terms of another value, one that looks decidedly objective and external to the self in question. That value we might call 'complexity' or 'substantiality' or 'richness'. We now need some further explanation of, for example, how we should weigh the single-minded pursuit of some ideal against a complicated life full of compromise, together with an explanation of why we should prioritise the latter. It is difficult to see how this explanation can proceed without appealing to further external values in deciding between the different ideal selves.

---

[10] Of course, given the fact of underdetermination, I deny that trying to find out whether our desires are systematically justifiable *entails* conservative, Rawlsian reflective equilibrium.

The final problem is that conservatism is *not* just a proxy for richness. By placing conservatism at the centre of our account of idealization, of course, we prioritise our actual self at any given moment, but what reason do we have to do that?

The obvious answer to this question is that our actual self is, as a matter of fact, important to us. This is the second reason that the idea that our ideal self is the ideally coherent self closest to our real self is tempting. We care about our actual selves in a special kind of way. Our concerns, values, preferences and dispositions at this point in time are important to us because they are *ours*. This is one of the reasons subjectivism is attractive in the first place. Objectivism ignores our special concerns and plans and gives priority to alien rules and requirements in their stead. One reason, then, to give the ideal self that is closest to our actual self priority is that it is the *least* alien, "external" ideal self. Correspondingly, its authority seems both more legitimate and less demanding than all our other ideal selves. Not only that, but we also might have a better chance of grasping what our ideal selves want us to do if they are not too remote from our actual evaluative standpoint. If this were true, then perhaps this might justify giving the closest ideal self priority and, if we can give one ideal self priority, then underdetermination is no longer a problem.

Unfortunately, I do not think this will work, for three reasons. The first reason comprises the familiar set of problems associated with indexing what you should do to your present desires and beliefs. The rationale for idealizing in the first place is that we are often mistaken about what we think we desire, know and value. For some people (either very lucky or very wise or both) it may be that some or all of their ideal selves are very close to their actual self. But for most of us, it seems very likely that most of our ideal selves will look very different from who we are now. Closeness to our actual self might be intuitively attractive in some cases and unattractive in others – how far each idealization takes us from our current evaluative standpoint will depend on the details of each person's psychology and place in the world – but there does not seem to be a non-circular way of justifying the priority of our actual self, at least not without appeal to some external value. Privileging those selves that are closest to the starting point would only make sense if the starting point had some normative significance over and above its being a set of psychological facts about us, but, by hypothesis, it does not.

But even supposing I was mistaken about that, giving priority to an ideal self because it is closer to an actual self appears to involve the following error: it supposes that some of your ideal selves are more you than others. But that is a mistake. Each of our ideal selves is us, but better, in the sense of having full information and reasoning infallibly. It is no objection to idealizing subjectivism generally that some people's ideal selves look very different from their

actual self. On the contrary, that fact demonstrates the power of idealization in making sense of our normative practices. Some people lack more information, make more mistakes and start with more incoherent sets of values, preferences and dispositions than others. It is therefore a mistake to think that, given underdetermination, one ideal self enjoys lexical priority over others as a result of sharing surface similarities with an actual self at a given moment in time. Each ideal self has an equal claim on us in virtue of being a fully-informed and error-free version of ourselves. That is why underdetermination is a big problem for the subjectivist too.

If the argument so far is correct, then no ideal self enjoys priority over any other just in virtue of its proximity to our actual self. While I believe this result exhibits deep consonance with the subjectivist project, the subjectivist's underdetermination problem remains. In the final sections of the paper, I sketch two potential solutions to it. The first is democratic and the second is modelled on trusteeship. I argue that potential technical problems with the democratic solution are soluble. But I also argue that, once solved, the solution looks difficult to square with a key reason for becoming a subjectivist in the first place, namely, the sense that there should be some close connection between an agent's actual desires and their reasons, such that the latter are capable of *motivating* actual agents in the right circumstances. This is not intended as a decisive objection to the democratic solution. But it does, in my view, give us a good reason to prefer a solution that looks more amenable to this characteristic subjectivist commitment. The trusteeship solution, on the other hand, satisfies this motivational desideratum quite well. In my view, this gives us a reason to prefer the trusteeship solution.

7. The democratic solution

I will label the first potential solution to the subjectivist's underdetermination problem 'the democratic solution', the bare bones of which are as follows. Imagine that each of our ideal selves gets one vote regarding what we should do on any given occasion. While they all disagree about what to do initially, let's suppose that through a yet-to-be-specified procedure of judgment aggregation they can arrive at a decision about what we should do. The democratic solution to the underdetermination problem maintains that we should do whatever the outcome of that hypothetical procedure says we should do.[11] That is, we should do what our ideal *selves* want us to do, where they are deliberating together as a *demos*.

[11] Thanks to an anonymous referee for suggesting this possibility. I note that (Southwood, 2019)'s discussion of 'contractualism for us as we are' might be re-purposed to help with some of the problems outlined here.

There is much to like about the democratic solution. But, of course, the devil will be in the details. Judgment aggregation is not a simple matter, as Condorcet's paradox shows (Condorcet, 1785).[12] But, as a first pass, we should at least *start* simple. The simplest voting procedure applicable to our scenario is Plurality Rule, also known as 'First Past the Post'. In any deliberative scenario, each idea self would vote for the candidate action they believe the actual self should perform. The action the actual self ought to perform is the action that receives the most votes. So, if you are deciding between granola and blueberry pancakes for breakfast and more selves want granola for breakfast than blueberry pancakes, granola is what you should have. This doesn't mean you should always have granola, of course. It may be that the demos feels differently about the relative merits of granola and blueberry pancakes on weekends, for example. It just means granola wins on this occasion. But, of course, there are also some well-known problems with Plurality Rule, including that it sometimes picks a Condorcet loser as the overall winner. Some voting methods seek to address this problem by introducing quotas, for example, requiring a Strict or Absolute Majority winner. But addressing Plurality Rule problems with quotas does not guarantee a winner (and of course may make it less likely in some scenarios). So, it looks like there will still be some decisions where the underdetermination problem remains. Of course, the real problem here is that there are more breakfast foods available than just pancakes and granola and, if we raise the stakes to deciding on *careers* or the like, the possibility of deadlocks looks very real. In short, the existence of multiple ideal selves is predicated on their having conflicting views about what is best for the actual self, so *by hypothesis*, in many important deliberative scenarios, Plurality Rule will produce no result. And this is precisely what we should expect, if the argument of the first part of this paper is correct.

But there is another reason to reject Plurality Rule, which is that it is too coarse-grained. We will get a much more nuanced picture of the general will of the ideal selves if they are required to *rank* the candidate actions for the actual self to perform. Of course, in non-ideal voting scenarios, a significant drawback to ranking systems of voting is that they are very demanding: they require more time, effort and information than other means of decision in order to accurately reflect the preferences of the voters. But, since the ideal selves are ideal and, we can assume, have all the information required to make their rankings, know their own preferences immediately, and can compute all of this instantaneously, we need not worry about these drawbacks here. The best-known ranking system of voting is the Borda Count, in which,

[12] For a helpful overview of the judgment aggregation literature see: (List & Puppe, 2009).

if there are *n* candidates, the voter assigns *n – 1* points to their preferred candidate, *n – 2*, to the second, and so on (Borda, 1784). So, applying this to our scenario, the Borda Score of a candidate action is the total points assigned by all ideal selves to that action and the action the actual self should perform is the candidate action with the most points. Will selecting this voting procedure – or another like it – ensure that there is always a unique answer to the question what the actual self should do? Let me start with an optimistic answer with an interesting result, before moving to pessimism as a reason to at least consider an alternative solution.

The optimist answers as follows. Yes, a procedure like Borda count; or a multi-stage voting method that eliminates least preferred options; or a voting procedure that allows ideal selves to provide more information about the strength of their preferences by allocating points in a different way between candidates; or some other appropriate procedure, will *always* provide a determinate answer.[13] We just have to find the right procedure. Supposing there is one to be found, can we say anything more about the sort of answers it might give to questions concerning what we ought to do? One hunch is that we might expect some tendency towards the mean. That is, the action the ideal selves will want the actual self to perform will be, on balance, the action one of the "middling" or "average" or "less radical" selves would have wanted. So, in the case of Greenie discussed earlier, we might expect the terrorist and cult-leader options to be ruled out, while the conservationist, marine biologist, climate scientist and green-tech investment banker look much more likely. In fact, *if* these professions could be ranked along a *single* spectrum and *if* a voting method were selected that satisfied the Condorcet Criterion (i.e., a Condorcet winner is always chosen if one exists), then Median Voter Theorem tells us that that the successful candidate will be that closest to the median voter (Black, 1948). But, of course, these are big 'ifs'. For example, it is not obvious that Greenie's candidate professions could be ranked along a spectrum of something like 'more to less extreme', even if we have the sense that terrorist and cult-leader are more extreme than the others. And even if they could, why is 'extremity' the relevant spectrum along which the candidate careers should be arranged? But perhaps we do not need the technical assurance of the Median Voter Theorem. Instead, there seems an intuitive sense of preferential "cross-

[13] These options are not intended as exhaustive, of course. There are many possible voting methods and the literature on their relative merits is vast, technical and well beyond the scope of this paper. Perhaps, if the democratic solution is independently plausible, it can wait for the best solution to these problems to be provided in that literature.

cancelling" that suggests that the more radical options will get less support overall than those less radical, and so, the optimist says, there is some intuitive support for the idea that the demos of ideal selves will select the sensible centre candidate action from among the various possibilities. If that is true – and *if* it is possible to identify the sensible centre candidate – then not only does the democratic solution solve the subjectivist's underdetermination problem, it does so in a way that might give us, for any given agent, some sense of which action they ought to perform, provided we have an antecedent sense of who their possible ideal selves are.

The pessimist denies that any judgment aggregation procedure will always deliver a single winner. Their pessimism may stem from the vast literature on Arrow's Theorem (Arrow, 1963) *or* it may follow more straightforwardly from a sense that that the different options a person has about what to do with their life are not the sort of thing that can be ranked: there are simply too many incommensurabilties and incomparabilties and ties in play to hold out any hope of always getting a determinate answer from a vote. For the pessimist, the problem with the democratic procedure is that it looks like it might give determinate answers to questions that don't really matter – like what to have for breakfast – and then leave us high and dry when making the decisions that really count. So, the pessimist worries, the democratic solution is at best a partial solution to the subjectivist's underdetermination problem. As it turns out, the questions that really matter do not have positive answers, though some courses of action will be ruled out.

But while I think the pessimist is right to worry that a straightforward judgment aggregation procedure is unlikely to generate determinate answers all the time, I also suspect that the rich resources of social choice theory will ultimately be up to the task. That is, I am fairly confident that some procedure could be found that solves all the technical issues with the democratic solution to the subjectivist's underdetermination problem. But even supposing that it can – and in part precisely *because* resolving technical issues in social choice theory is complex – I think there is a much more pressing, principled objection to the democratic solution.

The principled objection is that the degree of complexity and abstraction the democratic solution involves cuts actual, fallible, cognitively limited agents off from their reasons in two ways: *motivationally* and *practically*. Part of the initial attraction of subjectivism was supposed to be preserving a close connection between our actual desires, motivations, values and concerns and our *reasons*, such that those reasons were somewhat *tractable* for actual agents and, when apprehended, actually *motivated* us to perform the action they support.

Now, of course, the metaphysics of reasons need not straightforwardly deliver a decision-procedure, but one might hope it doesn't make deciding correctly look beyond the reach of

most actual people either. But if the correct answer to what one should do at any given moment requires instantaneous, complicated social-choice-theoretic computations concerning preference aggregation across large numbers of ideal selves, then most people are – *as a practical matter of epistemic access* – cut-off from their normative reasons most of the time. That is, most of us will never actually know what we're supposed to do at all, let alone in time to make a decision on the spot.

Worse still – and even supposing I were mistaken concerning practical epistemic access – the democratic solution risks severing a crucial connection between reasons and motivation. That is, supposing one could ascertain the decision handed down by one's demos of ideal selves, there doesn't seem to be anything *irrational* about the agent who is then utterly *unmotivated* to act on the reason they're given. In part, this will be due to the bewildering technicality of the metaphysics of their reasons. In part, it will be because the agent will know that while one action may ultimately have won the day, the reasons each ideal self may have for preferring that action may be completely different from and perhaps even straightforwardly inconsistent with the reasons other ideal selves have for preferring that action. This seems to go some way to stripping their verdict of its authoritative motivational appeal, such that it no longer seems bizarre or irrational to fail to be motivated by it, given one's other motivations. But preserving this sort of close motivational connection was one of the main reasons to become a subjectivist in the first place. So, if this worry is on the right track, the problem with the democratic solution to the subjectivist's underdetermination problem is that it deprives subjectivism of much of its initial appeal *and* is unlikely to be found congenial by many subjectivists. This is not a decisive objection to the democratic solution. It may be that this is just the way things are, normatively speaking. But I think it is troubling enough to hunt for another solution that scores better on these practical and motivational metrics. And I think the trusteeship solution does just that.

Let me briefly recap. The democratic solution to the subjectivist's underdetermination problem faces technical problems concerning the possibility of ties. That is because if there are often ties, then underdetermination and indeterminacy permeate the ideal level of decision-making too. But it is likely these technical problems can be solved. There is, however, a more pressing objection, namely, that the democratic solution cuts actual agents off from their reasons as a practical matter of epistemic access and by severing the close connection between reasons and motivation that is at the heart of subjectivism. This is not a decisive objection to the democratic solution, but it does give a subjectivist who wants to preserve the close

connection between reasons, motivation, and actual decision-making a reason to look for an alternative solution.

8. Choosing yourself

Let me therefore suggest an alternative solution to the underdetermination problem. I will label it 'the trusteeship solution'. I will begin by explaining it in the abstract as a solution to the underdetermination problem confronting the idealizing subjectivist and then argue that, so understood, this account vindicates a tidied-up version of how some reflective people already do deliberate in their everyday lives. This, in turn, suggests that a decision-procedure closely connected to the account is both possible (because actual) and attractive.

The basic idea, in the abstract, is as follows. *We should do what the ideal self we would choose to be, were we aware of their existence, would want us to do.* Let me explain. Suppose for a moment – *per impossible* – that each actual deliberating agent were aware of every one of their possible ideal selves: a virtual smorgasbord of candidates for who to become as the best version of themselves. Now, since no ideal self enjoys *automatic* priority over any other possible ideal self, the actual self is not subject to the automatic rational authority of any one of its ideal selves either. Each is a candidate for the allegiance of the actual deliberating agent, who can decide which ideal version of themselves to entrust with rational authority over their life. Each ideal self represents an ideally coherent, fully-informed version of them, with a clear plan for what they should do, integrated into an evaluative standpoint that determines what is valuable in life and how to live well.

Why is this second solution to the underdetermination problem modelled on *trusteeship*? In brief, the idea is that the relationship between actual and ideal agent is akin to a form of non-delegatory representative government.[14] It is not a form of *delegation* because the ideal self does not simply follow the present preferences of the actual self. They are not a mere mouthpiece for the actual self because, by hypothesis, many of the preferences of the actual self are inconsistent with the ideal self, otherwise idealization would be otiose. Instead, the trustee ideal self is bound by fiduciary duty to govern on the basis of its best understanding of the interests of its beneficiary, the actual self. Like a trustee, the ideal self has unfettered discretion to select both the ends to be pursued and means appropriate to pursuing them.

These brief remarks also suggest an answer to the following objection to the trusteeship solution. If it is up to us to decide which ideal self determines our reasons, is it even *possible*

[14] On the distinction between delegation and trusteeship in political representation see (Dovi, 2018).

for us to make a mistake by switching between ideal selves post-decision? After all, if we commit to one ideal self, but then behave perfectly in accordance with the demands of another, we're still acting in perfect accord with the recommendations of an ideal self, so how could we possibly be making a rational mistake? But if we are not making a rational mistake, then doesn't this show that the trusteeship solution to the underdetermination problem fails?[15]

Responding to this objection requires thinking carefully about the diachronic relationship between synchronic ideal selves. Suppose, simplifying grossly, that at t_1 you have a choice between two ideal selves, A_1 and B_1, and that you choose A_1. Shortly thereafter, at t_2, you begin behaving in accordance with the recommendations of B_1. In what sense have you made a rational mistake, since B_1 was among your ideal selves at t_1? The answer is that B_1 is no longer among your ideal selves, because it is no longer t_1. There *may* be an ideal self at t_2 that resembles B_1, and for argument's sake, we will call this ideal self "B_2". What are the salient differences between B_1 and B_2? Well, unlike B_1, B_2 is an idealization of your actual evaluative standpoint at t_2, which includes your prior commitment to A_1, any investments undertaken in actualizing the recommendations of A_1 and so on. Now, *if* – even allowing for that commitment, and the value you place on commitments (if you do), B_2 is an ideal self – *then* no, you have not made a mistake in behaving in accordance with B_2's recommendations, because that ideal self has "built-in" the breaking of rational commitment, the switching costs and so on. But it *would* be irrational to behave in accordance with the recommendations of B_1, because at t_2, B_1 is no longer among your ideal selves.

But perhaps the *really* troubling feature of the trusteeship solution is not the unfettered discretion of the trustee ideal self, but the *unfettered discretion of the actual self* to select their ideal self. It is to this worry I now turn. Recall that in the hypothetical decision scenario we have complete freedom to choose the ideal self we entrust with rational authority over ourselves on the basis of the intrinsic appeal of that ideal self, viewed from the actual evaluative perspective of the fallible, cognitively limited and less than fully informed and coherent agents

[15] Thanks to an anonymous referee for pressing this objection. Of course, depending on one's view of the role of intentions in practical reasoning, it may be that it is straightforwardly irrational, because means-end incoherent, to intend to φ but then fail to take the steps necessary to φing. If this is true then the actual self that intends to commit one way but then behaves inconsistently with this commitment has made a rational mistake. Whether this is also therefore a normative mistake is a point of significant contention. Cf., for example, (Bratman, 1987); (Broome, 2001); and (Kolodny, 2011).

we are.[16] This is because each candidate is "equally" ideal: our rational freedom is only constrained by the structural rationality that renders each potential self ideal, which, by hypothesis, it already is. But shouldn't we ask why the actual self should have any say in the matter *at all*. Doesn't granting the final say to the actual self undermine the solution altogether? If the motivation for idealization in the first place was that the actual self is poorly informed, irrational, weak-willed, and motivationally incoherent, why on earth should we grant this unimpressive specimen ultimate normative authority?[17] But although this objection touches on something deep and troubling about how best to justify assigning any value whatsoever to the desires and preferences rational agents just happen to have, I do not think, on closer inspection, it undermines the trusteeship solution to the *underdetermination problem*. This is because, I suggest, on the best understanding of the dialectic in this paper, the problem is: the underdetermination problem, given idealization; and *not*: why idealize, given the underdetermination problem.

The rationale for idealization in the first place is the irrationality and imperfect information of actual deliberating agents. Idealization is not generally directed towards solving the underdetermination problem, rather, it is directed towards solving the problem that, on the one hand, the subjectivist believes reasons and an agent's welfare are grounded in facts about what they (would) desire, while on the other hand acknowledging that uniformed and irrational desires are a poor basis for determining an agent's reasons and welfare. Idealization is supposed to cure the deficiencies in our subjective motivational set to avoid gin-petrol cases and the like. An assumption shared by many objectivists and subjectivists is that idealization will generate a unique result – a single ideal self – and that underdetermination is therefore not a pressing problem for the idealizing subjectivist. If the argument of this paper is correct, that assumption is false.

But I am not suggesting this is a reason not to idealize, rather, I am suggesting that further machinery is required for those troubled by the recurrence of the underdetermination problem at the level of ideal *selves*. But once we are operating at the level of equally ideal selves, I see no reason why the preferences of an actual agent *surveying those ideal selves*

[16] Cf. (Chang, 2009; 2013), who argues that we create voluntarist reasons where our given reasons run out. My view is similar to Chang's in some respects, though, of course, I am not arguing that there are given reasons explained by facts about objective values, nor am I arguing that we create voluntarist reasons by choosing to take *individual* considerations as reasons, nor am I arguing that there are deontic constraints on what an agent can recognize as a reason. Cf. (Gert, 2018)'s criticisms of Chang.

[17] Thanks to an anonymous referee for pressing this objection.

cannot be relevant to deciding who to be and which life to lead. By hypothesis, every option is equally rationally permissible – the field has already been radically trimmed by the requirement that each option be integrated into an ideally coherent evaluative perspective – so it seems right to me that, at this point, it is just "up to us". Our options have been tidied-up in light of full information and the requirements of rationality.

So understood, the trusteeship solution to the underdetermination problem offers a degree of rational freedom that is not overwhelming, but empowering. This, in turn, gets the balance right between rationality and alienation. An animating concern of subjectivism is to ensure that we are not alienated from our own good, while also not allowing our decisions to be dominated by false beliefs, incomplete information or irrationality. But a persistent concern with idealizing accounts of subjectivism is that our ideal self (or, if I am right, *selves*) would have to be so unlike us in order to weigh the various options that they are no longer really *us*. What they command "from on high" now looks too remote from our present concerns to be consistent with a key motivation for being a subjectivist in the first place. But the trusteeship solution answers this concern by placing the authority to decide what is best squarely on the shoulders of the non-ideal agent. We decide whom to *entrust* with our authority thereby preserving our freedom. We ensure that the authority is wise, because ideal and we ensure that it is *trustworthy* because, being us, it has our best interests at heart. Now, of course, there remains the possibility that the actual deliberating agent will find *all* the options prescribed by their ideal selves alienating. In this scenario, it may be of little comfort that the decision ultimately rests on their shoulders. But even in this scenario, it still seems a point in favour of the trusteeship solution that some choice between alternatives is left to the agent. Of course, it is better to have a choice between alternatives you find agreeable, but if the only alternatives available are disagreeable, it still seems better to have a choice among them than none at all, precisely because, if nothing else, the choice gives you ownership over your life in a way that respects your autonomy as a rational agent. Besides, one might add, in most cases selecting between ideal options it is not unreasonable to hope that your participation in selecting it, together with the knowledge that the option is genuinely ideal, might be enough to bring you around to it.

It is, of course, at least possible that there will be some agents for whom *every single one* of their ideal selves appears utterly remote *and* unappealing, and therefore alienating. For these agents, it may be small consolation that they get to choose between many remote and unappealing options. But such agents are likely to be very rare. Each of us will have many, many ideal selves and, since each ideal self is a function of the desires, preferences and values

of the actual self, it would be surprising (to the say least) if *none* of them were at all *appealing* to us. Granted, they might seem *remote*, given how far short of ideal rationality we actual agents generally are. But for most agents, I suspect, this will *add* to the appeal of at least some of them, as they recognize their ideal counterparts for what they are: a better version of themselves. This, in combination with the normative and psychological significance of *choosing between them*, greatly reduces the alienating aspect of their residual remoteness.

Now, of course, a further objection might be that the trusteeship solution therefore defaults to conservatism. One might worry that *inevitably* the actual self will be biased to their "most proximate" ideal self. But I do not think this is true. Of course, some people will choose the self that is most familiar and does least violence to their existing plans and self-conception. Perhaps Greenie will be most tempted by the life of a climate scientist in New York after all. But this need not be the case. One can imagine Greenie surveying her candidate selves, contemplating the eco-terrorist and thinking, with a rush, 'Wow, I never knew I had it in me. This is what I have been looking for all along.'

This point also helps us to see the significance of the following fact: it is precisely because the actual self would respond in a determinate way in the hypothetical choice scenario that we can resolve the subjectivist's normative underdetermination problem generally, i.e. determine the actual agent's normative reasons. If indeterminacy permeated the hypothetical choice as well, then the underdetermination problem would remain unresolved. According to the trusteeship solution, in the hypothetical choice scenario, the actual self appraises the various ideal selves and decides which to entrust with rational authority. This is not the same as surveying the possible actions prescribed by the ideal selves and choosing on the basis of the intrinsic appeal of each action alone, judged by the actual self's own lights. This is not to say that the actual self *could not* decide whom to entrust on this basis. If the desire to perform some action were sufficiently strong, that might rule out every ideal self as a candidate, bar one, namely, the one prescribing that action. But the point is that further information is available to the actual self in the decision scenario in question, namely, how that action fits into an overall plan for their life endorsed by a particular ideal self. So, that an actual agent would choose some action under non-ideal conditions does not entail they would choose it under conditions contemplating their ideal selves, even if it were available. This speaks to the normative significance of their choice under these counterfactual conditions, which is part of the rationale for idealizing in the first place. The ideal self mediates and organizes the considerations which may be marshalled in favour of pursuing some course of action by integrating them into a holistic evaluative standpoint.

Is the trusteeship solution to the subjectivist's underdetermination problem hopelessly abstract? I hope not. In fact, I believe the account vindicates a tidied-up version of how some reflective people already do deliberate in their everyday lives. This, in turn, suggests that a decision-procedure closely connected to the account is both possible (because actual) and attractive. Of course, actual people do not go around trying to get imaginatively acquainted with their ideal selves and choosing between them. Subjectivists never intended idealization to supply a decision-procedure: that would be to over-intellectualize the relationship between actual and ideal selves. But nor is something like this procedure completely alien to our actual normative practices. Anytime someone selects a role model; or turns down a high-powered job because they realise they do not want to become the person they would have to be to succeed in it; or picks up a self-help book encouraging them to 'be the best version of themselves'; or advises a friend against a course of action because 'you don't want to go down that road, you might not like who you'll become', they implicitly engage in the sort of reasoning I have suggested is key to solving the underdetermination problem. This is literally *self-conscious* deliberation, where one considers the impact of some course of action on who one will become as well as on its other merits. Naturally, much more would need to be said to actually develop a decision procedure that best approximated the trusteeship solution while correcting for our various biases and other cognitive limitations, but I believe that our present normative practices indicate that the prospect of such a procedure is both possible and appealing. Careful deliberation about what to do often requires attempting to integrate our reasons about what to do *now* into a coherent picture of how we want to live our life *generally* and over the long run. A very important heuristic we already employ to do this is to ask, 'What sort of person do I want to be?' and one way of asking and answering that question is by imagining different ways our lives might go and the different people we would need to be to make the best of them.

9. Conclusion

In the first half of the paper, I argued that underdetermination is just as big a problem for subjectivists as it is for objectivists. In the second half of the paper, I presented two solutions to the subjectivist's underdetermination problem. The first was democratic, which confronted both technical and principled objections. I argued that the technical objections could be solved, but that the principled objections remained; and, if anything, were exacerbated by the solutions to the technical objections. I then presented an alternative solution I labelled 'the trusteeship solution'. I argued this solution avoided the in-principle objections to the democratic solution and was independently attractive, given the subjectivist's other commitments. I concluded by

observing that the trusteeship solution vindicates a tidied-up version of how some reflective people already do deliberate in their everyday lives. This, in turn, suggests that a decision-procedure closely connected to the solution is both possible (because actual) and attractive.

CHAPTER 2: THE SELF-IMPROVEMENT MACHINE

Abstract

Suppose that in the not-too-distant future, scientists invented a self-improvement machine. The machine adds a chip into your brain which contains lots of new information and helps you conform more readily with the norms of epistemic and practical rationality. As a result of the procedure, many of your priorities, interests, values and tastes are quite different from what they were before entering the machine. Despite the obvious benefits, I am reluctant to step into the self-improvement machine. That is, I am reluctant to risk losing my present attachments, values, preferences and desires even where I *know* that losing them in the machine meant they were *irrational*. This sounds odd and yet, I suspect I am not alone in feeling a degree of disquiet at the prospect of undergoing such a procedure. My hunch is that exploring the source of this disquiet can help us better understand the relationships between agency, identity and self-improvement of the more "organic" variety.

Introduction

Self-improvement is usually thought of as an unqualifiedly good thing. Its only obvious and oft-noted drawbacks are the time, effort, and willpower required to achieve it. Enter the pundits of the vast self-improvement literature. Popular psychologists and other social scientists jostle with self-styled life-coaches, self-help experts, and motivational speakers to offer advice on cultivating the habits, motivation, willpower, and smarts to become the best version of yourself and "live your best life".[18] Many philosophers share these concerns. After all, becoming a better person and addressing the deficits of motivation, willpower, time, and expertise required are integral parts of ethics and moral psychology. I am skeptical that the primary – let alone, *only* – reason not to self-improve is that it's hard work. On the contrary, I will argue that there are more troubling reasons not to self-improve, especially absent this alleged drawback. To test that proposition – and interrogate its significance for our understanding of the relationship between agency, identity and self-improvement – I introduce a thought-experiment that takes all the hard work out of self-improvement.

Before doing so, let me say something about the conception of self-improvement I aim to criticise in this paper. My target is the so-called 'rationality movement', whose adherents champion a focus on rational, bias-free, probabilistic, optimising, preference-satisfying, clear thinking as the core of improving yourself and living a happier, healthier, more efficient, more

[18] '#hashtag' optional.

productive life. In philosophy, its most vocal champions are members of the effective altruist movement, with which the rationalist movement is deeply entwined. For example, a recent *New York Magazine* interview with Julia Galef contains the following:

> "I often publicly identify as an 'effective altruist,'" she says, referring to the rationalist spinoff movement focused on optimizing philanthropy, "because it's easier to explain, and it doesn't rub people the wrong way the way 'rationalist' does. It's simpler and friendlier"[19] (Wallace, 2021)

The movement has many prophets, acolytes and devotees in silicon valley, start-ups, popular social psychology and sociology, economics and beyond. The core commitment of this worldview is that you can become a better person – both morally and "generally" – by devoting your energies to becoming more rational. What self-improvement is *really about*, the committed rationalist believes – for new-age rationalists is what they are – is empowering oneself to become and live better by becoming more rational. Here is a summary from the *New York Magazine* article quoted above, which is entitled 'The Tech Elite's Favorite Pop Intellectual Julia Galef on bringing the rationalist movement to the mainstream':

> In 2012, Julia Galef, the host of a podcast called *Rationally Speaking*, moved from New York to Berkeley to help found a nonprofit called the Center for Applied Rationality. It was the early days of the rationalist movement: a community formed on the internet whose adherents strove to strip their minds of cognitive biases and subject all spheres of life to the glare of scientific thought and probabilistic reasoning. Galef and her CFAR co-founders — mathematician Anna Salamon, research scientist Andrew Critch, and math and science educator Michael Smith — wanted to translate these principles to everyday life…
>
> Over the next several years, as rationalism became not only the de facto brand of self-help in Silicon Valley but also an intellectual movement followed by pundits and executives alike, CFAR's profile grew; soon, the nonprofit was running

[19] Julia Galef is the author of *The Scout Mindset* (2021), which is freely available from the Effective Altruist website *via* Impact Books: Effective Altruism, https://www.effectivealtruism.org/resources/books. Accessed 2 September 2022. See also resources from Berkeley's 'Centre for Applied Rationality'; 'Lesswrong.com'; and 'The Max Planck Research Group for Rationality Enhancement'. Of course, this approach is also championed by many authors in the self-improvement and popular social science literatures, compare: (Burnett & Evans, 2016); (Kahneman, 2018); Kahneman, Sibony & Sunstein, 2021); (Pinker, 2021); (Taleb, 2001); and (Watts, 2012).

> workshops across the country and teaching classes at Facebook and the Thiel Fellowship. (Wallace, 2021)

Why target such a view? After all, being rational is obviously important to living well, both morally and generally. Perhaps it is not the be-all-and-end-all but then, the modern rationalists do not (expressly) claim it is.

Herein lies the difficulty of criticising the rationality movement. It is not, at least *explicitly*, a totally exclusionary cult. There is, of course, room for other commitments and attachments, provided they are not irrational and do not interfere with the pursuit of ever-greater rationality. That is, rationality emerges as a benevolent dictator, reigning supreme as both gate-keeper and point of salience in one's deliberations about how to live. But it does so only in your avowed interests, and so cannot possibly be either paternalistic or alienating. In a nutshell, the rationality movement is an *ethos* – one that holds significant sway both within and without philosophy – that demands rationality be the focal point of one's limited attention, but not its *exclusive* focus, and this makes it difficult to criticise. Part of the trouble is that it is difficult to *argue* against an *ethos*, since an *ethos* need not be precisely expressible in propositions one can demonstrate to be true or false. But this should not, of course, render them immune from philosophical scrutiny.[20]

In broad outline, my claim is that the rationality movement's focus on rationality-as-self-improvement leaves little room, in the end, for the kind of better life and personal development it was supposed to facilitate. We are cognitively limited creatures – as the rationality movement is fond of reminding us – so focusing on one creed (rationality) tends to come at the cost of others, for example, aiming at being a more *moral* person directly or doing more of what you enjoy. Arguments about opportunity costs and whether the strategy really is the best *means* to the end of self-improvement sound like just the sort of justificatory matrix to which the rationalist might appeal, which is fortunate, as it gives us some common ground.

In the paper's final section, I argue that the constant quest for optimisation in one's life is not optimific. In a nutshell, my argument will be that while you want your way of life to be

[20] A suggestive analogy is with free-market capitalism's lionization of the accumulation of money. This too is an *ethos*, and with a similarly tricky get-out-of-jail-clause. To value money, the capitalist says, is not to value it for its own sake. On the contrary, the value of money is instrumental and lies in how it enables and facilitates the good life for those in possession of it. But, of course, the express aim of facilitating the good life seems to get lost somewhere in the focus on accumulating the money that will fund it.

rational, you do not want rationality to be your way of life. Self-improvement is about much more, and much less, than becoming as rational as you can be.

1. The thought experiment

Suppose that in the not-too-distant future, scientists invented a self-improvement machine. The machine, you are assured, is perfectly safe and the procedure is painless.[21] The machine will add a chip into your brain which contains lots of new information: the basic knowledge package includes much of what we know about science, history, world demography and current affairs. The chip assists your brain to integrate these facts into your existing belief set, revising your existing beliefs where inconsistent with the introduced facts. The chip also assists your brain to conform more readily with the norms of epistemic and practical rationality. Outside improving your ability to reason in accordance with these structural rational requirements, the chip does not update your normative beliefs.[22] The chip does not implant new desires but, to the extent that there is tension between your various desires, and between your desires and beliefs, the chip assists your brain to revise them in line with the requirements of structural rationality (paradigmatically, consistency). The machine, that is, does what many proponents of idealization in the metaethical literature have in mind when they speak of *idealizing* your present evaluative standpoint.[23]

At the risk of sounding like a used-car salesman, let me pause here to emphasize just how attractive this opportunity is. You are being offered the opportunity to self-improve by becoming a far more rational, knowledgeable person than you ever could without the machine's help *and* you will not have to spend any money, time or effort to do it. The oft-noted drawbacks to many journeys of self-improvement – the greatest obstacles to becoming a better version of yourself – have been removed. If self-improvement is all it's cracked up to be, this should be

[21] Let me further stipulate that you do not have to pay for the procedure: perhaps you are the winner of a lottery drawn for this once-in-a-lifetime opportunity. I say more about whether others will also get a chance to step into the machine below. But for now, let us assume you are the first or, at least, that you do not know anyone else who has undergone the procedure.

[22] It is not, that is, at least directly, a moral testimony machine, and does not implant new norms or evaluative concepts.

[23] Of course, there are different accounts of what idealization involves: Peter Railton, Michael Smith, David Sobel and Sharon Street, for example, are going to tell very different stories about how you would be changed as a result of stepping into this machine, and the degree to which the desires of any two people who stepped into the machine would converge as a result.

a slam dunk! Now, this opportunity should be attractive to anyone. But it should be all the more attractive to *you*, since, if you are reading this paper, there is a very good chance you're a philosopher. And there seems something suspicious, perhaps even morally or intellectually *cowardly*, about any philosopher that would not take free knowledge – together with specced up capacities to deploy it – at no cost whatsoever to themselves.

But, then again, the change will not really be *entirely* without cost. Many things you used to think were important or enjoyable, and others that you thought amusing or insightful or worthwhile, will no longer seem so, although these will be replaced by new – and better informed – passions. Not only will you likely wish to spend your time quite differently, you may well (though this is by no means inevitable) reappraise some or all of your personal relationships. You may find spending time with some old friends nowhere near as enjoyable as you once did. This may also extend to your romantic partner or partners (if you have one or some). Nor is there any reason to think this readjustment in personal relationships will be one-sided. People from your old life may have difficulty adjusting to the new you and, if they liked the old you, they may be upset by the change. After all, many of your priorities, interests, values and tastes are quite different from what they were before entering the machine. Plenty more could be said, but you get the picture: changes are in the offing. You are much more epistemically and practically rational, better informed, and operating with a much more coherent evaluative standpoint than before. But will the new you wish to remain in your old life? My guess is that the odds are even, at best.

In light of these risks – and in spite of the obvious benefits – I am reluctant to step into the self-improvement machine. That is, I am reluctant to risk losing my present attachments, values, preferences and desires even where I *know* that losing them in the machine meant they were *irrational*. This sounds odd and yet, I suspect I am not alone in feeling a degree of disquiet at the prospect of undergoing such a procedure. My hunch is that exploring the source of this disquiet can help us better understand the relationships between agency, identity and self-improvement of the more "organic" variety.

How does considering this thought experiment further my express aim of criticising the rationality movement, who think living better is about becoming more rational? The basic idea is as follows. I am targeting a view according to which self-improvement is about becoming more rational and efficient, as rationally and efficiently as one can. My strategy for doing so is to offer the chance to do that *maximally*, by making the value of the rational self-improvement both very significant and costless to the improver or, at least, *intrinsically* costless. I will then argue that this costless and significant improvement is, in fact, unattractive, thereby showing

pursuit of the goal to be suspect, since any goal the attainment of which would be all-things-considered bad for the attainee, is a bad goal, at least if it is their interests with which we are primarily concerned.

One objection to this strategy, at least considering this thought experiment, is that pushing the attainment of some good to its limit can often lead to negative results. That is, one can have *too much of a good thing*. So, the rationalist might counter, we should instead think of the aim of increasing our rationality as a regulative ideal, one of which we will inevitably fall short, but to our benefit. But, I argue, the more limited rationality attained by our everyday efforts is not only not as valuable as that offered by the self-improvement machine, but comes at an equal or greater opportunity cost, due both to the time and effort required and – I argue in the final section – its tendency to collapse into a teleological paradox. In the end, I argue, self-improvement is not really about becoming more rational. If you want to be morally better – or want to convince others to be better – the way to do it is not by sneaking morality in the trojan horse of rationality. And if you want to be *happier*, you should stop trying to optimise. Or so I will argue. The thought experiment will help me do so. Of course, the other reason to wonder whether aiming at rationality directly is a good way to self-improve comes from observations like the following:

> To the extent that the rationalist movement has been written about, its eccentricities have tended to get outsize attention: Some rationalists live in group houses with names like Event Horizon and Godric's Hollow; polyamory and a preoccupation with the existential risk posed by AI are both overrepresented. In opposition to mainstream online culture, which believes that certain arguments should be off-limits, the rationalsphere wants to be able to talk about anything. Slate Star Codex — recently renamed Astral Codex Ten — the most prominent rationalist blog, has caused controversy by countenancing free-flowing discussion of topics such as race science and female harassment of men. (Wallace, 2021).

The rest of the paper proceeds as follows. In the next part, I try to come to grips with my reluctance to step into the self-improvement machine. I consider four candidate explanations concerning:

(i) transformative experience;

(ii) the value of imperfection and the impossibility of transition without loss;

(iii) the idea that becoming yourself is an agential *achievement* that would be undermined by stepping into the self-improvement machine; and

(iv) the risk that self-improvement poses to entanglement in valuable networks of personal relationships.

In the third and final part of the paper, I argue that aiming at rationality in order to self-improve very often undermines our efforts at genuine self-improvement and is therefore self-defeating.

2. Reluctance, agency and uncertainty

When it comes down to it, there are four salient features of the self-improvement machine I find particularly worrying. The first is my uncertainty regarding what the person who emerges post-procedure will be like. The second is that I am entirely passive during the transformation. My self improves, but not through my *agency*. The third is a lingering feeling that my many imperfections and irrationalities are partially constitutive of who I am in a non-trivial sense. To the extent that I value myself, therefore, there seems to be some reason to value them. It is not just that I value myself *in spite of* my imperfections and irrationalities (though sometimes there's that too). It is that I value being the person that I am and recognise I wouldn't really be me without them. Losing all my imperfections and irrationalities would be, it seems to me, to become someone else entirely. And I am not sure that's something I want. Fourth, and finally, there seems to be something wrong with someone willing to risk all their personal relationships for their own ambitions for improvement. I don't think I want to be that person, nor the person they become as a result.

But how cogent are these worries really? Consider an analogy with *physical* self-improvement. If there were a machine into which I could step that made me much healthier, fitter, stronger, faster and trimmer than I am now, I would step into it without requiring a moment's further thought. It would matter not a jot to me that I exercised zero agency in this process. The choice between actually spending the next year running, eating salads, abstaining from alcohol and so on, or being as healthy as I would be if I did all that *but not have to do any of it*, is, for me, not a difficult one.[24] So, what makes the decision to step into the self-improvement machine any different?

[24] I should add that this does not mean this would be an easy decision for everyone. For example, if physical achievements were central to my life's meaning – perhaps as a professional athlete or the like – then there might be very weighty reasons not to step into a physical self-improvement machine, just as there are weighty reasons for professional athletes, for example, not to take steroids to improve their physical performance. There is, of course, a significant literature concerning the human enhancement debate generally. Cf. (Kahane & Savulescu, 2013).

One immediate response to which you might be attracted is the following. Many of us have – or think we have – a pretty good idea of what it would be to be healthier or fitter. So, if the physical self-improvement machine offers us that – and does so without cost – then we should go for it. But most of us do not have an equivalently clear picture of what it would be to just "self-improve" *in general*. So, part of the problem seems to be that it is far from obvious that what the self-improvement machine offers can be directly equated with self-improvement generally and, in particular, it is far from obvious that it offers anything like a guarantee that the person who steps out will be a *morally* better person. Let me concede immediately that the self-improvement machine offers no guarantee that the person who emerges post-procedure will be morally better.[25] And let me add to that concession an even more embarrassing admission: I have *no idea* whether stepping into the self-improvement machine will make you a morally better person. There are at least two reasons for my uncertainty on this point. First, I am unsure whether my ideal – or your, or anyone's – hyper-rational and very well-informed epistemically "ideal" counterpart will arrive at a morally praiseworthy evaluative outlook *just because* they know lots more and reason far better than they did previously. Perhaps they will, if you have a sufficiently moralized starting point. But the possibility, for example, that the machine turns you into a hyper-rational, callous, prudential terminator or an ideally coherent Caligula is, I think, sufficiently chilling to give some pause.[26] Second, I am not sure whether my epistemically ideal self could be *akratic*. I can – or, at least, think I can – imagine a scenario in which I exit the machine painfully aware of my cataclysmic moral failings and utterly paralysed by them, unable to find the will to lead the life I know I now should, and horrified by the life I have led up to this point. That scenario is something close to a nightmare and, if it's a serious possibility, seems like a good reason not to step into the machine. If things are that bad, then perhaps ignorance is bliss after all. So, given my uncertainty about whether idealization entails a morally praiseworthy evaluative outlook and my uncertainty regarding the truth of judgment internalism, I already have some concerns about stepping into the machine. But, of course, if you are sufficiently confident in your prediction about the effects

[25] Of course, not all philosophers would agree. Michael Smith, for example, would likely think that the agent who stepped out of the self-improvement machine would be morally better and that the desires of multiple agents all of whom underwent the procedure would converge (Smith, 1994).

[26] See (Street, 2009) for a discussion of the possibility of an ideally rational Caligula, though admittedly I am less sure that his starting point would need to be so very alien from our own.

of idealization on your evaluative standpoint and the truth of judgment internalism, then perhaps these worries can be disregarded.

In what sense, then, is this machine rightfully described as the 'self-improvement machine' rather than 'the rationality-epistemic-improvement machine?' After all, and as already mentioned, there seems to be far more to self-improvement than becoming more rational. Is that 'something more' exhausted by acknowledging the moral dimension to our understanding of self-improvement? In my view, no. In fact, while I think some philosophers are primed to hear 'self-improvement' in a moral register, the vast – and extremely profitable – popular self-improvement literature suggests that its pundits and readers are not thinking about moral improvement *at all* when they think about self-improvement. While, as one might expect, there are many different approaches to self-improvement in popular culture – from working on emotional regulation, to appropriations of "Eastern" spirituality. and everything in-between – the rationality movement has its own take.

In broad outline, much of this literature aims to help its readers reason better about themselves, others, and what they want out of life, and assists them to set and then achieve their goals, whatever they may be, once they have worked out what they are. These pundits suggest that knowing more and thinking clearly – free of various irrationalities, biases, self-defeating habits and unhelpful patterns of thinking – is a route to a better you and a better life as judged by your own lights. Indeed, the entire self-improvement industry is premised on the idea that its pundits have knowledge that can be transferred via testimony which will enable the recipient to live a better and more fulfilling life by becoming a better version of themselves.

If that interpretation of the very general point and promise of the popular self-improvement literature – and the rationality movement's spin on it – is correct, then it is strikingly similar to what the self-improvement machine offers, which is in turn strikingly similar to what some proponents of idealization in the metaethical literature think is definitive of or authoritative with regard to, variously, one's reasons or welfare. This unexpected consonance between the conception of self-improvement implicit in much of the rationality movement's popular self-improvement literature and a metaethical device deployed for thinking about the preferences of our ideal counterparts and their relevance for how we should live our lives is, perhaps, slightly less surprising if one considers a venerable philosophical tradition that pre-dates both. For a great many philosophers from Plato's Socrates to Aristotle on have also thought that the route to a better life was to become a better person *and* that the route to becoming a better person was reflection, contemplation and the acquisition of knowledge. As it turns out, the contemporary self-improvement literature has hallowed

philosophical predecessors. So, I think the self-improvement machine deserves its name.[27] Like many more 'organic' journeys of self-improvement, it offers *not* pre-determined goals or an objective list of goods, but the chance to do better at working these out for yourself, together with uncertainty about where you'll end up and who you'll be at the end of the process.[28] But although this sounds exciting and tempting, the uncertainty at the core of this decision also leads to the first of my four candidate explanations for my reluctance to step into the self-improvement machine.

2.1 Transformative experience

Errol Lord opens *The Importance of Being Rational* with the following observation:

> We devote a good portion of our lives striving to figure out what it makes sense to do, believe, intend, desire, and hope for given our often limited information and abilities. In other words, a good portion of our lives consists in striving to be rational… we are not mistaken for devoting so much of our lives to being rational. Rationality, in my view, is of fundamental deontic importance—i.e., in my view, rationality plays a fundamental role in determining how we ought to react to the world. (Lord, 2018, p. 3)[29]

Since rationality determines 'how we ought to react to the world,' becoming more rational entails reacting to the world better. So, if you want to become a better person, you should become more rational. That is, if your *end* is becoming a better person – morally, epistemically, prudentially – then the *means* you should adopt is to become more rational. Something like this brief argument lies at the core of much contemporary rationalist self-improvement thinking.

But although it is true that we spend much of 'our lives striving to figure out what it makes sense to do, believe, intend, desire, and hope for given our often limited information and

---

[27] Of course, this is not to say that different self-improvement machines are not possible. One can imagine a self-improvement machine that changes your desires, preferences or dispositions directly, for example. Perhaps a committee of moral philosophers could come together and decide the ideal desiderative profile for the maximally virtuous agent (or cognitive scientists for the happiest, and so on). But, for reasons I suspect are obvious, I find the prospect of stepping into such a machine even more troubling.

[28] Indeed, if you *knew* in advance how the machine would change you, this would go some way to rendering it otiose.

[29] The centrality of rationality to our agential lives is, unsurprisingly, a topic of perennial philosophical interest. Cf. (Broome, 2013); (Fogal, 2020); (Gigerenzer, 2010); (Kiesewetter, B. 2017); (Southwood, 2008); and (Worsnip, 2021).

abilities' and that, therefore we spend much of our lives 'striving to be rational,' it is not true that we pursue the former aim *by* pursuing the latter. Nor is it obvious that we *should*. Aiming at increased rationality directly inverts the relation between means and end precisely because now what we are intending, desiring, and hoping for is to become more rational, rather than figuring out what it makes sense to intend, desire or hope for. The problem with being a cognitively limited creature is you have limited resources to deploy in furtherance of your aims and aiming at ideal rationality makes stringent demands on us: to become far more rational than we are now is not just beyond the powers of most of us, it would be to become an entirely different person. This is not self-improvement, but self-replacement. Let me explain.

It is difficult to predict what an "idealized" version of myself would be like: how my beliefs, preferences, desires and attachments would be affected by the self-improvement machine is likely not only uncertain but indeterminate.[30] I have argued elsewhere[31] that there are many possible idealizations of any given nonideal evaluative standpoint. If that is right, then attempted predictions concerning who will step out of the self-improvement machine are very likely to go awry (Paul, 2014; Callard, 2018) and the machine will be required to resolve considerable indeterminacies in selecting between possible idealizations.

The self-improvement machine is an example of what Laurie Paul calls 'transformative experience' (Paul, 2014). Transformative experiences are supposed to generate special problems for orthodox theories of rational choice because it is impossible to predict what they will be like in advance, thereby undermining our ability to price the relevant option. On Paul's model, concern for my *self-preservation* is to be weighed against the value I place on *revelation*: on finding out how my life will go if I decide to step into the self-improvement machine. The basic idea is that while there *seem* to be weighty reasons in favour of stepping into the self-improvement machine, they cannot be weighty reasons *for me*. But, against this application of Paul's theory, one might worry that the self-improvement machine comes with a feature that most other transformative experiences do not: a guarantee that I will *know* more and be more *rational* as a result of the experience.

This might instead make the decision instead amenable to the sort of reasoning involved in what Agnes Callard (2018) describes as '*proleptic* rationality'. Sometimes we engage in long term transformative projects, *viz.*, those in which we aspire to transform ourselves by striving to understand values we grasp only inchoately at the outset. The reasons on which we

[30] Given some plausible assumptions about multiple-realizability, rationality and incommensurability.

[31] Reference omitted for blind review process.

act are *proleptic* versions of the full-blooded reasons we hope, eventually, to understand as our future self. While appealing, the problem with applying Callard's model to this thought experiment is that the self-improvement machine involves giving myself over to a process *beyond my control* and granting authority to a future self whose values I may not even grasp inchoately at present. It is therefore difficult to see how proleptic reasons – posited to explaining the *agency* of becoming – can help here. In fact, the lack of agency in the process is part of the *problem*. It is to this worry I turn in section 2.3. But before I do, I want to consider two further points.

The first is whether my discussion of self-improvement or, perhaps, discussions of self-improvement generally, unhelpfully collapse the distinction between someone's becoming better and something's being better for someone. One might object that *since* it is far from obvious that becoming better – in the sense of being more rational and knowing much more – will make the person in question *happier*, it is also far from obvious that the allegedly strong *prima facie* case in favour of stepping in is nearly as strong as it initially appeared.[32] But, of course, I am very happy to concede this point. As I argued at the outset, many people think the only drawback to rational self-improvement is the hard work it requires. Part of what makes the self-improvement machine thought experiment so helpful, in my view, is that it clearly shows how our becoming better might be worse *for us*. Indeed, it seems possible we might be miserable as a result of being far rationally "better" than we ever thought possible. If that is correct – and I think the self-improvement machine illustrates how it might be – then hard work is *not* the most important drawback to every project of self-improvement. This point is also significant because the rationality movement proceeds on the assumption that becoming more rational, better-informed, knowing yourself and others better, becoming more efficacious and so on *will* lead to a happier and more fulfilling life.[33]

---

[32] It is also perfectly consistent with ideal advisor accounts of welfare or reasons that one's ideal self would counsel against stepping into the machine, if that would make one unhappy. But that result is far from guaranteed. One venerable reason for suspecting they might not is encapsulated in J.S. Mill's (in)famous observation that 'It is better to be a human being dissatisfied than a pig satisfied; better to be Socrates dissatisfied than a fool satisfied. And if the fool, or the pig, is of a different opinion, it is only because they only know their own side of the question' (Mill, 1957). It is an interesting further question how this situation differs from one in which the person who steps out of the machine regrets having stepped in.

[33] Of course, *if* you think the self-improvement machine will also make you morally better, then there is further a difficult question to answer. Are you morally required to step into the self-improvement machine if you suspect you will be less happy but more virtuous as a result? One might even argue, more ambitiously though

Final point: if stepping into the self-improvement machine is a genuinely transformative experience, then you have maximal rational latitude in deciding whether to do it.[34] That is, *that* the self-improvement machine constitutes a transformative experience offers a complete justification for not stepping into it if, upon reflection, you decide you'd rather keep your current, unimproved (and relatively uninformed and irrational) evaluative outlook rather than find out what it would be like to be an improved version of yourself. Given how strong the presumptive case in favour of stepping into the self-improvement machine is, I think this is quite surprising. More importantly, I think it teaches us something further about the relationship between rationality and self-improvement.

The upshot of the foregoing analysis, in my view, is that aiming at rationality as a means to self-improvement is a muddled approach *unless* you think the aim of self-improvement *is* to become more rational. But even the diehards of the rationality movement do not think this. They think aiming at rationality is a good way of *achieving* self-improvement. But the foregoing discussion suggests that very significantly improving your rationality could take you in a great many different directions – and some of them are not plausibly regarded as improvements in any sense other than as conforming with some of the demands of informed structural rationality. A perusal of some of the views espoused on Lesswrong.com or Astral Codex Ten will support a similar conclusion.

What the self-improvement machine demonstrates, in my view, is that one can have right-kind reasons to remain less than ideally coherent. That is, right-kind reasons to be structurally *irrational*, and that, among them, are reasons pertaining to the hope that one might become a better, if still vastly irrational, person.[35] If that is true, then we should both reject claims that equate self-improvement with improved rationality and claims that improving one's rationality is the best route to self-improvement. Of course, it may often be true that by becoming a better person, you also conform better to some of the demands of rationality. But that is a very far cry from thinking that the best way to become better is to become more

---

somewhat less plausibly, in my view, that because the self-improvement machine will make you far more rational you are rationally and perhaps therefore also normatively required to step in, even if you will be less happy as a result. This is, of course, a point of significant contention. Cf., for example, (Broome, 2013) and (Kolodny, 2005).

[34] Assuming, of course, that Paul's argument goes through and, perhaps, subject to your answers to the questions posed in the preceding paragraph.

[35] On the distinction between right-kind and wrong-kind reasons to be (ir)rational see (Worsnip, 2021, Ch. 3). Cf. (Elster, 1983).

rational. In the next section, I say more about how concern for our own agential identity and well-being might block the claim that self-improvement consists in improved rationality.

2.2 The value of imperfection and the impossibility of transition without loss

The previous section's conclusion could be worked up into the following dilemma for the rationalist. Either rationality does not require stepping into the machine, i.e., does not require becoming *more* rational, given the associated costs, or it does. If it does not, then it is perfectly rational not to dedicate one's limited cognitive and temporal resources to the project of rationalism. If it does, then what the self-improvement machine shows is that there are sometimes very compelling right-kind reasons *not* to attempt to become far more rational than you are now. *Either way*, it looks like the rationalist project is at best an optional and dubious method for pursuing one conception of cognitive self-improvement and, at worst, a direct inhibitor to pursuing other, more important projects of self-improvement, namely, those aimed at moral self-improvement directly.

But is my reluctance to step into the self-improvement machine just explained by a bias towards the person I just happen to be right now? If so, this is precisely the sort of wrongheadedness rational self-improvement will annihilate.

A tempting response is that to the extent I value myself, this value transmits to the many imperfections that make me who I am.[36] But this leads to *prima facie* odd results when considering my various cognitive and affective defects in isolation. Surely, I have good reasons to ameliorate each of them individually. It cannot be a reason against reading a science or history book that it will unsettle my current sense of my place in the world, can it? If not, then why should the value I place on myself now block elimination of my many false beliefs and irrational attitudes?

Drawing an analogy with the *Ship of Theseus* is suggestive. The basic idea is that where my values, beliefs, dispositions and attitudes are updated gradually and due, in part, to my agency, there is a sense of continuity that preserves the original, albeit *improved*. But at some point – it is difficult to say precisely where, but certainly well before the degree of sudden change envisaged in the self-improvement machine thought experiment – the changes begin to look like *substitution* rather than *improvement*. In this regard, it is worth noting that although the changes from one state to another are improvements, they nevertheless entail loss of the

[36] I think this dovetails with Guy Kahane's (2021) more general – and to my mind, persuasive – argument that, roughly, we should not regret that the world is not perfect.

parts of the self with which they are incompatible. Perhaps sometimes this will be an altogether good thing, but often it will not. Consider the transition from the naivety of childhood, to adolescence, to adulthood. Of course, each progression involves improved understanding of the world and one's place within it, but it would be foolish to suggest that this improved understanding is costless. Given the degree of violence likely done to one's starting attachments in the self-improvement machine, it is highly likely that much that one now values will be a casualty of the experience. This alone is some reason to be reluctant regarding the change. But there is a compounding factor. The progression from child to adolescent to adult can plausibly be construed as the overcoming of successive stages in a linear development. But the coherence the self-improvement machine purchases comes at a much greater cost. There are likely many different revised coherent combinations of the values and preferences now constitutive of your nonideal, incoherent evaluative standpoint. But each route to coherence will therefore involve pruning certain values and attitudes that might have been preserved, had another route been taken. If this is true, then – like an artwork (likely, all artworks) that resists a single authoritative interpretation – the transformation to the new, improved and coherent you comes at the cost of the *lost potential* to be many other selves: the different yous you might have been. The opportunity cost of rational self-improvement is therefore high. Of course, the lost opportunity to develop certain valuable aspects of yourself is not uncompensated. You gain in rationality and, perhaps, in the possibility to develop further in new and unexpected directions. But compensated loss is different to no loss at all. Even recognising that one cannot have it all, the value you place not just on *who* you are today, but the *options* you have as a result, is no small matter (Setiya, 2017). Not only this, but the decision to step into the self-improvement machine requires adopting an attitude to our past, self, and personal attachments and relationships that, one might worry, is itself deeply alienating (Wallace, 2013).

It is tempting, at this point, to adapt a point made famous by Susan Wolf (1982). Would you really want to *be* a moral (or rational) saint? My point is, would you really want to *become* a moral (or rational) saint, especially if the cost of doing so was losing yourself in the process? I want to emphasize that I do not intend to make a strong claim about the metaphysics of personal identity. I do not claim that, say, on a Lockean or Parfitian model you would have no reason for prudential concern vis-à-vis your improved self (Parfit, 1984). They will still remember being you even if you cannot now imagine what it will be like to be them. But you do also have legitimate prudential concern for the self you are now which, in my view, extends to not dying a sort of normative death as a result of losing many of your most basic attachments. In one sense, this should be obvious. But in another, as I mentioned at the opening of this

section, it has counterintuitive implications. Can it really be better for someone, say, an anti-vaxxer or climate-change denialist, to retain the false beliefs and values around which they have organised their life? Perhaps this point could be worked up into the following dilemma. If a person is sufficiently irrational, then it will be good thing that the machine fundamentally alters them in ways that would make them unrecognisable to their past self. But if, on the other hand, a person is not *all that* irrational, then the machine will not do too much violence to their evaluative outlook. So, either way, you ought to step in. Now, I do not doubt that it could be better, all-things-considered, *that* a deeply irrational person step into this machine and thereby lose the uninformed and harmful views that are central to their sense of self. But I do doubt that it can be better *for them*, since I do not think they remain the same ethical self throughout the change. This is not a reason not to step into the machine if they otherwise desire to do so. Some people actively seek out such utterly transformative experiences because they are dissatisfied with their current evaluative outlook or sense there is more value out there than they have yet experienced. Nor is it a reason for all of us not to wish that such a machine were widely available *and* that such people would step into it. But it is, in my view, a reason available to each and every person that could justify refusal to do so, even though this might be worse for everyone else, all things considered. We all have a vested interest in being ourselves, even if it would be better if we were someone else.

2.3 The idea that becoming yourself is an agential *achievement* that would be undermined by stepping into the self-improvement machine

In this section, I explain why the hope for greater achievements in our lives – and the achievement of becoming a better version of ourselves – is actually *undermined* by stepping into the self-improvement machine, *unless* our aim is to benefit others. I then argue that if that is the real aim of improvement, we should aim at that directly instead.

'Self-improvement' is suggestively ambiguous. On the one hand, it involves improving a self – any self – by any means or agents possible. On the other hand, it could also be reflexive: meaning it should be the self who *works on itself*. If the value of self-improvement inheres solely in the difference in value between original and improved self, then the means and agent of change are irrelevant. But this, I argue, is an unacceptably *alienated* conception of self-improvement. If the *agential process* of self-improvement is itself valuable, then we have a neat partial explanation of what is so disturbing about the self-improvement machine. As it turns out, this explanation is highly suggestive for thinking about self-improvement generally. The telic process of self-improvement has significant *atelic* value. This explains why, for

example, it can be a central part of a meaningful life well-lived to work hard on my verse, backhand, jokes or painting, even if I will never be even a half-decent poet, tennis player, comedian or artist. Much of what gives life its meaning involves striving to *do better* with no thought or hope of *being best*. The self-improvement machine *maximises* my potential only by depriving me of responsibility for the achievement this represents, not to mention the enjoyment I could have had along the way. Moreover, having the values, beliefs, dispositions and attitudes I do is much more important to me than my dubious prowess on the tennis court: and my hope that they will be replaced by improved values, beliefs, dispositions and attitudes in the future as a result of my efforts is an important reason to get out of bed in the morning. The self-improvement machine undermines my agency in becoming a better agent and, in my view, this is a potent antidote to its Faustian allure. This is one of several points at which comparison with Robert Nozick's (1974) famous 'experience machine' thought experiment is suggestive. There, the worry was that simulated achievements are less valuable than real ones. Here, the worry is that becoming yourself is no longer an achievement at all, at least, not one attributable to *you*.

In response, one might object that once I step out of the self-improvement machine, I will have new opportunities to develop yet further. In fact, I will be better equipped to *achieve* post-machine. Accordingly, concern with agential achievement is actually a weighty reason to step into the self-improvement machine.

I believe there is something importantly right about this response, about which I say more later. For now, let me make three observations in support of my lingering doubts to the contrary. First, when assessing agential achievements, causal histories are generally considered highly relevant in assessing whether the success is attributable to the agent in such a way that they meet the credit condition (Bradford, 2015). Take Lance Armstrong's cycling "achievements". Where a contestant cheats to gain an advantage over their competitors either before or during any contest, they will usually also have trained extremely hard, exhibited a high degree of skill, and so on. Yet the fact of their cheating – say by doping to allow them to train even harder for the race – *undermines* their achievement *despite* their skill and effort. One might worry that, analogously, one's later agential achievements will be undermined by the prominence of the self-improvement machine in one's causal history post-procedure. In short, your future achievements are not attributable to *you* but to *it*, even if you put plenty of work in too.

My second observation is closely related to the first. Causal histories are not just relevant to the credit condition, they are also highly relevant when accounting for the relevance

of *difficulty* to achievement. Suppose that stepping into the self-improvement machine is not akin to cheating. Perhaps it is more like having state of the art training facilities, equipment, financial support and coaches. Our expectations of performance – where we set the bar for genuine achievement – are often relativised to the skill, opportunities and support available to the agent in question. The basic thought is that the better your starting point, the more you need to do for your efforts to count as a genuine achievement. For me to one day play the *Moonlight Sonata* would represent a remarkable achievement, since I presently cannot play piano *at all*. For a virtuosic concert pianist, it is not presently an achievement at all (or only a very minor one).[37] In part, what makes one a virtuoso is the ease with which you can perform feats that would represent an immense achievement for others. Naturally, *becoming* a virtuosic concert pianist *is* an immense achievement, but this is precisely the sort of achievement undermined by stepping into the self-improvement machine.

If our concern in agential achievement is to perform well relative to our starting point, then the problem with stepping into the self-improvement machine is that it shifts the goal posts. What would have counted as an achievement pre-procedure will likely no longer count post. So, if the objection raised earlier – *viz.* that a concern with agential achievement is actually a weighty reason to step into the self-improvement machine – is to gain traction, it must counter the thought that doing so is actually *self-defeating*.

My third observation concerns a possible objection to the line of argument just sketched. Some achievements are not just valuable relative to the capacities and resources of the agent(s) in question – my piano performance or a 100-year-old running a marathon – they are valuable intrinsically or, perhaps, relative to humanity's efforts as a whole. Picasso's *Guernica*, flying to the moon, the discovery of penicillin, splitting the atom: these are not just valuable achievements because they required tremendous skill and effort. They are valuable because their products are valuable or push the frontier of humans' understanding of the world and our place in it. Let's call these "agent-neutral" achievements.[38] Perhaps stepping into the self-improvement machine will significantly increase one's chances of succeeding at some agent-neutral achievement.

For some agents, I suspect this *is* a good reason to step into the self-improvement machine, subject to two very important caveats. First caveat: in considering the value of agency

[37] This is not, of course, to suggest that a specific and breathtaking performance of it *might not* represent an immense achievement for the virtuoso, but this will not be just the bare fact of playing it.

[38] That is, they are or would be genuine achievements regardless of who performed them.

and its relation to self-improvement it is important that we not unduly *instrumentalise ourselves* in service of the achievement of some external good. There may well be occasions where one has good reason to sacrifice some or all of one's present attachments, desires, plans and values in service of some higher aim, but my hunch is that for most agents these will be vanishingly rare. Especially – as here – where the higher aim in question is presently unknown. Second caveat: there has been a slide in the discussion of this section from self-improvement as an agential achievement in itself, and self-improvement as a means to other agential achievements. Even if it occasionally makes sense to instrumentalise our own agential development in furtherance of other aims, including by diminishing our agential control over the process of improvement, this will nevertheless come at the cost of the unique agential achievement developing ourselves represents. On this view, self-improvement is not just one 'ground project' among many (Williams, 1976). It is *the* ground project underpinning all others. If the self-improvement machine removes our agency from this fundamental ground project – and the discussion of this section provides some reasons to think it does – then this is a powerful reason to resist its siren song.

If sound, how does the argument of this section bear on whether we should aim directly at improved rationality as a means to self-improvement? Let me make two suggestions. First, if the *rationale* for improvement is to benefit others, we will likely do far better to aim directly at that, especially given the opportunity costs associated with the time and effort required to become a far more rational and knowledgeable version of yourself before then working out how to help others. Why attempt to spec yourself up, given the time and effort involved, if there are far more efficacious ways to benefit humanity? This sounds like an argument that goes to the core of the alleged *rationale* for rational improvement those in the rationality movement, including effective altruists, should take seriously. My suggestion: if it is all about helping others in the end, you need not – perhaps *should* not – dedicate time and resources to becoming a far more rational version of yourself. Becoming a better person is not really an *intellectual* exercise, nor is it one that requires exemplary rationality. It might be that those who behave badly can always be shown to be irrational (though I am skeptical). But eliminating irrationality is not obviously therefore the best, let alone *only*, way to become better. This is especially true given that the sort of improvements we can achieve alone are far less than what the self-improvement machine offers. But if, on the other hand, the aim of becoming a far more rational version of yourself is simply to become a more *rational* person for its own sake, the question is, *why*?

2.4 The risk that sudden improvement poses to entanglement in valuable networks of personal relationships

The final candidate explanation is the interpersonal analogue of 2.2: if I value my relationships with others and my embeddedness in various communities, the risk that improvement will damage those relationships or my membership in those communities might seem like a weighty reason *not* to improve. But this has troubling implications along both the personal and interpersonal dimensions.

Consider the following advice: any good relationship will leave you room to grow. Sometimes, even in the best relationships, people *grow apart*. But attachment to one's present relationship(s) is not a reason to stunt one's personal growth. Furthermore, any relationship *premised* on your not improving yourself is, just for that reason (and, likely, for many others) a bad relationship. All the more reason to jump ship. So, a litmus test for weighing improvement and attachment may well be whether the improvement contemplated is consistent with maintenance of existing valuable personal ties – and is therefore a form of situated personal growth – or whether it will require breaking those ties. If the latter, it always worth asking whether ties that require this form of sacrifice are worth keeping. After all, sometimes you really should *put yourself first*.

This advice sounds sensible. Does its good sense mean that the danger the self-improvement machine poses to my various personal relationships and my membership in the communities that are an important part of my identity *should* be disregarded? I am less sure. It is one thing to acknowledge that it is often morally permissible and sometimes also prudent to prioritise one's personal development over one's personal relationships. It is quite another to suggest that deciding not to is irrational. Leaving some relationships and communities can leave one in a form of exile: no longer accepted by one's old community but not accepted anywhere else (Pickard, 2020). And some relationships are irreplaceable (for example, with your parents or children). A troubling upshot of scenarios like these is that one may have very weighty reasons to remain in a relationship or community in which one's congenial relations are maintained, in part, by resisting the temptation improve oneself. In the end, the lesson is once again that the transition is unlikely to be without loss. How to weight that loss is likely a matter for each individual. This problem is compounded by the uncertainty inherent in the self-improvement machine, though beginning a journey of self-improvement is often unpredictable in many ways similar in kind, if different in degree. Finally, the *opportunity cost* inherent in the improvement countenanced in 2.2 reverberates through its interpersonal analogue here: the many valuable relationships you *could have had* before you stepped into the self-improvement

machine. Many of our most important and valuable relationships in life come from the journey of specific attempts at discrete tasks of self-improvement: teachers, mentors, friends we make in our formative years of hardship and personal development. A final worry with the self-improvement machine is therefore that it cuts us off not just from the atelic value we experience along the road to improvement, but the people with whom we might have shared it. Of course, if you join the rationality movement, then you will be a member of a new community: one that shares your quest for optimisation. But we should be wary of joining any group that makes relating to and remaining in others that much more difficult; and the rationality movement is one such group.

3. Improving self-improvement (is not about optimisation)

Might there be something a bit *off* about someone who would risk all their valuable personal relationships to realize their dream of rational self-improvement? I think so. Just as we do not always admire people who are willing to sacrifice everything (and others) for their own ambition, so the desire to become a more rational, better version of yourself at all costs can seem pathological. My hunch is that what is at work in the pernicious rational self-improvement case is a charge of *self-absorption*. The problem is that sometimes a healthy concern for personal development devolves into an unhealthy and obsessive main-character syndrome, with predictably deleterious consequences for those closest to the self-styled protagonist. So, a final, and significant, potential drawback to self-improvement is the risk that it leads to self-absorption. Compounding – or, perhaps, exemplifying – this worry is that self-improvement of the kind discussed in this paper carries with it a whiff of *teleological paradox*. This risk obtains precisely because directing our attention towards our own self-improvement can prevent us from attending to the matters required to achieve it. This problem looms largest in the context of moral self-improvement – where focusing on ourselves directs attention away from genuinely focusing on others – but it seems almost as pressing when considering attempting to improve along other dimensions as well, most notably for our purposes, rationality. This is troubling, not least because a significant takeaway from the argument so far seems to be that ensuring our agential control over the process of self-improvement is a key desideratum of realizing its value for us and bulwark against the pitfalls the self-improvement machine thought experiment makes so stark.

So, how could spending time on becoming a more rational agent ultimately interfere with the task of self-improvement? By now, you know what my answer is: aiming at rationality directly is not only a suboptimal way of becoming a morally better person, it is a suboptimal

way of becoming a happier one. But if the aim is not to become morally better or happier, then what is the goal of self-improvement the rationality movement is selling us? The most obvious alternative is: becoming more rational for its own sake or, perhaps, because this is just what rationality requires. By doing what you are rationally required to do, you are "better" than those who do not (including your past, less rational self). In this paper, I have given some reasons for thinking this is a deeply misguided way of thinking about both rationality and self-improvement. Aiming at rationality for its own sake does not obviously lead to self-improvement worth the name, nor necessarily conduce to the happiness of the self-improver.

Conclusion

If I had the chance to step into the self-improvement machine, I would not do it. I suspect rationalists would take this as conclusive evidence of just how badly I need the rational self-improvement the machine offers. But among my reasons for not wanting to step into the self-improvement machine are that it seems to risk my chances of genuinely improving in the ways that really count. That is not to say that irrationality is the key to living well or better than you are now. It is just to make the more circumspect claim that living well – and better – is not really about focusing on rationality. If you aim instead at the things that really matter in life, it is not unreasonable to hope you might thereby conform to some of the key demands of substantive rationality along the way. If living well is what you desire, trying too hard to optimise is not, after all, the best optimisation strategy.

CHAPTER 3: SELF-CREATIONISM

Abstract

Is genuine self-creation – understood as self-directed value-acquisition – possible? Many philosophers think not. Agnes Callard disagrees. I argue Callard's solution to the problem of self-creation fails. Then, I propose an alternative solution. The solution has wide-ranging implications for our understanding of value-acquisition and agency.

PART I: THE PROBLEM OF SELF-CREATION

1. Self-endorsement, self-cultivation and Strawson's dilemma

Is genuine self-creation – understood as self-directed value-acquisition – possible? Many philosophers have worried it is not. Some cite with approval Nietzsche, who pronounces the idea of self-creation a 'rape and perversion of logic' and mocks the philosopher's pretension 'to pull oneself up into existence by the hair, out of the swamps of nothingness' (*BGE*, sec. 21).[39] One popular solution – variations of which are found in the work of Harry Frankfurt (1971; 1988), Christine Korsgaard (1986; 1996; 2009), and Gary Watson (1975; 1993) – is to look to the varying roles *self-endorsement* in assessing our individual values. But, as Agnes Callard (2018) argues, endorsement cannot *create* new values. At most, it can lead us to retain or abandon a value we find we already have. More promising, argues Callard, are *self-cultivation* models, proponents of which she identifies as Jean Hampton (1993), Michael Smith (1994), Michael Bratman (2000; 2003), Richard Holton (2004), and Joseph Raz (1988). But Callard adapts an argument of Galen Strawson's (1986; 1994) to suggest that all value cultivation is subject to the following dilemma. The new values must either be rationally continuous or rationally discontinuous with the values an agent already has. If they are continuous – in the sense that my new values are entailed by my old ones – I don't really change. But if they are discontinuous, then any change must come about accidentally or from external forces – in which case I don't really change *myself*. Either way, it looks like genuine self-creation is *impossible*.

Since the original argument of Strawson's that forms the basis of this dilemma was designed to show the impossibility of the kind of freedom ultimate moral responsibility

[39] Cf. (Noggle, 2005) and (Leiter, 2001).

requires, an important question to address at this point is whether the problem of self-creation is just the free-will problem for the self-absorbed.

It is not. The situation is in some ways worse, though in some also better. Starting with worse: many solutions to the free-will problem nevertheless rule out the possibility of genuine self-creation, including those that side-step arguments like Strawson's because they make the possibility of (a certain conception of) self-creation look remote. So, while the impossibility of free will rules out self-creation, proving our freedom will not guarantee it.

But the situation is also better in other respects because the self-creation problem need not, and should not, be understood in terms of *ultimate* responsibility for oneself. The self-creation problem asks something like whether there could be two selves, call them 'S_2' and 'S_3', such that S_3 is a genuinely new self, distinct from S_2, and S_2 created S_3. But this does not require taking a stand on whether, for S_2 to be responsible for S_3, there must be an earlier self 'S_1' who created S_2, and for whom S_2 is also responsible, and so on. That is, the problem of self-creation need not address the possibility of an infinite regress of creator selves to determine whether a local act of creation is genuine and of the sort for which someone can be *locally* responsible.

This is important. Not just because it allows us to defer "solving" the free-will problem, at least as Strawson conceives it, to another day, but because the possibility of this kind of local change and responsibility is often more important to us anyway. The possibility of self-creation is what a person cares about when they ask – recognising some deep flaw(s) in themselves – whether someone can ever *really* change. Solving the problem of self-creation is not about whether the change in question was itself already written in the stars or whether they can be held ultimately morally responsible. It is about whether it is reasonable for us to hope that by working on ourselves we can become better; or see life from an entirely new perspective; or break free from the confines of an evaluative perspective in which we no longer feel comfortable. So, while the free-will problem is of immense metaphysical and moral significance, it may not (as is often, but not un-controversially, noted) be of much practical significance. But the problem of self-creation is practically significant and, if the argument of this paper is correct, solvable.

2. Callard's solution

Here is Callard's summary of the problem of self-creation and her solution to it:

> The new values, acquisition of which constitutes my act of self-creation, must be either continuous or discontinuous with the ones I already have. If they are

> continuous, I am not changing but rather working out the implications of the person I already was. If they are discontinuous and the new values contradict or come at a tangent to my old values, the change is not a product of my agency. I change, but I do not change myself. This paradox, adapted from the work of Galen Strawson, can be solved if we allow that the direction of value-dependence may be teleological: the aspirant's values depend on, and are entailed by, those of the person she is trying to be. The aspirant does not fashion, control, or make the self she creates. Instead, she looks up to that self, tries to understand her, endeavors to find a way to her. (Callard, 2018, p. 179)

At the core of Callard's solution to the problem of self-creation is her more general account of proleptic rationality. Callard argues that many of life's most important experiences, choices, achievements and relationships are only possible because we engage in large scale transformative pursuits: pursuits in which we get ourselves into some valuable state of affairs without having known from the outset exactly where we were headed. She labels this phenomenon 'aspiration'. Proleptic reasons are the reasons that explain the rationality of aspiration: '[a] proleptic reason is an acknowledgedly immature variant of a standard reason.'[40]
γ is a proleptic reason for an agent to ϕ iff:[41]

(1) the agent desires to ϕ for the sake of γ;
(2) the agent does not yet know the value of γ;[42]
(3) the agent knows that (2), i.e. the agent knows that she does not know the value of γ;
(4) the agent desires to know the value of γ; and
(5) it is rational for the agent to ϕ for the sake of γ.

Where 'ϕ' stands for engaging in any 'large-scale transformative pursuit'.
A pursuit is large-scale and transformative iff:

(1) 'one cannot know beforehand' the full value of the pursuit; and

[40] See (Callard, 2016, p. 142) and (Callard, 2018, p. 88).

[41] Callard's (2018) comments on p.88-89 suggest that she believes that this is roughly equivalent to 'An agent has a proleptic grasp of her reason, γ, to ϕ iff…'. More on this later.

[42] Either because the agent has no idea what the value of γ is or because the 'knowledge' or understanding the agent has of the value of γ is incomplete, inchoate, anticipatory, indirect or deficient in some way. See (Callard, 2016, pp. 131–2) and (Callard, 2018, p. 72).

(2) the pursuit will 'require years of sustained effort, both in the form of preparation and in the form of the work attending the completed state.'[43]

With this more general account of proleptic rationality in hand, Callard's solution to the problem of self-creation is as elegant as it is ingenious. In brief, she explains the rationality of process of value acquisition by pointing to the future self as authoritative with regard to the present self. She argues that explaining the rationality of value-acquisition does require explaining a dependence relation between creator and created self, but that normative priority and temporal priority are reversed. We explain the rationality of the actions of our present self as licensed and justified by the endorsement of the self we will become. The future self takes *normative* priority even though it is temporally subsequent. Her claim is that it is not impossible that causal, temporal priority does not track normative priority and we should therefore accept that it does not (given its role in explaining self-creation) unless a good argument to the contrary is provided. In her words:

> When an agent acts on a proleptic reason, she takes steps toward bringing about the normative source of her present condition. She works to make herself grasp the reason (R2) that is, normatively speaking, the full or complete or nondefective variant of the proleptic reason causally driving that work (R2). Her grasp of R1 is causally responsible for her grasp of R2, though R2 is the source of normativity for R1— R2 is what makes R1 a good reason. My claim, then, is that in the case of aspiration, causal or temporal priority fails to track normative priority. Is this possible? I am inclined to think it must be, since it is actual. The theorist of aspiration gives us what is the only way to account for self-creation in the face of Strawson's dilemma; and self-creation is a fact of life. We ought, therefore, to accept that such a reversal is possible, unless someone can show otherwise. (Callard, 2018, pp. 213–214)

Her basic idea is to acknowledge that while a normative dependence relation does indeed exist between the aspirant's present and future self, the rationality of the actions of the aspiring, value-acquiring present self depends on the endorsement of their future, created self, rather than the other way around. The self whose endorsement counts is the future, created self: they take responsibility for the development of the earlier self who has been working their way towards this future self, guided by their developing sense of who they will become. The future self is therefore responsible for its own creation – it guided the past self towards it and thereby

[43] See (Callard, 2016, p. 133) and (Callard, 2018, p. 74).

rationalizes the actions of the past self that did not yet understand the reasons provided by the value they were attempting to acquire. We can therefore escape Strawson's dilemma, because the earlier self 'is not the lender but the borrower of normative authority' (Callard, 2018, p. 205).

3. Three problems with Callard's solution

3.1 Self-creationism

I think we can agree with Callard that this is possible and that a way out of Strawson's dilemma is desirable, but disagree that her solution works. In short, Callard's argument does not resolve the dilemma. It shows how the future self can be a source of value, but it does not show how it can be created by the agent. This is because creation is a *causal* relation, even though normative dependence perhaps need not be. Now, Callard rightly rejects backward causation – which she describes as 'absurd' (Callard, 2018, p. 213) – but the problem is then not whether we can wrap our heads around normative and temporal priority running in opposite directions (which I am very happy to grant *arguendo*), but whether we can accept that creation and causation can run in opposite directions. This seems much harder to swallow.

Strawson's dilemma is a problem for the theorist of self-creation because the temporally antecedent self must be the one doing the creating of the future self, if there is any self-creating happening in real time. The metaphor of being guided by features of a future self one does not yet fully understand is both interesting and suggestive, but whatever one's image of that self is and whatever work one performs in its service, the image is a function of the evaluative standpoint of the present self. It is not, because it cannot be (as Callard agrees), emanating from the future self backwards to the past. Of course, as we work towards that new self, we will also come to fill in its details, as we become better acquainted with the new values that comprise it. But if that is all Callard has in mind, there is nothing here with which Strawson need disagree. It is indeed often helpful to imagine ourselves as we wish we were, and then take steps towards becoming that person. But it is the imagination and the steps of the present self doing the work, as it must be, if we are to understand the process of self-creation as an agential activity taking place in the bounds of spacetime, at least as we presently understand them.

My diagnosis of why Callard's solution goes wrong is that it involves a subtle fallacy of equivocation. We started out by recognising that the problem of self-creation – understood as how one can change oneself *over time* into a genuinely new self and be the one responsible for it – is really the problem of how self-directed, novel rational value-acquisition is possible. We are then given a normative story about when it is appropriate to describe the acquisition of

a new value as a self-directed and agential activity. According to Callard, we now have a solution to the problem of self-creation because we have explained how one can rationally acquire new values. But the new problem is that we seem to have a story of rational value-acquisition without a satisfactory story of how self-creation is possible, at least as we understood it prior to its substitution for the idea of rational value-acquisition.

I am part of the process described in the normative story of my future self's creative prowess, but I do not *create* my future self, I simply *become* my future self.[44] If the initial concern was to explain how one can change oneself into a genuinely new self and be the one responsible for it, the answer is: you cannot. Your future self is relevantly responsible for its own features, you fall out of the picture.

What we really have is a story of how *rationalising* value-acquisition is possible. That is, of how we can make sense of and justify a temporally extended process of value acquisition after the fact. But although we are now in a position to see in virtue of what the process was rational, we are not in a position to see in virtue of what it amounted to self-creation, in the sense of one self (whether present or future) *creating* something new. Put differently, one can object that if *that* is what it means to rationally acquire a value, then self-creation is not, after all, equivalent to rationally acquiring a value. The original problem of self-creation with which we are left is, instead, that encapsulated by the thought motivating Strawson's dilemma: how can I create something genuinely new if all I have is myself to work on?

3.2 The solution embraces both horns of Strawson's dilemma

I have just explained that the first problem for Callard's proposed solution is that the causal story she tells is entirely orthodox and the problem of self-creation is, at least in part, a casual problem. A related problem is that the solution therefore, in effect, embraces both horns of Strawson's dilemma in reverse order. So, on Callard's account, an aspirant comes to realise through a chance encounter, with classical music, for example, that she does not understand, appreciate or apprehend its value. She is aware that others find it valuable and she hopes to develop an appreciation of that value she lacks currently. Gradually, she works on her inchoate grasp of the value in question, attempting to deepen her engagement with it, by attending music appreciation classes, symphonies, and so on, until, one day, she realises that she has come to appreciate the value of classical music in a much more robust, full-blooded way.

[44] The book's subtitle is 'The Agency of Becoming'. But, as it turns out, the agency of becoming is so atypical that it no longer ties that agency to a causal story about how I create the future self.

Now this is an excellent and subtle account of how an aspirant works themselves from a defective into a superior valuational state, and it does a nice job of explaining how the later self can appropriately be understood as a product of the labours of the earlier, aspiring self, but the fact remains that the impetus for the value acquisition was external and accidental and then, once this stimulus had done its job, the aspirant works away at cultivating this value, in light of her others.

So, the novelty of the value acquired is explained by externality or accident, while her responsibility for developing it is conventional cultivation. But if embracing either horn of Strawson's dilemma was supposed to be a problem, how is embracing both better? In my view, Callard's solution to the problem of self-creation, here understood as Strawson's dilemma, also fails on its own terms.

3.3 What is rational (dis)continuity?

I have just argued that Callard's solution to the problem of self-creation fails on its own terms. Now, I want to criticise those terms. Specifically, I want to question the idea of rational continuity structuring Callard's adaptation of Strawson's argument to the specific problem of self-creation.

In a nutshell, my objection is that the concepts of rational continuity and discontinuity involved in explaining Strawson's dilemma do not withstand scrutiny. Callard explains the rational continuity of value-acquisition as occurring when my new values are entailed by my old ones. But what concept of 'entailment' is at work here? Suppose we grant – as I think is plausible – that, as Michael Smith (1994) argues, systematic investigation and justification of our current evaluative standpoint could require the addition or elimination of some further value(s). It seems intuitive that we should count something like this as a paradigm of rational value acquisition, but of the sort that Callard labels cultivation.

But now note that, given the presence of conflicting values in our starting evaluative standpoint – as surely all of us must have – any values that follow rationally from some of them (and are therefore rationally continuous with those values) will also be rationally discontinuous with others. This delivers the strange result that the same value-acquisition could be both rationally continuous and rationally discontinuous with our current values. Perhaps we currently value music, drama, and glitter, but are unaware that coming to appreciate opera would be a natural consequence of thinking through our current values. But now suppose that our antecedent value set also includes the values of frugality and accessibility of the arts. Now opera does not look like a natural choice at all. Indeed, whether these values are consistent, as

they appear to be, but deliver inconsistent verdicts regarding the value of opera in this case or whether the opera case reveals a pre-existing tension in our evaluative standpoint, assessing the rational (dis)continuity of our value-acquisition looks like a very difficult task. Perhaps rational discontinuity with *one or more* of our starting values will be sufficient to generate the requisite novelty or, conversely, perhaps continuity with one or more of our values is enough to undermine it.

A plausible response to this criticism of the set-up of the problem of self-creation is that it is too *atomistic*. The real question is whether there is, to paraphrase Bernard Williams (1995), a sound deliberative route from the present self's evaluative standpoint to the acquisition of the new value. If there is, then the value is rationally continuous, in the relevant sense, with an agent's starting values. If there is none, then it is not. This sounds attractive. So, for Owen Wingrave to come to value a life of honour and tradition in the military would be for him to acquire a new value and thereby become a genuinely new self, precisely because there is at present no sound deliberative route from his current evaluative standpoint to that new one. Such a value is therefore rationally discontinuous with his current values and, if this value is acquired, Owen is a new self, the question then being whether it is also possible for Owen to have created this new self (given that its values are discontinuous with his).

But there is a significant problem with this otherwise attractive refinement of the problem of self-creation, now understood as arriving in a new valuational state in the absence of a sound deliberative route for getting there from one's starting point. The significant problem is that, as I see it, this construal of the problem of self-creation makes it genuinely impossible. This is because introducing the normatively loaded 'soundness' criterion seems to rule out the kind of responsible agential involvement required (by hypothesis) for genuine self-creation. That is, if there is no *sound* deliberative route, then Owen could only get there himself by accident. Now novelty of value really looks like it really will indicate that the creator self could not have been responsible for the creator self.

But there is a refinement available that, I think, delivers the right result. So far, we have construed the sound deliberative route synchronically and statically, as a matter of possibilities flowing from an agent's current evaluative standpoint. But our notion of possibility was quite limited, since it did not include the possibility that the evaluative standpoint itself could change as a result of deliberation. That is, we did not initially construe the sound deliberative route dynamically, diachronically and reflexively. But we must be open to the possibility that sound deliberation can involve changing one's mind about one's existing values, desires, preferences and commitments. In fact, if our concept of sound deliberation did not include the possibility

of changing our mind, then it would beg the question in the context of self-creation, since there could never be a sound deliberative route that led to genuine value change. But if sound deliberative routes can include changing one's mind about features of one's starting evaluative standpoint, then a sound deliberative route can take you anywhere. Is this tantamount to depriving it of any useful limiting role whatsoever?

I think not. For we have now isolated what I think is a plausible criterion for genuinely acquiring a new value in such a way as to constitute an act of self-creation. The criterion is that there is no *direct* sound deliberative route from one's current evaluative standpoint to the value in question. If there were, the value would be acquired *via* cultivation. But where there is an *indirect* sound deliberative route to acquiring the new value, this constitutes an act of self-creation, where an indirect sound deliberative route just is a sound deliberative route that requires revision of one's starting evaluative standpoint. That is, what is really required is that acquiring that value required one to *change one's mind* about one's former evaluative standpoint. As will become clear in Part II, I intend this to be understood quite literally. To preface: we sometimes acquire new values that integrate seamlessly into our existing evaluative standpoint, presenting as refinements of it or pre-existing commitments we are now better able to articulate and identify or as the entailments, natural concomitants or complements of views we already hold. The new values, that is, can be simply *added* to one's current evaluative standpoint, without doing violence to the extant set. I do not deny that this sometimes means that we have acquired new values but, like Callard, Strawson, and others, I agree that this is not an act of *self-creation*. But where the acquisition of the new value involves *re-structuring* of our former evaluative standpoint to accommodate the new value, we are engaged in self-creation.

PART II: VALUE ACQUISITION

In this Part, I will do three things. First, I will explain what it is to acquire a value. Second, I will argue that acquiring values is not usually a matter of deductive reasoning, and so we need not be troubled by rational continuity or discontinuity understood in terms of entailment. Third, I will argue that even if the values acquired are entailed by some of the agent's core values, this does not preclude acquiring them constituting an act of self-creation.

4. What is it to acquire value?

If self-creation is about self-directed value-acquisition over time, what is it to *acquire* a value? I propose to understand this idea attitudinally and dispositionally. Valuing, on this account, is a complex dispositional bundle: one is disposed to believe, feel, desire, and act in accordance with the value in question, such that the value becomes a fixed point in one's reasoning, both explicitly and (often more importantly) implicitly. To come to acquire a value is therefore to have one's agency *re-structured.* So, what it is to acquire the value of loyalty is to be disposed to be loyal where fitting, to approve of loyalty in others and disapprove of its absence. This entails, moreover, that considerations of loyalty will feature in your practical deliberations and, for example, in the advice you give to others. That is, acquiring the value of loyalty is to come to have, variously, beliefs, desire, attitudes – and be disposed to have like attitudes – that reflect the value of loyalty. Note that these attitudes need not be – and should not always be – *de dicto*, i.e. the value of loyalty should not necessarily be front of mind under that description in your motivations when behaving, thinking, feeling or desiring in a loyal way. This is a variation on Bernard Williams' (1976) 'one thought too many'. If asked for a justification for why you did not backstab your friend to gain a promotion, for example, a loyal person may simply respond 'Because they are my friend', without more. The value of loyalty need not feature in their explicit deliberation or justification at all. Note, moreover, that often ideas inconsistent with an agent's values will not occur to them. For example, in thinking about how to secure a sought-after promotion, many of us would consider putting in extra hours or attempting to demonstrate our value and abilities. Most of us would not consider murdering our direct superior to create the relevant vacancy. That is, we do not have to consider whether we really ought to murder our colleague and then, upon reflection, *deduce* that we ought not, since we disvalue murder. Instead, the murder-my-colleague option does not appear on the deliberative smorgasbord. Our values often structure our deliberation in this way without our needing to reason by directing our attention towards them and deducing what follows. They are the vast implicit background that structures our explicit deliberations about what to do: values *dispose* us to deliberate in various ways, just as much as they feature explicitly in our thoughts.

It is important for self-creation that what it is to acquire a value is therefore to develop a disposition to respond to the world and act in new ways. The reason this is important is that acquiring a new value involves a structural change in the internal composition of an agent, akin to the introduction of a new force in a physical system or, perhaps, a new fixed point that changes the shape of the field of forces in its vicinity. This is the evaluative equivalent of the view of beliefs according to which they should be understood in terms of their inferential

patterns rather than their atomic propositional content alone. To acquire a value, then, is to come to be disposed to think, feel and act in new ways that reflect the content of the value in question.[45]

5. We do not usually acquire values *via* deduction

If coming to value something is a matter of having a complex bundle of dispositions reflecting the value in question, this helps explain my next claim, namely, that we also do not usually acquire values *via* deduction either. I think this claim is independently plausible. The phenomenology of coming to value something does not, at least usually, feel like deducing that it is valuable. This is likely because, as Samuel Scheffler (2010, pp. 27–28) and others have argued, values are not beliefs.[46] We can believe that many things are valuable, but because there are many more valuable things out there than we are capable of valuing in our short lives, we will not actually *value* all of them. Often, we come to acquire a value by encountering a particular instantiation of it or realizing that the value in question explains the pleasure we take in some activity or performance or the respect we have for some person, for example.

That is, the phenomenology of much value-acquisition feels more akin to *inference to the best explanation*. The relevant value is posited by the agent as an inference to the best explanation of something they realise requires explanation: an experience of the good that typically attends some positive personal, cultural, artistic or social experience and that is in some way unfamiliar, unexpected or otherwise novel. Once posited, if the feeling in question is sufficiently strong; the value in question best explains it; and acquiring the value requires revising the agent's current evaluative standpoint; then that is precisely what the agent will do.

As far as I can tell, we do this often enough: when we come to realise what everyone's on about with New Wave Cinema or how indirect discrimination works or what self-sacrifice means. But this is just to say that we often acquire new values that require us to modify our existing tendencies and dispositions to reason in accordance with our previously held values. We have come to value New Wave Cinema when we are disposed to attend screenings of it,

---

[45] This account of value-acquisition just sketched shares affinities with virtue theoretic accounts of value. But it is wider in scope, since coming to value various objects or practices – art, sport, fashion, food, for example – need not lead to paradigmatically *virtuous* dispositions. Rather, it will lead to dispositions that reflect evaluative engagement with the object or practice: promoting, respecting, participating-in, enthusing-about, and generally evincing (and evidencing) one's cognitive and affective acquisition of the value in question. Cf. (Arpaly, 2003).

[46] Cf. (Bratman, 2000); (Frankfurt, 1971; 1988); (Kolodny, 2003); (Korsgaard, 1986; 1996; 2009); (Langton, 2007); and (Wallace, 2013).

discuss it with like-minded enthusiasts, count it as a reason to attend one film festival instead of another, enjoy and understand its films and so on.

Note also that while acquiring a value will mean altering one's patterns of slow, deliberate reasoning when explicitly attending to the relevant value, it will also change our more intuitive or reactive agency as well. Someone who has come to appreciate the value of art is far more likely to risk injury saving a precious artwork from a fire in an emergency, even if they do not carefully "decide" to do so before rushing into the burning building and *mutatis mutandis* for a cultural artefact for someone who has come to appreciate the value of cultural heritage; or a nonhuman animal, for someone who has come to genuinely value nonhuman animals' lives. But this just is to recognise that we sometimes engage in agential, rational value-acquisition in such a way that we also alter the deliberative patterns our more intuitive agency then manifests. That is, we sometimes really do change ourselves for the better.

If we do not acquire values by deducing them, it is difficult to see why the test for whether they are genuinely new should be understood as whether or not they are entailed by an agent's existing values. This is not to say that we *never* acquire values by realising that, given what we presently value, we are logically required to value something else. Philosophers, for example, are prone to this sort of evaluative world-building. This is also the sort of argument we might use to convince others to acquire a value we think they should share with us but do not.

Callard is entitled to respond that there has been a sleight of hand in my argument. The rational-continuity test envisioned in Strawson's dilemma, she might argue, is not a matter of *decision-procedure*. Instead, rational continuity is matter of what is or is not entailed *as a matter of fact*. I have slipped from metaphysics to decision-procedure and back again, so it is I who am guilty of equivocating to make my argument.

But now note something interesting. Suppose we encounter someone deeply mistaken about the entailments of their present values who is, consequently, prone to all kinds of faulty and *immoral* reasoning. They do not realise that, say, because they care deeply about their future self, they should also care about others (Nagel, 1970). Instead, they constantly treat others as means rather than ends; violate others' personal dignity; lie, cheat, steal; and are generally no fun at all to be around. Now suppose that after a period of sustained reflection they come to realise that what their core value(s) – namely, concern for their own future welfare – *really* requires, that is, entails, is that they have this same concern, or something approaching it, for others. They undergo a radical transformation: instead of now spending their weekends ripping-off pensioners and puppy-farming, they do good works for the needy. They renounce

their life of crime and perfidy, settle down, and follow their *true* vocation, namely, sweeping for landmines in former conflict zones. It seems here that we have a radical personal transformation. Indeed, it is difficult to imagine one more radical. It is also the product of the agent's deliberation concerning what their core values *really* require, that is, entail.

This case is a paradigm instance of self-creation of the very best kind. The agent is radically different from how they were before and brought about these changes in themselves as a result of personal reflection on what follows from their most important core values. But if we are to subscribe to an account of self-creation according to which rational continuity precludes genuine self-creation, then we must concede that all this agent has really done is work out the entailments of who they were already. That is, they have *cultivated* themselves without creating anything new. In fact, accepting this view of what self-creation requires has the consequence that according to a number of plausible accounts of morality, becoming something approaching a moral saint is already entailed by our starting point values, regardless of which values we happen to have: so, it is impossible to genuinely *create* a better version of yourself, no matter how bad your starting point and how good your end point.[47] All you can do is realise that, deep down, you already were a moral saint, you just did not act like one.[48] This cannot be right.

6. The possibility of self-creation (or, what happened to the dilemma?)

Allow me to summarise the conclusions of this Part. So far, I have argued that acquiring a value is a matter of coming to have a complex dispositional bundle one did not have previously. I have also suggested that we do not usually acquire values by deducing them from our other values, but that even when we do, if the change is sufficiently radical, it is an act of self-creation. That is, rational continuity is not really the test of whether an act or acts of self-creation have occurred. This requires rejecting the terms of Strawson's dilemma, at least as it presented by Callard. Instead, the test of whether an act of self-creation has occurred is that advocated in Section 3, namely, whether the acquired new value(s) were available only *via* an *indirect* sound deliberative route, such that acquisition of the value(s) requires *revising* the agent's starting evaluative standpoint.

---

[47] Cf., for example, (Korsgaard, 1986; 1996; 2009); (Nagel, 1970); and (Schroeder, 2007).

[48] Cf. (Wolf, 1982).

This resolves Strawson's dilemma for self-creation. But it does not entirely diffuse the challenge that motivates it. There are at least two reasons for this. One I have already bracketed. The second will be the topic of Part III. The bracketed reason is that we have made no headway on the free will problem or the question of ultimate moral responsibility. Nothing I have said in favour of local self-creation will support an argument to the conclusion that we can be ultimately responsible for our character or entirely the author of our own self. At the outset, I said we would have to leave that question for another day. All I wanted was to establish the more circumspect claim that we really can change ourselves, sometimes hopefully for the better.

But the second reason the motivation for Strawson's dilemma remains troubling is as follows. I agree with Callard that often the impetus for radical agential change is *external*. It does not take place in a moment of sober reflection, but rather gets going when we witness something tragic or beautiful; hurt or come to love someone; or are smacked with the sheer complexity, improbability, enormity, oddity or significance of some series of events, whether great or small. Such catalysing experiences can have an accidental or passive character to them, such that it seems inappropriate to label all the changes that follow agential acts of self-creation, of even a local kind. Unless the relevant distinction between accidental or passive changes *versus* those of a sufficiently agential kind to constitute self-creation can be drawn, something like Strawson's dilemma might yet carry the day against the would-be theorist of self-creation.

PART III: INDIRECT EVALUATIVE VOLUNTARISM

In this Part, I explain my alterative account of self-creation, which I label 'indirect evaluative voluntarism'. I begin by distinguishing between two different ways of reacting to an experience and argue that while one is a good candidate for self-creation, the other is not. In the final section, I temper this exclusion, and conclude that the prospects for self-creation are bright.

7. Acting and reacting

The changes wrought in us by some experiences are transformative, but passive: even this natural phrasing supports the idea that some changes simply *happen to us*. Sometimes we *react* to witnessing or experiencing tragic, joyous, beautiful, painful, or unsettling events in ways that seem entirely out of our control: certain emotions, beliefs, values or conclusions are simply "forced" on us by the experience – or so the phenomenology suggests – and there is little we

can do about it. We are changed as a result, but not through our agency.[49] These are not good candidates for self-creation.

But other events, though momentous and subsequently identifiable as the catalyst for great changes in who we are, do not *determine* our reaction in the same way.[50] We are left with the feeling that we need to make sense of what has happened; or that something calls for explanation or justification. That is, while the event will clearly change us, *how* it will change us feels as yet undetermined. We are, at least initially, *unsure* what to make of the novel experience. There is, consequently, *work* to do. While the passive transformative experience seems to occur *via* something close to revelation; this more active version involves the agent *evaluating* and attempting to understand the new data in light of their existing values. Sometimes, this process of evaluation and understanding requires the agent in question to posit a new value and *revise* the evaluative standpoint they had in order to accommodate it. When that happens, consistent with the criterion developed in Section Three, the agent is working on themselves to create something new. They have arrived at a new value *via* an indirect sound deliberative route. That is, *that* some change is to be explained in part by external stimulus does not entail that the change cannot amount to an act of self-creation. In other words, it is not just what happens to you, but what you make of it, that counts for assessing the possibility of self-creation. How to respond to an entirely new aesthetic experience, or an unthinkable breach of trust, or the unexpected and untimely loss of a close family member is not usually *fixed* by the values we already have. There are many different ways we might go, and deciding which – and who we might become as a result – is not something that simply happens to us.

More encouragingly still for the would-be theorist of self-creation, not all *passive* transformative experiences preclude an element of self-creation either. Sometimes, we put in agential work of the right kind *prior* to the passive transformation's taking place. In the final section, I provide an account of when this amounts to self-creation, and use it to complete my account.

8. Indirect evaluative voluntarism

Lurking behind Strawson's dilemma are, I think, two implicit theses we would do well to reject. The first involves bundling together the evaluative equivalents of direct and indirect doxastic

[49] Compare (Fischer, 2006)'s conclusion that it is how one plays the cards one is dealt that counts.

[50] There are clear links to this evaluative permissivism and epistemic permissivism generally. I say more about the ethics of belief in Section 8. Cf. (Horowitz, 2019) and (Dandelet, 2021a).

voluntarism and concluding that, because the former is implausible, the latter must be too. I provide an argument to show that something like indirect doxastic voluntarism is just as plausible in cases of evaluation as it is in assessing responsibility for our non-evaluative beliefs, which lays the groundwork for a unified account of doxastic responsibility.

I will understand direct doxastic voluntarism as, very roughly, the view that belief is directly within an agent's control. That is, we can simply *decide* what to believe and can be held responsible for our beliefs accordingly. It is widely considered to be implausible and, in my view, rightly so.[51] But indirect doxastic voluntarism, *viz.* the view that an agent's beliefs are *indirectly* under their control, is much more plausible.[52] According to indirect doxastic voluntarism, explaining an agent's responsibility for their beliefs is a matter of assessing their compliance with epistemic norms. So, agents are responsible for gathering, weighing and believing in accordance with the evidence, for example. To the extent that they have complied with the relevant epistemic norms, they are epistemically praiseworthy and to the extent that they have violated them, blameworthy.

One reason self-creation – understood to involve an agent's being *responsible* for acquiring a value – seems impossible is because it appears to require the evaluative equivalent of direct doxastic voluntarism, which I will label 'direct evaluative voluntarism' though note I do not thereby mean to conflate it with voluntarist theories of normativity. Direct evaluative voluntarism would be the view that our values are just "up to us". That is, that what we value is directly under our control such that we can pick-up and drop values as we see fit. Some crude caricatures of, variously, postmodernism, existentialism and subjectivism equate them with or take them to entail something like direct evaluative voluntarism. But direct evaluative voluntarism is, in my view, as implausible as direct doxastic voluntarism. It certainly does not match the phenomenology of our evaluative lives. But *indirect* evaluative voluntarism, now understood as the evaluative analogue of indirect doxastic voluntarism, is much more promising. The basic idea is that we are responsible for acquiring values – even those that we acquire as a result, in part, of external stimulus – when we acquire them in the right way.

---

[51] The *locus classicus* of the ethics of belief debate is the eponymous paper by William Clifford, usually read with William James' commentary on it. Cf. (Clifford, 1877) and (James, 1896). A very helpful overview of the literature can be found in (Chignell, 2018).

[52] Cf. (Adler, 2005); (Audi, 2001; 2008); (Conee & Feldman, 1985); (Feldman, 2000); (Hieronymi, 2006; 2008); (Leon, 2002); (Pascal, 1670); and (Yee, 2002).

Sometimes, we set out with the intention of exposing ourselves to new values, evaluative perspectives or ways of life. This establishes a nexus between our agency and any value we acquire as a result. Note, moreover, that if the acquisition meets the indirect sound deliberative route criterion, we have thereby brought about a change in ourselves that amounts to an *act* of self-creation. So, if you travel to a country where another religion or radically different way of life prevails with the intention of coming to understand the value of that way of life better *and thereby come to acquire new values*, you are appropriately held responsible for having created yourself as a result.

This is obviously not to say that all novel value acquisition that meets the indirect sound deliberative route criterion, no matter how accidental or external, will nevertheless count as self-creation. Far from it. Often, the rupture with your previous evaluative standpoint is clearly the result of chance or externally engineered exposure to new values, perspectives or ways of life, in the way Strawson's dilemma correctly rules out as too external to count as genuine self-creation. If, as a youth, I am dragged kicking and screaming to the opera by my parents every month for a year, then one day – despite my resistance – hear an aria that blows my mind, my subsequent love of opera and rejection of rock music is not plausibly construed as an act of self-creation. The question is whether the circumstances surrounding the value-acquisition included your intention to expose yourself to the new value, perspective or way of life in order to come to understand it better. If you did, I think it makes sense to hold you responsible accordingly.

An analogy is with activities like getting a sun-tan or going to see the northern lights. If you happen to get a suntan or see the northern lights on the way to the shop to get a pint of milk, it does not really make sense to hold you responsible for the achievement: sure, going to the shop is agential, so in once sense the suntan or seeing the northern lights are a product of your agency. But they are a *by-product*. But if you get out into the appropriate climactic conditions – with your lotion or binoculars, respectively, and presumably not in the same part of the world – with the express *intention* of tanning or seeing the northern lights, and you succeed, you are responsible for the achievement, and this is true *even if* you do not know what they look like in advance.[53] This is despite the fact that, for example, when you *get* a tan, you

[53] This allows us to answer Anscombe's 'Why?' without needing to rationalise the process of value-acquisition by reference to the content of the value after the fact. That is, we need not (and could not) *specify* the content of the values to be acquired in advance, but this does not prevent their acquisition being agential. (Anscombe, 1963). Cf. (Callard, 2018, p. 70–71).

don't actually have to *do* anything other than put yourself in the right place at the right time (since it is the sun and your cells *doing* all the work).

The second implicit thesis lurking behind Strawson's dilemma is a bright-line distinction between evaluation and value-acquisition, where one can be responsible for the former but not the latter. The model of value-acquisition I favour suggests it is akin to *evaluating* new values. What else could our encounters with new values be? We come to see there is or might be a new value in the vicinity – one that best explains the experience we are having – we reflect on its character and we appraise it in light of the values we already have. But if that is true, then it is not different in kind than appraising the value of anything else in light of the values we already have. But that sort of appraisal is a *paradigmatically* agential activity, and one for which we hold agents responsible all the time. So, insofar as we can be responsible for creative, evaluative interpretations of, for example, new artworks, actions or people we encounter in our lives, so too can we be responsible for the acquisition of new values we encounter as we live them.

Note that a final and peculiar corollary of accepting the picture of value-acquisition implicit in Strawson's dilemma is that we can really only be responsible for self-creation if it follows from an evaluative *mistake*. For if I mistakenly think that some new value is rationally continuous with my current values and acquire it accordingly, I will be responsible for it, since I have acquired it agentially (by deploying my deliberative capacities) but it is also rationally *discontinuous* with my previous values, and so not caught by either horn of the dilemma. But if I reason correctly from my current values, I can never be responsible for any self-creation, because any agential value-acquisition indicates rational continuity and therefore pre-existing implicit endorsement. This evaluative responsibility asymmetry is odd. We should reject it in favour of indirect evaluative voluntarism.

Conclusion

In the first half of this paper, I argued that Callard's solution to the problem of self-creation, understood as Strawson's dilemma, fails. In the second half, I provided some reasons for thinking that the dilemma was ill-conceived. Finally, I argued that indirect evaluative voluntarism is both independently plausible and explains how self-creation is possible. Of course, none of this makes self-creation any *easier* or provides a how-to manual. But at least we can be optimistic that attempting to do so is not *guaranteed* to be a waste of time.

CHAPTER 4: SOCIAL NORMATIVITY: FORMERLY, MERELY FORMAL NORMATIVITY

Abstract

This paper defends two claims. First, I argue that you *always* have some reason to comply with the social norms applicable to your situation, no matter how immoral or ridiculous. This is not a baroque, technical fact about social norms: it is fundamental to understanding what they are, how they work, and how they can be a source of grave injustice. Second, I argue that this fact about social normativity requires re-thinking the distinction between formal and robust normativity. Social norms are not robustly normative. But nor are they merely formally normative. Since these two categories are widely assumed to be exhaustive, this requires revising our normative taxonomy.

1. Introduction

Contrarians are fond of reminding "sheeple" that many social norms are arbitrary, stupid, inefficient or immoral. Unsurprisingly, this sentiment is echoed by many philosophers working in ethics, philosophy of social science, political theory and jurisprudence. Accordingly, 'one of the central puzzles of social theory [is] that of explaining why people comply with social norms' (Anderson, 2000, p. 170). Explanations of why we comply with social norms often focus on internalization mechanisms and provide rational reconstructions of social norms' capacity to provide *motivating* reasons (Bicchieri, 2006). In brief, social norms alter our behavior because we *care* about them and take them to be normative for us. We also care about social norms because our behavior in their light affects what others think about us. Enter rational choice theory: most of us desire esteem for its own sake *and* because it is useful in realizing other desires. So, there are often *instrumental* reasons to do as social norms require (Brennan and Pettit, 2006). Of course, this feature of social norms can be leveraged for good or ill, just as social norms themselves can be good or ill. But there are limits to social norms' power. After all, some transgressions go unnoticed, unpunished or ignored (Langton, 2009). Some transgressions are even applauded. And some people just don't really care what others think. So, nothing *necessary* or *universal*, it is thought, can be said about social norms' power to generate reasons. There is no magic formula the decision theorist can hand the deliberating agent who must work out whether to comply with a given social norm.

Most metaethicists working on social norms agree. Unlike, say, the norms of morality, epistemology or prudence, there is nothing *general* we can say about the reason-entailing

power of a social norm *qua* social norm.[54] In contemporary metaethical jargon, social norms are not *robustly* normative.[55] The distinction between robust and "merely" formal normativity is often marked deploying something like the following condition. A given norm is *robustly* normative iff it satisfies:

> **Robust-reason-entailment** Necessarily, if N is a robust norm requiring an agent A to perform an action ø in circumstances C, then A has a reason to ø in C.

Social norms, on the other hand, are *the* paradigm example of merely formal normativity. That is, it is taken to be *obvious* and *uncontroversial* that you do not *always* have a reason to do as social norms require. For example, in a recent survey of the literature on normativity, Stephen Finlay notes that:

> This distinction is marked with various terminology; here I'll adopt the relatively theory-neutral labels of *formal* versus *robust* normativity$_{ont}$." "Mere" formal normativity$_{ont}$ is said to be ubiquitous (found in law, etiquette, games, shopping lists, and more), naturalistic, and nonsubjective, but also of little philosophical interest. Robust normativity$_{ont}$ is said to be the important kind of normativity$_{ont}$ of special interest to ethics, epistemology, and other branches of "normative" philosophy." (Finlay 2019, p. 205).

He cites (McPherson 2011), (Parfit 2011), (Joyce 2001), (Copp 2007) and (Broome 2013) as also deploying the distinction, all of whom seem to regard social norms as instances of merely formal normativity. To these we might add (Baker, 2018), (Berman, 2019), (McPherson, 2018) and (Southwood, 2019).

Likewise, the best metaethical accounts directly focused on social normativity also argue for its formality. Nicholas Southwood, for example, thinks it is 'manifestly false' that all social norms give the people subject to them reason to do as they require. He adds:

> To be sure, we may *often* have reason to comply with conventional norms. The point is just that this is not *necessarily* so. Whether we ought or have reason to comply with conventional norms depends, inter alia, on their particular content

[54] I employ the slightly awkward 'reason-entailing' here rather than the more natural 'reason-giving' to avoid terminological confusion with David Enoch's influential discussions of reason-giving in (Enoch, 2011a; 2011b; and 2014), which I discuss in more detail in section 2.

[55] See (Baker, 2018), (Berman, 2019), (Enoch, 2019), (Finlay, 2019), (McPherson, 2018), (Southwood, 2019) and (Woods, 2018).

> and the particular circumstances in which agents find themselves. (Southwood 2019, p. 37, his emphasis)[56]

Likewise, Kate Manne observes:

> Of course, *nobody* would want to hold that every social practice—however cruel or pointless—generates practical reasons… (Manne, 2013, p. 52, my emphasis)[57]

This, then, is the orthodox view. One might encapsulate it in the slogan: 'You do not necessarily have any reason whatsoever to do what social norms require'. The aim of this paper is to show that the orthodox view is false. On the contrary, a thesis analogous to Robust Reason-entailment is true for the social domain:[58]

> **Social-reason-entailment** Necessarily, if N is a social norm requiring a human agent A to perform an action ø in circumstances C, then A has a reason to ø in C.[59]

In fact, establishing that social norms *are* necessarily reason-entailing is a very useful starting point for systematic investigation into their proper role in decision-making; evaluation; and our understanding of normativity generally. The argumentative strategy for establishing that conclusion is analogous to one that has delivered promising results in epistemology (Kornblith, 2002; Dyke, 2020), metaethics (Schroeder, 2007) and metasemantics (Cappelen, 2018): show that the relevant instrumental normativity is simply contingent on an agent's having any desires

---

[56] Context indicates Southwood is using the term 'conventional' in this quote in a way that includes social norms more generally and *not* in the sense distinguished in (Eriksson & Southwood, 2011). See (Southwood, 2019, p. 25).

[57] Of course, one might construe the authors I cite in section 2 (for example, Enoch, Southwood, Manne etc) as denying that social norms always and necessarily give you a *pro toto* reason for action (as opposed to a *pro tanto* reason for action). But this strikes me as implausible, since no one would or should find this this maximalist view of social normativity remotely appealing. Besides, one might add, the dualism of practical reason suggests that even the strongest imaginable prudential reasons are only ever *pro tanto*, since they must be weighed against those of morality. And, of course, unless you endorse are a particularly stringent form of moral rationalism, the same will be true of moral norms.

[58] To be clear, I do not intend to suggest that there are special sorts of reasons – "social reasons" – entailed by social norms. I adopt this language to echo that of Robust-reason-entailment. The reasons entailed by social norms are instrumental prudential reasons.

[59] Pro tanto, not pro toto (Broome, 2013).

whatsoever, on the model of a universal hypothetical imperative.[60] This approach locates the normativity of social norms in the interplay between collective response-dependence and individual agency. Its significance is that it can explain how that collective response-dependence can generate reasons for individual agents *independent* of what they happen to desire or believe. In the final part of the paper, I argue that social norms therefore possess quasi-robust normativity, and sketch how recognizing this intermediate category might illuminate other interesting questions in normative theorizing.[61]

2. The orthodox view

The orthodox view, then – that social norms do not necessarily give you a reason to do as they require – is as intuitive as Social-reason-entailment sounds implausible. There are many immoral, inefficient or otherwise irrational social norms that benefit no one or benefit some but not others also subject to them. This follows straightforwardly from the fact that social norms are a matter of social fact rather than the immutable deliverances of Reason. That is, social norms are a function of the attitudes of the members of the group in which they obtain. My preferred account of social norms – on which I will rely in making my argument in this paper – is the 'norms-as-normative-attitudes' account developed by Geoffrey Brennan, Lina Eriksson, Robert Goodin, and Nicholas Southwood. According to it, we should think about social norms as clusters of normative attitudes, together with beliefs regarding the existence of those attitudes. Accordingly:

A normative principle P is a norm within a group G if and only if:

(i) A significant proportion of the members of G have P-corresponding normative attitudes;[62] and

[60] See also (Dyke, 2020) for an intersubjective elaboration of the epistemic instrumental strategy, with interesting parallels to the intersubjective approach to social normativity in (Manne, 2013).

[61] Echoing (Blackburn, 1993), though strictly speaking, I will be developing an account of what one might label 'anthropocentric normativity' which is itself is a sub-category of quasi-robust normativity more generally.

[62] How 'significant' must the proportion in question be? (Brennan et al., 2013) describe this as a 'vague matter' but note that this 'is a good thing, for it is also often a vague matter whether or not a given normative principle is a norm of a group.' (Brennan et al., 2013, p. 31).

(ii) A significant proportion of the members of G know that a significant proportion of the members of G have P-corresponding attitudes.[63] (Brennan et al. 2013, p. 29)

According to the authors:

> normative attitudes include at least the following: (a) normative beliefs, judgements, and other cognitive states, (b) normative expectations, (c) reactive attitudes and dispositions to have such attitudes, and (d) any other attitudes that entail (a), (b), or (c). (Brennan et al. 2013, p. 29)

What, if anything, *follows* from accepting something like a norms-as-normative-attitudes account of the *nature* of social norms with regard to their reason-entailing power?[64] In a recent paper, Laura Valentini suggests the following:

> • P1: One has a pro tanto obligation to respect people's genuine commitments (i.e., "to give *agency respect* to people"), provided those commitments are morally permissible.
>
> • P2: If one does not respect the socially constructed norms that exist in a given context, one fails to respect the commitments of norm-supporters.
>
> • Conclusion: One has a pro tanto obligation to respect the socially constructed norms that exist in a given context, provided the commitments of norm-supporters are genuine and morally permissible. (Valentini, 2019, p. 6)

Accordingly, Valentini argues, social norms possess 'moral force'. She contrasts her own account of the moral normativity with 'the deflationary view' according to which 'the fact that certain actions are required or prohibited by socially constructed norms, per se, does not affect their moral status' (Valentini, 2019, p. 4). Suppose Valentini is correct, is her view *inconsistent* with the orthodox view, according to which social norms do not entail reasons for action?

No, in fact, it represents an important elucidation of one aspect of the orthodox view. To see why, consider the following. The strongest motivation to hold both the orthodox and deflationary views is the incontrovertible existence of immoral social norms. Valentini's account shows that the orthodox view and the deflationary view can come apart, precisely because the fact that social norms do not *always* provide one with moral reasons does not mean

---

[63] (Brennan et al., 2013) note that they are drawing on HLA Hart's famous discussion of the distinction between norms and mere habits (Brennan et al., 2013, pp. 28-29). Cf. (Hart, 1961, pp. 55-56). Cf. (Pettit, 2019).

[64] I consider how my argument would be affected by adopting an alternative definition of social norms in section 3.

they are devoid of a distinctive moral normativity.[65] But because one must inquire into the moral permissibility of the commitments underpinning the norm, whether it gives one a moral reason to comply will still depend on the norm's 'particular content and the particular circumstances in which agents find themselves' (Southwood 2019, p. 37). Valentini's is the orthodox view, updated.[66]

Further support for something like the orthodox view is to be found in David Enoch's influential work on the normativity of law (2011a; 2011b; and 2014).[67] According to Enoch, the law provides reasons for action (when it does) by manipulating the non-normative circumstances of agents in ways that trigger pre-existing conditional normative reasons. But the law does not *always* trigger reasons, moreover:

> once we see that we're in the territory of contingent triggering reason-giving, we see that there is no reason to expect any uniformity in the explanation of how it is that the law gives (that is, triggers) the reasons it does in fact give. *Had it been a necessary truth about law that it gives reasons, perhaps a uniform account of law's reason-giving force would have been called for.* (Enoch, 2011b, pp. 27-28, *emphasis added, footnote omitted*)[68]

I understand my project to be taking up Enoch's counterfactual suggestion italicized in the quote above, but with regard to *social norms* generally rather than law specifically.[69] That is, I am attempting to show that it is a necessary truth about social norms that they entail reasons and I aim to provide a uniform account of how.

---

[65] On this point, compare (Manne, 2013), who argues that only *valid* social practices can be sources of practical reasons, where validity requires that "they are conducive to human flourishing at large" and satisfy both consequentialist and rights-based constraints, *viz*., that the practices do not cause suffering and do not trample the rights of anyone subject to them (Manne, 2013, p. 69).

[66] Cf. (Woods, 2018) who also argues that we have respect-based reasons to comply with formal norms generally. Fortunately, just as it is possible to separate the orthodox view from the deflationary view, it is also possible to hold Valentini's view regarding the *moral* force of social norms while rejecting the orthodox view of their normativity generally. In fact, I think Valentini's attractive account of the moral force of social norms could be combined with the more general theory of social normativity developed in this paper. This is, in my opinion, a welcome result.

[67] Cf. (Schroeder, 2004; 2005).

[68] While I have no objection to Enoch's triggering account in broad outline, in the final section of this paper, I will suggest a friendly amendment required to accommodate the intermediate normativity of social norms.

[69] Note that Brennan et al. argue that laws are social norms according to their definition of them, but where the relevant normative attitudes are *de re* rather than *de dicto (*Brennan et al., 2013, pp. 46-50).

Summing-up: so far, I have suggested that many philosophers, reasoning from the obvious response-dependence of social normativity, draw the conclusion that it is a ragbag, about which very little can be said that is both general and interesting.[70] In the next section, I combine a plausible story about human agency with the norms-as-normative-attitudes account to deliver the beginnings of a universal account of social normativity for creatures like us. The basic structure of the argument is one establishing a claim of *a posteriori* necessity regarding social norms' reason-entailing power. It follows from what social norms *are* and what being a cognitively limited, socially situated agent involves, that social norms always entail genuine reasons for action.

3. Norm violation is risky for creatures like us

In Australia – as, I imagine in many countries – there exists a (widely disregarded in practice, one suspects), social norm against picking one's nose in public. There is also a social norm requiring one to face the doors when riding in a lift. Entering a lift and facing one's fellow passengers *while* picking one's nose would be a double-violation and frowned upon by most. Even if people were too embarrassed to say anything, they would disapprove, gossip about you subsequently and so on. Even if they have no idea who you are, they will of course still disapprove (perhaps more so) and *if* they encounter you again in the future, they may well remember you. *Not* in a good way and with predictable damage to social capital in the relevant setting. This risk could materialise sooner rather than later. Perhaps you fall over and break your leg as you exit the lift and need help. Perhaps one of the people in the lift show up on a job panel next month. Perhaps you're in the presence of future father-in-law. You just *never* know, and that's the point.[71] Likely, this has even happened to you in some – hopefully mild – form. Being a social creature has its advantages, but it has some pitfalls too.

So, there is always some chance we will require the cooperation and goodwill of other people and, when whenever we violate a social norm, we run the risk of reducing the esteem

[70] Of course, if Valentini is correct, then her account boosts the systematicity of the orthodox account considerably by delivering a form of 'content-independence' without 'content-insensitivity' (Valentini, 2018). But whether a given social norm gives those subject to it *any* reason to comply still depends on *what* it requires, and this is the orthodox view (Southwood, 2019). On my account, contrary to the orthodox view, the existence of a reason to comply with a social norm is *content-insensitive*.

[71] This uncertainty pertains not just to the possibility of being observed but also to how observation in this context could be relevant to the realization of our ends in the future. Cf. (White, 2020).

required to secure that co-operation.[72] In game-theoretic parlance, it's a mistake to assume your interactions with others are one-shot if they might be repeated. Of course, the probabilities involved regarding future detrimental consequence on any one occasion of norm violation are slim indeed. Likely you are thinking that the probabilities are sufficiently low that you do better in the long run to be a strategic norm-violator over time and particularly when the cost of compliance is high and the risks low and *of course* all this is true. Sometimes the balance of reasons favours social norm violation. But this is entirely consistent with the (surprising) truth of Social-reason-entailment. This has a number of important practical and theoretical implications, to be explored below. But one immediate implication appears to be that all claims of the form 'Sure, N is a social norm, but what's that to me? The fact that N happens to be a social norm around here is *no reason whatsoever* for me to do as it requires' are false. And that's the sort of claim one hears often enough both inside and outside philosophy (Enoch, 2019). The falsity of these sorts of propositions is theoretically and practically significant even if, on principle of charity, we understood them as licensed by the pragmatic irrelevance of the relevant prudential reasons in many cases. Contrarian bravado can tend to give the impression that the truly rational agent is often rationally permitted, perhaps required, to ignore social norms. But being in a position to ignore what others' think of you is a (vanishingly rare) privilege. A moment's reflection on pernicious social norms' role in maintaining iniquitous regimes of social discrimination should dispel the appeal of this attitude. Besides, as it turns out, establishing that social norms are necessarily reason-entailing is crucial to delivering a complete account of how they should feature in the deliberations of the rational human agent.

The account begins by recognizing that necessarily, *human* agents are cognitively limited, fallible, social creatures. This, of course, is not true of all rational agents, and my account of the normativity of social norms presupposes this important feature of what it is to be human. Being cognitively limited, fallible, social creatures means that realizing our aims very often requires some *cooperation* from other human agents. As I will argue later, even where those aims are individualistic or "siloed", at a minimum, their realization requires not being *interfered* with by other agents as we attempt to realize our aims. Human agents are also the sorts of creatures whose opinion of one another affects their willingness to aid, cooperate

[72] Of course, most people also desire esteem for its own sake *and* wish to be polite, respectful etc. I am not assuming people are callous, instrumentalising egoists. Rather, I am attempting to show that the reason entailed by each and every social norm does not require the presence of common prosocial traits. Cf. (Brennan and Pettit, 2006).

or (not) interfere with one another. So, the esteem in which others hold us is *crucial* to realizing our aims, regardless of what those aims are. *Observable* violation of a social norm, according to the norms-as-normative-attitudes account, *inevitably* occasions the risk of being the subject of *negative normative attitudes*. Just what it *is* for some principle to be a social norm in a given group is for members of the group to be disposed to disapprove of violations of the conduct the principle requires. Note that *observability* does not require that the violation is actually *observed*. All that is needed is some appreciable *risk* and this, I will argue in greater detail below, is *inescapable*.[73]

Let me therefore preface a more formal statement of the argument with an important caveat: the weight of the reasons social norms entail will often be very *weak*. While this might look like a problem for my argument, I hope to convince you it is exactly what we should expect. Sometimes, the reasons one has to comply with social norms *are* very weak. Just as sometimes, they are overwhelmingly strong. At this stage, all I want to establish is that there is *always* some reason. Part of the story I want to tell is that we are *distracted* from attempts to provide a uniform account of social normativity by our sense that it is an utterly contingent matter whether the fact that a social norm requires øing gives you any genuine reason at all to ø (Enoch, 2011b). If one could establish that social norms *do* entail reasons for all human agents, then the focus can shift to providing a uniform account of how, which will form the basis for determinations of their weight and integration into rational choice theory more generally. Happily, as it turns out, testing the *limit cases* that loom as potential counter-examples to the argument I set out below is an excellent way of proceeding to develop an account of precisely that structure and the weight of the reasons social norms provide. In fact, *even if* you remain unconvinced that social norms entail reasons for action after evaluating the argument below and the cases considered in the next section, my hope is that you agree that they come so *close* that a uniform account of their role in deliberation is an attractive prospect: the beginnings of which I hope to deliver in the second half of the paper. With this brief *précis* and our definition of social norms in hand, we can set things out a little more formally. The central claim of the paper is that the following thesis is true:

> **Social-reason-entailment** Necessarily, if N is a social norm requiring a human agent A to perform an action ø in circumstances C, then A has a reason to ø in C.

The basic argument for Social-reason-entailment proceeds as follows.

---

[73] Though, of course, I am not suggesting this fact *alone* grounds normativity *via* rational inescapability of the sort criticized in (Silverstein, 2012).

(1) Assume N is a social norm requiring a human agent A to perform an action ø in circumstances C.[74]

(2) A significant proportion of members of the group in which N is a social norm have negative normative attitudes towards violations of N.

(3) Being the person subject to negative normative attitudes can lower the esteem in which the person is held by people possessing those attitudes.

(4) If A violates N, there is a non-negligible risk of being observed by an observer O.

(5) When a violation of N is observed by O there is a non-negligible risk A is disesteemed by O or others, on the basis of O's testimony.

(6) When A is disesteemed there is a non-negligible risk this will impede the realization of A's actual or possible desires now or in the future.

(7) If A violates N, there is a non-negligible risk this will impede the realization of A's actual or possible desires now or in the future.

(8) A always has a reason not to do that which non-negligibly risks impeding the realization of A's actual or possible desires now or in the future.

(9) So, A has a reason to ø in C.

That is, Social-reason-entailment's truth is grounded by a universal hypothetical imperative, contingent only on human A's (and all human As') having any actual or possible desires whatsoever.[75] Note some important features of this argument. First, the risk involved needs to be 'non-negligible' (Smith, 2016). But as my discussion of cases below will show, the bar for negligibility is not all that high, i.e. that we are social creatures that might be observed unawares by people that might affect our interests in the future is sufficient to generate the reason, given the wealth of evidence that generally, people disapprove of social norm violation

[74] To re-iterate, I take it as part of the definition of 'human' that the agent in question is a cognitively limited, fallible, social creature, with no special powers rendering them immune from interference from others or ability to perfectly predict their future desires and so on. This argument would not work were it applied to all rational agents.

[75] Must the argument for Social-reason-entailment appeal only to the normative significance of desire-satisfaction? No, in fact, I believe there is a closely related argument that might appeal to those with more heavyweight or alternative normative commitments. For example, assuming that norm compliance and violation always carry with them a non-negligible chance of adding or subtracting from one's social credit, and assuming that one's social credit is directly tied towards how much freedom one has to act in the social domain (by being considered trustworthy, reliable, norm-compliant, non-deviant, reasonable and the like), then one always has a reason to comply with a social norms insofar as one has reason to value one's freedom which, given its intrinsic value, one always does.

within the group in which the norm obtains and with predictable effects on one's standing in the group (Brennan & Pettit, 2006).

Second, the agent in question need not be a member of the group in which the social norm obtains. An Irish tourist in China has a reason to comply with Chinese social norms. But the same person need not comply with Chinese social norms when they are at home in Ireland: the *circumstances* for the application of a given social norm are *spatiotemporally indexed* to the group in which the norm obtains, i.e. "When in Rome".[76] Next, note that premises (2) and (3) follow from the definition of social norm I endorsed earlier. Just what it *is* for something to be a social norm is for a significant proportion of the members of some group to have normative attitudes that reflect it, paradigmatically: taking a dim view of violations of the norm.

But suppose – contrary to what I believe – that (Brennan et al., 2013)'s account of 'norms as normative attitudes' is mistaken. Could the basic argument in favour of Social-reason-entailment be modified to suit an account less obviously congenial to it? In a word, 'Yes'. But further premises would be required, the exact content of which would naturally depend on the details of the view. For example, suppose it turns out that a *reductive* view of social norms is true, along the lines proposed by Christina Bicchieri (2006). In my view, a further premise would be required in my basic argument to establish that the risk of loss of esteem following observed norm-violation was inevitable. (Brennan & Pettit, 2006) and (Pettit, 1990) suggest how such a link might be established: the disappointment associated with defeated legitimate expectations, for example, should be enough to generate the relevant *risk*, even where the expectations are themselves a function of (conditional) desires rather than intrinsically normative attitudes.

---

[76] This also explains why, for example, we no longer have reasons to comply with *outdated* norms of our community. 1950s' social norms that were once genuine social norms in my community no longer give me reasons because the attitudes and beliefs of my community no longer reflect them. But spatiotemporal indexing will, of course, generate some complications. What if a *group* of Irish tourists are dining at a restaurant in China? My argument implies they are subject to Chinese social norms, but they may well also be subject to Irish social norms too, given the group setting. Overlapping contexts generate the possibility of normative conflict, as does overlapping group membership generally. All of us occupy a variety of social roles (parent, teacher, partner, club-member, citizen) and the norms pertaining to each role can conflict. When they do, they will give rise to normative "social dilemmas". Since I believe other genuine normative (e.g. moral) dilemmas are possible, I do not find this conclusion theoretically troubling (Williams, 1966). Of course, as a practical matter, it is often deeply troubling. I say a little more about this in Section 6. One solution to problems like these is to peg the weight of the competing reasons to the value of the social identity underpinning them (Korsgaard, 2009) but I suspect the prudential normative situation is better modelled on the sub-formula I label 'Weight' in Section 5.

More troubling for my argument are possible counterexamples to premises (4) and (6). The sort of tricky cases I have in mind should be pretty obvious. First, one might object that we are very often *un-observed* and, when those circumstances obtain, social norms give us no reason to comply. Second, some of us do not care about what other people think *at all* and none of our present or future possible ends require the possible cooperation of others. Third, some of us *relish* causing disapproval in others or enjoy bringing attention to the arbitrariness and pettiness of social norms by violating them in the presence of the people who hold them most dear. Since all these types live among us, and since we can often cloak our misdeeds in secrecy, the objection might run, nothing even close to Social-reason-entailment holds. At the very least, we need to know some important facts about the relevant agent, *viz.* whether they care what other people think or have present or possible ends that require social interaction or are a subversive or activist *and* we need to know a crucial fact about the circumstances, *viz*., whether there is a possibility the agent will be observed.

But although these objections sound extremely plausible, closer inspection reveals them to affect not the *existence* of the reasons in question but their *weight*. Or so I will argue. More importantly, as it turns out, interrogating *how* these limit cases come close to extinguishing the reason(s) in question provides the beginnings of a substantive account of which factors are relevant to determining the weights of social reasons generally, an account I then develop in more detail in section 5. The three cases considered are naturally not exhaustive, but they will illustrate an argumentative strategy that, I think, generalises quite well.[77] I will call the cases 'Solo norm deviant', 'Omniscient norm deviant' and 'Jolly norm deviant'. The point of *Solo*, *Omni* and *Jolly* is to show that human agents are *never* in circumstances where a social norm applies yet they have *no reason whatsoever* to comply. That is, the circumstantial caveat often used to denigrate the normativity of social norms is ultimately toothless.[78] There are important differences between their normativity and that of moral, epistemic and prudential norms (as I discuss in section 7) but the difference does not consist in the fact that they are not reason-entailing for humans.

4. Norm deviants

An obvious objection to the basic argument for Social-reason-entailment is that sometimes we are not observed; do not care what other people think about us; and have no reason to care

[77] I consider a fourth and even more troubling case in section 6.

[78] Cf. (Enoch, 2019, p. 72).

because we have no need of their goodwill. To consider the full force of that objection, consider a case like the following:

> **Solo norm deviant** tired of the annoying, inefficient and sometimes downright unjust norms present in each and every human society, Solo retires to a cabin in a remote part of Alaska, where he lives in total self-sufficiency. He relishes holding his knife and fork "incorrectly", wearing no clothes, picking his nose at will and is happy to die alone from minor infection rather than participate in human interaction to receive medical treatment.

We can imagine Solo happily going about his day, naked, nose-picking, and generally enjoying all the freedoms solitude permits, including, paradigmatically, freedom from the need to worry about what other people think of you. Let us further suppose that there are social norms in Alaska requiring holding one's cutlery a particular way, wearing clothes in public and not picking one's nose in public. Do these Alaskan social norms give Solo *any reason whatsoever* to behave accordingly? Let me start by granting something I think is false: Solo has genuinely removed himself from even the slightest, tiniest chance of being observed or of interacting with anyone. In that case, I agree, Solo (so long as that condition obtains), has no reason whatsoever to abide by the relevant social norms.

Is this a devastating concession? I do not believe so. Social norms are the norms that *govern social interaction*. Nothing could be less surprising than the fact that if you could *guarantee* that there will never be any social interaction of any kind where you are, then you are not subject to the norms of the society who occupy some of the territory that surrounds your carved-out, isolation-guaranteed segment of it. Any necessary reason-entailing capacity social norms possess is restricted to (potentially) social settings, in which all or almost all humans in fact live. But society may not be *inescapable* and so, social norms may not be either.

This first point is related to a second. Suppose a hiking group get lost in this remote part of Alaska and stumble across Solo's encampment. Although relieved they will not die of starvation or be eaten by a bear, they are nonetheless disturbed to find a weird man violating the "local" pro-clothes-wearing and anti-nose-picking social norms and think less of Solo as a result. Of course, Solo does not care what they think, because he has no present desires requiring their cooperation (except perhaps getting rid of them). Is this enough to ensure that he has not the slightest conceivable reason to care what they think *at all*? Two observations are in order. If they survive and return to society and report what they've seen, Solo's peace may be disturbed by people (pesky officials or others) who, for example, do not approve of his free-spirited naturist ways. Even if not, should Solo ever decide to reintegrate to society, which he

does not presently desire to do but may do in the future if he gets lonely, bored, sick or is unable to fend for himself any longer but reluctant to die, then his interests may be materially affected by their reports of him. For example, as he attempts to make it back into town, get food and find somewhere to stay. This is not a conclusive reason to change his behaviour, far from it! That's why he went into the woods in the first place. But that does not mean he has *no* reason to comply with the relevant norms, if this is how people will react. Rather, it means that the reason is overcome by his much stronger desires for solitude, nakedness and nose-picking. Of course, this tenuous case is ridiculously far-fetched and the reasons laughably weak, but then this also the limit case, and it seems like *even here* it is false to say that the social norms give Solo *no reason* whatsoever.[79]

Solo's case is silly. But the lesson for developing a substantive account of weighing social reasons is important. The *weight* of the prudential reason a social norm always provides is, in part, a function of the probability of observation, together with the probability this observation will affect one's interests now or in the future. In Solo's case, the weight of that reason is featherlight. But for the vast majority of us in our day-to-day interactions, it will be considerably heavier.

So much for Solo. If his isolation is complete, then so too is his freedom from social norms. Otherwise, not. But, we might think, whether alone in a lift or in the Alaskan woods or surrounded by same-minded norm eschewers, what really counts is what you *know* about your situation. If you *know* you're alone or you *know* the people who might see you don't care about the relevant norm or you *know* you will never, ever change your mind about living in the social world and having ends that could conceivably require human interaction, then there's no problem. This thought might lead us to consider a case like the following:

> **Omniscient norm deviant** Omni knows everything, including who is watching her, what they are thinking and what her future desires and plans are. As a result, Omni often knows that she can violate various social norms with impunity and does so, with an impish glint in her eye.

Unlike Solo, Omni need not remove herself to the woods to get away with norm-violating behaviour risk-free. She can walk among us violating norms when the conditions are right. We can deal with this case much more quickly than Solo's. Since no actual deliberating *human*

[79] Cf. (McPherson, 2012) and (Shafer-Landau, 2012), but see also (Schroeder, 2012)'s response.

agents are or could be omniscient, they cannot know what Omni knows.[80] It is paradigmatically irrational to behave as if one is infallible when one is very much not, so no one should take their cues from Omni here. Of course, one should counter here that one can be far less than omniscient and still know things: all that is really required is that one know everything about one's present surroundings, present and future desires and how the opinions of potential observers in the vicinity of present norm violation will affect them in the future. My point is that this is precisely the sort of knowledge we cannot have. Of course, that is not to say we cannot make reliable predictions: among the most reliable of which is, *ceteris paribus*, social norm violation, if observed, will likely decrease the esteem in which we are held, which will likely decrease the degree of cooperation people are willing to offer in the future as we attempt to realize our ends. And personal experience should teach us that we are often observed unawares: most of us have done something believing we were unobserved, then wished we hadn't, when we realized we were. Obviously, the risk is often sufficiently low that the reason it provides is *decisively outweighed*. But, as with Solo, that is not because there is *no reason*, but rather because the reason in question is sometimes very weak (though we are often guilty of underestimating its weight). Viewed in this light, Omni's lesson is simply that the strength of the prudential reason a social norm provides varies with how much we know about our situation and how stable our various desires are. I say more about this in section 5. Omni's case also helps us to appreciate an important point about the reason-entailing power of social norms. The reasons we have to comply are often generated by the fact that we are acting and deciding under uncertainty. If we always knew all the facts, then social norms would be deprived of some of their reason-entailing power. This is the crucial difference between Social-reason-entailment and Robust-reason-entailment: norms satisfying the former, but not the latter, are limited to agents like *us*. I say more in the final section of the paper.

For now, let me turn to a third potential counter-example which will – I hope, as with the previous two cases – help us to build our substantive account of the weight of the reasons social norms entail. In this case, it is contingent features of a human agent that seem to deprive the social norms of their reason-entailing power:

> **Jolly norm deviant** Jolly delights in violating social norms. Sometimes, this is simply because it is funny to watch people squirm with discomfort or to imagine them silently judging her. That is, sometimes, Jolly violates social norms *just for*

[80] There are, of course, various contexts in which uncertainty is relevant to what we should do or think. See, for recent examples, (Dorst, 2020) and (Podgorski, 2020).

> *fun*. But other times, Jolly violates norms because it is, she believes, the right thing to do. She does this whenever there is a norm that is immoral or stupid or inefficient or otherwise objectionable to bring attention to precisely this fact. At these times, Jolly takes herself to be engaging in a form of well-justified social activism.

There is a lot to like about Jolly. One of the things to like about her – relevantly for our purposes – is that she has her fun *despite* the fact that it might be personally costly. Indeed, it is fun, in part, precisely because it involves a refusal to allow the fact that people may think less of one affect what one does (or, rather, it does affect what one does, but only in a contrarian sense). But we can make a by now familiar observation. The fun is possible and Jolly admirable *because* the social norm provides her with a prudential reason to comply. So, this case is no threat to the universal reason-entailing power of social norms (any more than a rock-star's very cool and creativity-facilitating self-destructive behaviour is a threat to the reason-entailing power of prudential norms). What about Jolly's activist mode? Again – and in a close analogy to civil disobedience generally – I think the best view is that Jolly's activism is made possible by the existence of the prudential reasons the violated norms give her to comply (Rawls, 1971; Brownlee, 2012). It is because it is personally costly to violate the norms that Jolly's actions are particularly meaningful. The lesson from Jolly with regard to developing our substantive account: you can manipulate the normative relevance of social norms in a variety of situations *because* they entail reasons of a predictable sort.[81] The better grip we can get on how social norms give us reasons for action, the better chance we have of succeeding in the social world, regardless of what our aims happen to be. Including, especially, where our aim is to *change* features of that social world. I return to consider this point in a more troubling context in section 6.

Summing up, what can we learn from considering cases like Solo, Omni and Jolly? My suggestion: the basic argument for Social-reason-entailment is sound, even where the probabilities involved are low enough to explain some lingering incredulity.[82] Each and every social norm gives each and every *human* agent subject to it some reason to do as it requires.

---

[81] There is an alternative explanation of the Jolly case. Namely, Jolly's activism is made possible by, and is crucially parasitic on, the existence of reasons that we normally have to comply with relevant norms. That is, normally but not necessarily. I am happy with this alternative explanation, provided it negates the potential counter-example Jolly might otherwise pose to my argument, which I believe it does.

[82] That is, the probabilities are not rationally negligible (Smith, 2016).

So, the highly intuitive and very plausible orthodox view is false. Is this result merely a technical curio? I do not believe so. In fact, I think it might help when deciding what to do; evaluating conduct; and in theorising about the relationship between social normativity and other forms of normativity. It is to these three applications of Social-reason-entailment that I now turn. As I suggested earlier, my aim in considering cases like Solo, Omni and Jolly was not just to defend Social-reason-entailment. These cases are important because they help us to see basic features of *how* social norms provide reasons and the *weight* of those reasons.

5. Deciding in light of social norms

Suppose Social-reason-entailment *is* true. So what? Establishing that social norms necessarily give people reasons to do as they require, regardless of the content of the norm and the circumstances of the agent to whom they apply, might be surprising; but is it useful? As even the briefest perusal of the metaethical and social scientific literatures on social norms indicates, there is no shortage of highly sophisticated theorising concerning norms' nature, emergence, persistence or the dynamics of their change and decay, which accounts are supplemented by equally insightful theories regarding why agents do, in fact, comply with social norms. But, in part precisely because of what we know about social norms are and why we often do as they require, it has proven much more difficult for philosophers to say anything in general, let alone *systematic*, about how a deliberating agent *should* decide what to do in light of an applicable social norm, in a fashion comparable to the many attempts at providing analogous stories regarding the norms of prudence, ethics, epistemology or rationality itself. But the truth of Social-reason-entailment suggests that such a story might be told or, at least, that the attempt to tell it is no more Quixotic than those for prudence or morality. Many philosophers think that a norm's status *as* a moral or prudential or epistemic norm provides us with important extra information – namely that, *just in virtue of that status*, the norm gives the agent *some* reason to comply. So too, if the argument of this paper is correct, does a norm's status as a social norm provide us with this important extra information; and this fact also provides the basis for the following calculation: [83]

> **Weight** The weight of the reason an agent has to comply with a social norm, just in virtue of its status as a social norm, is a function of the probability the agent

[83] Of course, how exactly one should go about "weighing" reasons in deliberation raises many difficult questions and the attendant literature is vast. Helpful discussions include (Chang, 2001), (Schroeder, 2007), (Snedegar, 2017), (Baker, 2018), (Lord, 2018), (McPherson, 2018), (Nair & Horty, 2018) and (Tucker, 2021).

> will be observed transgressing the norm, the probability of loss of esteem as a result, and the expected value of that esteem to the agent's attaining their present or future ends.

Where, of course, for most but not all agents, one of their present or future ends is the attainment of esteem itself and where, if the argument of this paper is correct, the probability values for all three components of Weight are greater than 1.

This, I suggest, is the only reason that necessarily follows directly from a norm's status as a social norm, regardless of its content or the circumstances in which an agent finds themselves, which is a potentially helpful observation, in the following sense. The rational deliberating agent can disentangle this ever-present reason from the many contingent reasons the content of the norm and the surrounding circumstances in which the agent finds themselves might provide; for example, the extent to which compliance with the norm expresses a valuable practical identity of the agent's, the extent to which non-compliance might violate the legitimate interests of third parties or otherwise affect their interests in ways that trigger further moral or prudential reasons, and the extent to which, for example, the social norm in question represents the equilibria of various games of strategy, in such a way that either compliance or defecting is or is not an advisable manoeuvre. Of course, Weight is only a schema: calculating the relevant probabilities and expected values is where most of the action is, but having the schema at least makes this task much more manageable. That calculation can then be nested in further calculations with regard to the aforementioned moral, prudential and identity-based reasons that may or may not be salient, depending on the norm, circumstances and agent in question.[84] Weight is a sub-formula, designed to contribute to an all-things-considered decision calculus in which an agent must decide how to act in light of an applicable social norm. If, for example, the agent sees no other reason to (not) comply, outside of their own (dis)inclination to do so, then the calculation simply involves weighing the value of Weight against the strength of the agent's countervailing preferences. But, of course, the calculation may be much more complex, involving moral or other-regarding considerations, strong feelings of allegiance or identification with the group in which the social norm obtains and so on. The significance of Social-reason-entailment is not that it makes such calculations *easy*; though, one might hope, it might make them easier. The point is to help make the calculations *correctly*.

[84] It is at this point that one might attempt to deliver a more comprehensive account, integrating insights from, for example, (Korsgaard, 2009), (Manne, 2013), (Woods, 2018) and (Valentini, 2019).

6. Perpetuating injustice

Consider the following invidious scenario:

> **Persecuted norm-complier** lives in a racist society in which many of the norms unfairly discriminate against him. He complies with these norms because norm transgressors are punished by various forms of social ostracization that make life extremely difficult for the transgressor and their family. Many members of the oppressor group view members of the persecuted group as weak and craven for complying with this set of norms, contributing to the low esteem in which members of the oppressor group hold members of the persecuted group, which they then use as a justification for the discriminatory regime.

"Damned if you, damned if you don't" scenarios like this one are depressingly common, whether based on race, gender, caste or some other grouping. Does their ubiquity show Social-reason-entailment is false? In a word, no. In fact, they serve to further illustrate the utility of analyzing social normativity in the manner defended in this paper. From the perspective of the deliberating agent, it is important to know that failure to comply with the discriminatory norms in question always brings with it the risk of being observed and a consequent risk of the diminution in the esteem in which the agent is held by other members of the society. There is plenty of evidence that individuals of a persecuted class who fail to comply risk being branded 'trouble-makers' or 'difficult' in ways particularly injurious to the fulfilment of many of their actual or possible desires and they therefore have some reason to comply.[85] These perverse incentive structures are precisely how iniquitous regimes of social discrimination are maintained. This is consistent with the existence of a countervailing reason not to comply with the social norm that follows not from the social norm itself, but from understanding how iterated collective compliance is perceived by members of the oppressor group.

Social-reason-entailment is also useful in explaining and evaluating the actions of the oppressor group. The oppressor group in this example are paradigmatically *irrational* along multiple axes: both because they are guilty of a particularly galling form of fundamental attribution error in light of the unjust system of incentives they themselves have created *and* because they use the effects of that system to justify its imposition, thereby engaging in particularly objectionable form of bootstrapping. That irrationality infects the options available to the persecuted group, which is why the very norms that give them some reason to comply,

[85] See (McAdams, 1995), (Bicchieri, 2006), (Henrich & Henrich, 2007), (Brennan et al., 2013), (Bicchieri, 2017) and (Appiah, 2018).

just in virtue of their status as social norms, also give both groups (oppressor and oppressed) collective moral and prudential reasons to overthrow the unjust norms. This is not only consistent with Social-reason-entailment; it presupposes it, as we observed in the case of the Jolly norm deviant. It is in part because the social norms provide the prudential reasons that they do that they are unjust *and* that makes effective civil disobedience possible. Additionally, not only does Social-reason-entailment get the right result regarding the individual deliberating agent's reasons, it also helps to explain why the oppressor group's views are not only grossly unjust, but paradigmatically irrational. Finally, note that the irrational tendency to discount the reason-entailing power of social norms when evaluating the conduct of others is far from limited to these extreme scenarios and that here too, attention to Social-reason-entailment and Weight can assist. It is worth keeping in mind that people *always* have some reason to do as social norms require, that's just the sort of things that they, and we, are.

7. Normativity, socialized

Robust norms entail reasons for action, while merely formal norms do not. Merely formal norms *sometimes* give you reasons *depending on the circumstances*. But they do not entail reasons for action just *qua* norm. Therein lies the distinction.

Are social norms merely formal? They are usually thought to be the *paradigm* case of merely formal normativity. Yet, if the argument of this paper is correct, this conclusion is difficult to sustain. Social norms *always* give you a reason and this property is usually understood to be the *defining* property of robust normativity. Does that mean social norms are robustly normative? I think not. The explanation for *why* and *how* they always give you a reason is important: their normativity derives from a more fundamental normative source.

The significance of this for explaining robust normativity is as follows. Universal reason-entailment is *necessary* but not *sufficient* for robust normativity. What, then, should we now say about formal normativity? My suggestion: it bifurcates. *Merely* formal normativity is the generic form of normativity, as it is traditionally understood, entailing nothing further about reason-giving power. But *quasi-robust* normativity is an intermediate category between the two: quasi-robust norms entail reasons for action, but they do so conditional upon further normative facts.

Earlier I prefaced a friendly amendment to Enoch's triggering account of reason-giving. Here is what I had in mind. One skeptical response to my quasi-robustness proposal is the following. All I have established is that social norms universally and uniformly *trigger* reasons. According to the triggering model, mere triggering reason-giving occurs due to alteration of

the *non-normative* facts, while robust reason-giving is an instance of triggering reason-giving characterized 'in terms of the rather complicated intentions of the reason-giver and the normative background' (Enoch, 2011a, p. 2). Social norms are merely formally normative, on the orthodox view, and so merely trigger reasons by altering the non-normative facts.

But, of course, the distinction between merely formal normativity and robust normativity is itself usually cashed out in terms of whether the norms in question universally entail reasons for action. If the argument of this paper is correct, then social norms are not *merely* formally normative. So, it is awkward to characterize them as merely altering the *non-normative* facts. Instead, social norms alter the *normative* facts, but unlike fully robust norms they do so *indirectly*. Social norms' reason-giving power is *derivative*: it is parasitic on the robust normativity of other more robust norms. But derivative normativity is not non-normativity. On the contrary, to be derivatively genuinely normative just is to be genuinely normative, but not self-sufficiently so. If this suggestion is on the right track, it sharpens our inquiries into the nature of robust normativity too. Robust normativity is not only universally reason-entailing, but *non-derivatively* so. The distinction between different forms of normativity is therefore tripartite, where the taxonomical differences are a function of both reason-giving power and the *source* of normativity.[86]

8. Conclusion

This paper defended two claims. First, I argued that you *always* have some reason to comply with the social norms applicable to your situation, no matter how immoral or ridiculous. This was not a mere technical curiosity: it is fundamental to understanding what social norms are, how they work, and how they can be a source of grave injustice. Second, I argued that this fact about social normativity requires re-thinking the distinction between formal and robust normativity. Social norms are not robustly normative. But nor are they merely formally normative. Reviving our taxonomy accordingly not only allows it to accommodate this important phenomenon, it sharpens our understanding of the pre-existing categories too.

[86] Cf. (Korsgaard, 1996). This has the welcome result of integrating a contemporary metaethical taxonomy with a fruitful – but alternative – paradigm of normative explanation.

CHAPTER 5: EPISTEMIC ANGST IS A SOCIAL PROBLEM

Introduction

The aim of this paper is to investigate a phenomenon I will call 'the social problem of epistemic angst' or just 'epistemic angst' for short. Its main purpose is exploratory: to establish that there is a unified phenomenon worth further investigation. The idea is simple enough: sometimes our social interactions with others give us reason for radical doubt concerning our epistemic capacities. The feeling that attends or is partially constitutive of this doubt is visceral: it strikes at the core of our perception of ourselves as competent agents capable of understanding the world and making sensible decisions about how to live our lives. It can be, variously, bewildering, paralysing, anger-inducing, tragic, fascinating, scary and, of course, drastically undermines our epistemic self-confidence. It can also be exploited.

The philosophical topics in which epistemic angst is a key explanatory factor include, at least, peer disagreement, epistemic injustice, trust in expertise, gaslighting, epistemic coercion, epistemic anxiety, and identity protective cognition. I strongly suspect this list is not exhaustive. Epistemic angst features in day-to-day interactions too: the shame, fury or sadness we feel at being outwitted, deceived, or bested in argument, or the outrage we feel when lied to by politicians, large corporations, or others with great social, political or economic power we *trusted*, are also likely manifestations of the same underlying phenomenon. So too are the feelings of embarrassment and concern we feel when we could not see something that others could or mistakenly thought we had seen something others assured us was not there. Once you start looking for it, epistemic angst is everywhere.

The plan for the paper is as follows. In Part I, I provide a working definition of epistemic angst and explain how further investigation into the phenomenon might help us better understand a variety of philosophical and social problems. Part II comprises a worked example of epistemic angst's utility in deepening our understanding of existing philosophical and social problems. I demonstrate the role epistemic angst plays in mistrust of expertise and use it to diagnose a novel aspect of this problem fuelled by what I label 'the erosion of competence.' In Part III, I sketch a partial solution to the problem of epistemic angst as it manifests in mistrust of expertise and comment on the scalability of this solution across the many different domains plagued by epistemic angst. A partial remedy to the epistemic social ills posed by epistemic angst is recognising its role in our lives, good and bad. But a recurring theme is that epistemic angst is like quicksand. The more one struggles against it, the deeper one is enmired. The

paper's hope is that recognising the hold epistemic angst has on us might help loosen its grip, but care is required to ensure it does not tighten instead.

PART I: EPISTEMIC ANGST

In this Part, I provide a working definition of epistemic angst and situate it among the literatures on radical skepticism, feminist social epistemology, epistemic anxiety, and identity protective cognition.

1. What is epistemic angst?

Epistemic angst is a radical state of epistemic self-doubt.[87] A person in the grip of epistemic angst loses or is in the process of losing faith in their ability to know and understand the world and make sensible judgments about what to do. It is reflexive and generalised, in the sense that it is not epistemic doubt concerning one's reasoning or judgment on a particular matter or topic, such as when one wonders whether one is biased or insufficiently well-informed to have justified beliefs in some domain of inquiry or because one lacks the requisite expertise to understand the relevant questions or evidence. Instead, epistemic angst is radical and global doubt about one's status *qua* reliable knower or deliberator. It is also importantly different in some key aspects from the doubt attending more traditional, radical forms of Cartesian or external world skepticism.[88]

Most saliently, social epistemic angst does not arise from contemplation of one's inability to exclude global skeptical hypotheses or justify one's beliefs in non-circular ways. It arises in response to (what appears to be) *evidence* that one is wrong about some fact(s) about which one previously felt epistemically confident and in circumstances where *if* it turns out one really is wrong about those facts, one is very likely wrong about many other matters too. That is, epistemic angst arises when we are given reasons to think that our belief- and judgment-forming mechanisms are deeply unreliable and where we see no way of knowing *when*. The source of the epistemic angst in which we are interested is *social*. That is, the evidence that

[87] It is radical in the sense of having very wide scope. But the intensity of the affect comes in degrees: for example, some cases of peer disagreement or reflection on the genealogy of some subset of our beliefs might make us mildly uneasy about our epistemic capacities as whole, while other experiences in which we discover evidence we were wrong can be completely devastating to our epistemic self-confidence.

[88] I say more about the connection to Prichard and Sosa in a moment.

suggests that one is wrong concerning some matter about which one previously felt confident comes in the form of people – better informed, or, at least, equally well-informed but in large numbers – thinking you are mistaken and providing reasons why this is so and, moreover, purporting to explain not just *that* and *why* you are wrong on the topic in question, but purporting to explain your error *ad hominem*, by implicit reference to your inferior status or reliability *qua* knower: whether because of social group, education, epistemically compromised position, bias or downright stupidity, in way that cannot – and often is not intended to – be quarantined to the matter at hand.

This important fact about social epistemic angst, *viz.*, that it is fundamentally resistant to limitations of its scope, it what makes it so difficult to handle. Confronted with a choice between accepting evidence that you are radically mistaken on some topic in circumstances where you have deployed your very best epistemic capabilities and upon which you had come to hold your beliefs with very high credence or rejecting it, accepting it seems tantamount to giving up on yourself, epistemically speaking, unless you can see *why* and *where* you went wrong and have been provided with tools for ensuring it will not happen again.

Summing-up, epistemic angst as a social problem arises when social interactions challenge our sense of ourselves as reliable knowers. I define social epistemic angst as generalised anxiety and unease concerning one's epistemic capabilities and good standing in one's community. Angst concerning our good epistemic standing goes to the core of how we perceive ourselves: as competent agents capable of making sensible decisions about how to live our lives. I will argue this angst lies at the heart of some thorny issues in social and political epistemology.

This paper locates the phenomenon of epistemic angst *qua* social problem at the intersection of four literatures. First, there are discussions of epistemic angst (Pritchard, 2016; Sosa, 1991) as a deeply personal, cartesian problem, arising in response to global, radical skepticism. Second, are discussions of epistemic injustice (Fricker, 2007), epistemic coercion, gaslighting (Abramson, 2014) and self-gaslighting (Dandelet, 2021a) from feminist social epistemology. Third, there is the literature on epistemic emotions. Most notably, for our purposes, on epistemic *anxiety* (Cabrera, 2021); (Newton, 2022); (Nagel, 2010); (Vazard, 2018, 2021). Fourth, is the literature on peer disagreement and expert testimony (Goldman, 2011); (Benton, 2016); (Bokros, 2021); (Constantin & Grundmann, 2020); (Duijf, 2021); (Kelly,

2011); (Levy, 2019); (Nguyen, 2020).[89] Before discussing those literatures in a little more detail, let me outline in brief how my proposed account of epistemic angst builds on them.

In broad outline, my account of epistemic angst builds on the contributions of each of these literatures as follows. First, I argue that epistemic angst is *also* a deeply social, interpersonal problem. Second, I argue that epistemic angst is a more general, but not necessarily unjust, form of the epistemic self-doubt discussed in some parts of feminist social epistemology. Third, it differs from epistemic *anxiety*, as it is defined in the literature on epistemic emotions, in that it is reflexive, generalised, and detrimental to inquiry. Fourth, it aims to secure a place for epistemic angst as both a contributor to and consequence of distrust of experts and peer disagreement.

1.1 Radical skepticism: Prichard on epistemic angst

In this section, I briefly set out Prichard's account of the source of epistemic angst and his solution to it. Prichard argues that the challenge of global, radical skepticism confronts us – and should confront any responsible inquirer – as a deeply troubling, existential question. As he understands it, there are, in fact, not one but two different skeptical paradoxes, each of which requires a different solution. The paradox relevant to epistemic angst as a social problem is the following:

Closure-Based Radical Scepticism Qua *Paradox*

(I) One cannot have rationally grounded knowledge of the denials of radical sceptical hypotheses.

(II) The closure principle.[90]

[89] But, if my hunch about how widespread this phenomenon is proves correct, it will also intersect with a variety of literatures from other disciplines. Dan Kahan's work on identity protective cognition, for example, is of obvious relevance: our identity *qua* reliable knower and competent decision-maker is just as important to us (and perhaps upstream of) our valued socio-political identity, so, if the evidence for identity protective cognition is good, epistemic angst may help explain motivated reasoning that protects our identity as a reliable knower (Kahan, 2017).

[90] The Closure Principle

If S has rationally grounded knowledge that p, and S competently deduces from p that q, thereby forming a belief that q on this basis while retaining her rationally grounded knowledge that p, then S has rationally grounded knowledge that q. (Prichard, 2016, p. 71).

(III) One has a large body of rationally grounded knowledge of the external world.[91] (Prichard, 2016, p. 72)

Prichard argues that (his interpretation of) Wittgenstein's argument against the universality of rational evaluation and the centrality of hinge commitments can solve the closure-based radical skepticism problem, but not the underdetermination-based radical skepticism problem; while epistemological disjunctivism can solve the underdetermination-based radical skepticism problem, but not the closure-based radical skepticism problem. Combining the two, he argues, allows us to solve both in a such way that the two solutions support each other.

But Prichard acknowledges that even having solved the problem of epistemic angst, lingering concern will remain. We have been epistemically humbled, as it were, by engaging in the process of 'epistemic ascent' that leads us to realise *that* our rational evaluation must be local and *that* the support for our beliefs is insular. This leads to some residual anxiety concerning our overall epistemic perspective that he describes as 'epistemic vertigo'.

As outlined above, I think there is an interpersonal, social form of epistemic angst, namely, that associated with considering the asymmetry between experts' knowledge and our own. My suspicion is that this manifests as anxiety directed towards our own epistemic capabilities. The basic idea is that recognising the extent of what we do not know and cannot understand *in circumstances where there are others who do know it and can understand it* pushes us to confront a painful implicit epistemic comparison very much to our detriment. The worry, if, as these experts maintain, the matters on which they are pronouncing are of grave importance to how we should live our lives, is that it is deeply troubling and alienating to be told that you are so poorly epistemically equipped that you do not even *understand* how much you do not understand about these matters of immense importance for how you should live your life and so *must* rely on expert testimony due to your implicit epistemic inferiority. Note that this hits even harder if you have come to have a stock of beliefs about such matters that these experts claim are false (that have been further reinforced by merchants of doubt), such

[91] The underdetermination paradox, which is less relevant for our purposes, he sets out as follows:

Underdetermination-Based Radical Scepticism Qua *Paradox*

(I*) One cannot have a rational basis that favours one's everyday beliefs about the external world over radical sceptical hypotheses.

(II*) The underdetermination principle.[91]

(III*) One has a large body of rationally grounded knowledge of the external world. (Prichard, 2016, p. 75)

that you presently believe that you *do* understand the issues in play. You are now being told not only that you are *wrong*, but that you are not even smart enough to know when you do not understand what is going on (since it seems to you that you currently do understand what is going on).

The link to the radical skeptical paradox set out above is as follows. Implicit in the challenge presented to your epistemic capabilities by how much you do not understand (that others do) is the idea that *if* you cannot know how much you do not understand what you do not understand, then your epistemic agency is compromised *in toto*. That is, the outer limits of your understanding (and the limits of your understanding of what you understand) threaten the very core of your understanding. If the situation is as extreme as that, then you are not really the responsible epistemic agent you thought you were, capable of making decisions about how to live your life, using common sense, and possessing a wide array of true beliefs on everyday matters. The implicit challenge is therefore along the lines of:

If you do not know *that*, then you do not really know anything.

And this is a social analogue of the traditional skeptical problem, where the challenge is that *if* you cannot know the denial of the radical skeptical hypothesis, *then* you cannot know anything.[92]

What is needed, in my view, is way to block the inference from a failure to understand the limits of one's own epistemic capabilities to the idea that one is possibly radically in error about everything one believes. That is, what is needed is an argument that limits the fall-out from failure to recognise one's penumbra to epistemic chastisement, leading to epistemic humility, rather than radical loss of epistemic self-belief. This problem is both practical and theoretical. Psychologically, it is unsurprising that many people are resistant to adopting new beliefs that imply that not only *were* they radically in error, but that their epistemic apparatus is still deeply unreliable. Theoretically, given a choice between accepting a belief that implies my mechanism for belief formation is deeply and irremediably faulty and rejecting that belief,

[92] This is also analogous to Dworkin's (1988) challenge to Hart regarding the centrality of theoretical disagreement to legal practice. For Hart, hard cases represented the penumbra of legal reasoning, but their uncertainty or indeterminacy did not affect the stable core of settled law. Dworkin argued that these penumbral cases actually revealed deep disagreement as to the settled core of the law. So, analogously, here, the social skeptical challenge is that the inability to identify where one's knowledge *ends* is not just a penumbral issue, rather, it threatens the very core of your settled beliefs, because it turns out you do not really have the epistemic equivalent of a rule of recognition.

it is difficult for me to do anything but the former, since the latter requires me to accept a belief using an apparatus that acceptance of that belief indicates is faulty.

As I will argue, the solution to the social problem of epistemic angst requires delineating the reasoning we use in our day-to-day lives from that used in various expert domains, paradigmatically science. My thought is that securing a *degree* of discontinuity between the two (and thereby preserving the space for common-sense reasoning about how to live one's life) will help secure the space for trust in expertise once more among the groups most likely to doubt it. The social problem of epistemic angst is therefore also one of the *encroachment* of global rational evaluation into the domain of our local rational evaluations and the paucity of support for our individual beliefs, revealed implicitly by how much we do not know, that others do. Alleviating it requires showing how the knowledge asymmetry at the heart of recognising expertise does not diminish our everyday agency. The problem is not so much one of learning to trust experts as it is one of learning to trust *ourselves* in a world governed by expertise.

1.2 Feminist social epistemology

Feminist epistemologists have long argued that many of our social practices undermine the standing of marginalised groups in society's status *qua* knowers. Miranda Fricker, for example, has done much to explain how these practices undermine the other's standing as a knower, which in turn undermines the others' epistemic self-confidence. These can be more or less intentional or explicit. At the most intentional and explicit end, we have practices like those represented in the film *Gaslight*, the well-known cinematic representation of the eponymous phrase. Much of how society generally – and our knowledge-exchange practices specifically – are structured exploit or exemplify the undermining of people's epistemic access to the world, in ways that discredit them as knowers and, at worst, lead to total epistemic paralysis. In my view, these phenomena exemplify the broader phenomenon of epistemic angst I am attempting to get into view in this paper.

At the paper's outset, I noted that epistemic angst was a phenomenon that tied together and could help us explain many disparate phenomena. There is something strange – and *prima facie* objectionable – about linking the experiences of victims of epistemic injustice (Fricker), epistemic coercion, gaslighting (Abramson, 2014) and self-gaslighting (Dandelet), with the experiences of anti-vaxxers and climate skeptics. First and foremost, this is because the subjects of the former experiences are victims of injustice, while the latter are not. Of course, one way of explaining the parallel would be to note that one can *experience* victimisation or *feel* that

one is the victim of injustice without actually being victimised or any injustice being perpetrated. And, of course, I suspect that something like this is often true: in many weird corners of the internet, there are people from oppressor classes claiming to be victims, oppressed by the oppressed. Since experiences and emotions have an ineliminable subjective element to them, it might be enough to note that sometimes epistemic angst attends the experiences of people who mistakenly feel themselves to be epistemically – and otherwise – disenfranchised or discriminated against.

But that is not the whole story. Feminist epistemologists are, of course, right that the feelings associated with radical self doubt – the phenomenon I am calling epistemic angst – can come about as a result of unjust practices. But there is a more general phenomenon here to be interrogated, one that can come about through the epistemic dependence of the general public on expertise and the resultant division of epistemic labour. That is, the *source* of epistemic angst need not be unjust. I say much more about this in Part II.

1.3 Epistemic anxiety

Recently, a number of philosophers have argued that epistemic anxiety is a good thing. The basic idea is that state anxiety about our present lack of understanding of a state of affairs serves as the impetus to seeking an inference to the best explanation and is therefore a trigger for epistemically virtuous deliberation (Cabrera, 2021) or as a helpful emotional response to epistemic risk (Newton, 2022). These recent contributions build on earlier work by (Nagel, 2010) and (Vazard, 2018, 2021).

In broad outline, I agree that epistemic anxiety, understood to involve concern regarding what one knows or understands about a particular state of affairs, topic or choice is a potentially helpful practical and theoretical *motivator*. But it is therefore important to distinguish epistemic *anxiety* from epistemic *angst*, where the latter involves a global negative feeling regarding one's epistemic position, standing or capabilities *in toto*. Far from being specific, useful for inquiry and motivating, epistemic *angst* is global, detrimental to inquiry and practically and theoretically debilitating. While my focus is on epistemic angst, I want to note now (and return to the point later) that epistemic anxiety can lead to epistemic angst. Some lines of inquiry – or anxiety concerning how little we understand about some state of affairs – can generate more global, reflexive anxiety concerning our epistemic capabilities generally. At this point, epistemic anxiety is not, I argue, a good thing, nor is it helpful to our inquiries. In my view, therefore, epistemic anxiety is not an unequivocally good thing: to the extent it is a

catalyst for epistemic angst, it is bad. But with the right antidote to epistemic angst in hand, its helpfulness is assured. I aim to provide the beginnings of that antidote in the rest of this paper.

PART II: EPISTEMIC ANGST AND EXPERTISE

In this Part, I provide a worked example of epistemic angst's utility in philosophical explanation of a contemporary social epistemological problem, *viz.*, mistrust of expertise.

2. Expertise, identifying experts, and the erosion of competence

In this section I introduce an additional factor explaining contemporary mistrust of experts, namely, a failure to identify whether a given topic is one that *calls for* expertise.

A focal point of the social epistemological literature on experts is answering the question how a novice can identify which expert(s) to trust in cases where putative experts disagree. This is often referred to as the 'novice/2-expert problem'. The problem arises because, *qua* novice, the non-expert is not well-placed to assess the expertise of the experts, yet this seems to be precisely what is needed to tell which to trust. The literature then parses into various attempts at explaining how novices can nevertheless obtain a warrant, often indirectly, for believing one expert or group of experts over another or what to do when they cannot. A contemporary application of this literature has been in discussions of the erosion of trust in experts – especially scientific experts – in the last decade or more, on issues like climate change, vaccinations and 5G telecommunications and the like.

The problem this literature identifies is that novices cannot identify whom to believe: very broadly, either scientists from the establishment *or* the so-called 'merchants of doubt' prominent on Fox News, among politicians, and, of course, on the many platforms the internet affords them. This is supposed to be an iteration of the novice/2-expert problem, since the novices in question, themselves very often lacking the educational opportunities that might have engendered respect for the scientific method and the establishment governed by it, are ill-placed to assess whether that establishment really is a consensus on the facts as ascertained by the best methods available to us or instead a sinister, biased or otherwise mistaken echo-chamber of liberal elites. One oft-noted additional factor in this dynamic, about which I say more below, is then that because these issues are highly politicised, where one stands on them becomes a matter of political identity and therefore apt to be distorted by identity protective cognition or outright manipulation. This is then compounded by the existence of cognitive islands, echo chambers and bubbles.

But even allowing for these compounding factors, the novice/2-expert characterisation of the contemporary crisis of faith in experts is at best a partial characterisation. Just as important is identifying whether a given topic is one that *calls for* expertise at all. That is, whether one *needs* to listen to any expert at all or can, instead, make up one's own mind. Let me explain.

There are many topics on which most of us feel we do not need expert advice, either because it is not the kind of topic that requires expertise or because, to the extent that it does, we are experts. Such topics might include which breakfast cereal to choose, basic arithmetic or geography, how to be good friend, how to do our job, how to raise our children or look after our parents, how to deal with conflict at work or with neighbours, romantic relationships or, more generally still, how to structure our days – from eating, sleeping, exercising, to relaxing, spending time with our family or working. The topics on this list comprise a mixture of straightforwardly empirical matters and those requiring the exercise of one's practical judgment: and most of us do not distinguish between these in our everyday lives. Topics on which most of us know we are not experts *and* on which we would happily acknowledge requiring expert advice include topics like medicine, law, accounting, mechanics, robotics, engineering and architecture.

What distinguishes these two lists of topics? One tempting answer is that the latter list enumerates accredited professions requiring a high degree of specialised knowledge and skill, while the former lists non-expert topics that do not. This characterisation is not entirely accurate, however. The topics on the first list are all also topics on which we have experts in the contemporary world: dieticians, developmental psychologists, geriatricians, sleep scientists, relationship counsellors, management consultants, personal trainers, occupational therapists, life-coaches, and dispute resolution arbitrators all have expertise that we do not, bearing directly on the topics contained in the first list. To be such an expert also requires a high degree of knowledge and skill not possessed by the rest of us. A better way of distinguishing the two lists might be to say that the former enumerates topics that we expect competent adults to handle by themselves in the ordinary course of events, while the latter contains topics on which expertise is *required*.

Sharpening this distinction, the former list contains topics on where it is not obviously negligent or imprudent to attempt to act without expert advice, while the latter contains topics where it is. We are expected to be able to repair relationships, but not cars, for example. We *might* ask for expert advice concerning topics on the former list if we encountered significant problems or if we hoped to gain some advantage or attain some outcome we felt unable to

achieve by ourselves, but we otherwise feel these topics to be within our sphere of competence. They are topics on which, one might assume, it is not unreasonable to suppose that, in deciding for ourselves or those entrusted to our care "we know best".

But this assumption is unjustified. If you and a dietician disagree about which breakfast cereal is best; or you and a developmental psychologist disagree about how many hours of sleep your child needs, you should defer to the expert's judgment. And many relationships – romantic, workplace, friendship – might have been saved by the advice of counsellor or mediators, had it been obtained sufficiently early. Note, moreover, that to the extent you ignore – or do not bother to ascertain – developmental psychologists' advice on child-rearing or educational experts' advice on education, you run the risk of being a far worse parent than you might have been – and, perhaps, as guilty of harming your child through your wilful ignorance and negligence in the process. Likely, this last suggestion strikes you as offensive. You do not *need* a shrink to tell you what is *best* for your child. The topic does not *require* specialists' expertise.

I think this reaction is entirely understandable and justifiable, though strictly speaking I think it is also often false. It is understandable and justifiable because we would likely be practically paralysed if we sought expert advice on every topic on which it is available.[93] It is strictly speaking false because the specialisation and compartmentalisation of expertise on every topic in the modern world likely means that while you very likely perform adequately in your role as jack-of-all-trades; as master of none, you likely do not perform *best*.

I label this phenomenon 'the erosion of competence'. Every moment of every day, you become *relatively* a little less competent (and relatively more of a novice) in almost every sphere of your life, as a result of your entirely justifiable ignorance of what the growing body of expertise on each of those topics says. This thought is unsettling and, for me at least, anxiety-inducing. While we are largely able to ignore it, it rears its ugly head when we hear disagreements in parents' group about sleep-training; or the benefits of early childcare *versus* home parenting; or when we are told breakfast is not actually – and never was – the most important meal of the day; that Pluto is not a planet; or that microaggressions are a real thing.

These revelations are anxiety-inducing precisely because they imply that we are far less competent in our alleged sphere of competence than we thought we were. We watch as that competence erodes inexorably from the combined forces of expertise, specialisation and complexity and wonder if the headline message coming from expert corners is that we really

[93] Which, in the modern world, is almost every topic.

do not know anything and should just shut-up and do as we are told. Given the relative knowledge asymmetry between us and them and the possibility for third-party and collective harms, it is sometimes difficult to disagree.

The erosion of competence is a source of epistemic angst that affects all of us, to differing degrees. But it is particularly significant in scenarios involving disagreements concerning whether a topic belongs on the first or the second list. That is, when people disagree about whether a topic is one where expertise is *required* to adequately act on or understand the topic at all.[94]

It is at this point that we can re-introduce the novice/2-expert problem. For while there are undoubtedly many instances of competitor merchant-of-doubt "experts" attempting to convince the gullible that climate change is not real or vaccinations are unnecessary or harmful by directly peddling their own analysis of the evidence, this is not the only, and perhaps not even the most prevalent, way in which doubt concerning the deliverances of expert communities is manufactured or spread. Instead, the catchcry of climate-vaccine-5G skeptics on twitter, as elsewhere, is 'Do your own research!', sometimes attended by graphs of the colder than usual winters in the past decade or troubling anecdotes about the effects of vaccines on loved-ones. These skeptics are the unlikely inheritors of the Enlightenment's '*Sapere Aude!*'. In the background of this discourse is the idea that some of the topics the establishment are trying to convince you belong on List Two, lying outside the area of your possible understanding or competence, in fact belong on List One: they do not require specialist training or knowledge to understand.

This introduces a pernicious dynamic into this variant on the conflict between "experts" in the novice/2-expert problem. Instead of two experts offering competing claims about the subject matter of the dispute, as in the traditional set-up of the problem, we instead have one expert – "the establishment expert" – telling you what to believe about the subject matter in question, for example, climate change; and a second expert – "the anti-expert" – telling you that you should do your own research, rather than listen to the other expert. That is, while one expert is telling you to listen to them, the other expert is telling you that you are capable of informing yourself about these matters and should do so, given that they are important and affect how you and others should live your lives. This is an empowering message. The anti-expert speaks to the concern – one I have said we all share – that we are incapable of working

[94] As opposed to those where expertise is not required at all or is only required to optimise.

out what to do without relying on experts' say-so. The anti-expert, in this dialectic, becomes the champion of intellectual autonomy and individual responsibility, while the establishment expert tells you to fall in line. The problem is not how to identify which expert to believe or, not exactly. The problem is to decide whether to believe one expert who tells you that you cannot understand this topic by yourself and another who is telling you that you can.

That is, the anti-expert exploits our *epistemic angst*; and it is in part because epistemic angst makes people vulnerable to such exploitation that the phenomenon is worth investigating.

3. Epistemic angst, experts, and interpersonal doubt

I have argued that angst concerning our good epistemic standing in our community goes to the core of how we perceive ourselves: as competent agents capable of making sensible decisions about how to live our lives. In this section, I argue that epistemic angst helps explain some instances of identity protective cognition. I propose an apparently attractive solution to the problem, but – in a theme that will recur in my explanation of all three cases of epistemic angst – I then argue that the solution risks catalysing yet further and deeper epistemic angst.

The starting point for the problem is as follows. It is well documented that for some groups in society, climate denialism, 5G conspiracy theories or vaccine skepticism, for example, have become important and self-defining traits. Explanations of resistance to apparently clear expert consensus that these views are false have varied, but many have pointed to the fact that once the beliefs, values or theories have become an important part of a group's identity and the individuals in the group value that practical identity, they will be highly resistant to evidence, testimony and explanations that they would, in other circumstances, find highly compelling, and of the form that they do find compelling in other parts of their lives. This identity protective cognition lead to various forms of motivated and otherwise biased reasoning; generates domain-specific confirmation bias and mistrust of experts; and often involves impugning the motives of the experts in question.

An important and sometimes underappreciated aspect of the internal logic of these groups is the entirely human and altogether natural inclination not to be proven wrong. The more fervently you hold some belief; the more valuable you take something to be; the more time, effort and social capital you have invested in investigating and then promoting your view; the more embarrassing it is to be shown to be mistaken. Recanting is a very public demonstration of our inadequacy as a tracker of truth or value. No one likes to be called-out in this way, because it is embarrassing and diminishes one's social epistemic capital in the future. Equally important is the damage it does to one's epistemic self-confidence. Many of us will

have had the experience, at some point in our lives, of coming to realise that we were mistaken in having believed so fervently in something or someone. The natural response to this revelation is to feel embarrassed; sometimes even ashamed; and to wonder what else you could be wrong about, if you were so wrong about this. That is, confronting serious error can lead one to doubt one's own good judgment. These emotions will be familiar to anyone who has been through a bad break-up, lost their religious faith; or bought into an investment scheme that goes belly-up; but many of us will have other examples varying from the amusing-in-hindsight to the tragic.

The point of these observations is that revelations of this kind very naturally cause us to doubt our own epistemic capacities: our ability to apprehend and judge well, and to respond well to the evidence and makes decisions considering it. They are, as a result, usually quite painful and in hindsight we often recognise that we failed to weigh the evidence correctly; allowed our emotions and biases to interfere with our good reasoning; and had been misleading ourselves for some time before the revelation, putting off facing the challenge to our good epistemic standing and our sense of ourselves as a reliable tracker of the truth. This will be particularly true, for example, of cases where we have made our commitment to the now-known-to-be-mistaken view (or person) public or a significant plank in our identity. So, of course, we are strongly motivated to resist the force of evidence that suggests that we are badly wrong about something important to us. Not just because, for example, it will mean that we will have to leave a group that we value and that has given us a practical identity that is very important to us, but because it will involve admitting that we were *wrong*. And not just wrong, but wrong in a way that casts doubt (for ourselves and for others) on our credentials *qua* knower.

This is a very important aspect of the phenomenon I have labelled 'epistemic angst'. It is a *generalised* anxiety about our epistemic capabilities that threatens our sense of self generally: because it is crucial to our sense of self that we are generally capable of tracking the truth and making responsible decisions in its light. If we are unable to rely on our own good judgment, then it our autonomy is threatened from within. Ironically therefore, doubt concerning our *own* epistemic capacities therefore partially explains our doubt concerning the epistemic capacities of *others*; as it generates domain-specific confirmation bias; mistrust of experts; and impugned motives of the out-group.

Let me propose an initially appealing solution to this problem, one that will illustrate the perils of epistemic angst – and its quicksand-like quality – quite well. One apparently attractive strategy for alleviating the social problem of epistemic angst is as follows. We should

promote the idea that there is nothing wrong with being wrong. That is, being wrong does not discredit you as a knower or agent: *it happens to the best of us*. There should be no shame. You were not "beaten" by those holding the opposing view or *outsmarted* by those who took advantage of your epistemic trust. Ascertaining the facts is not a competition for knowledge or acclaim. It is – for cognitively limited creatures like ourselves – a joint, inevitably error-ridden and deeply imperfect enterprise.

The proposal is to promote the view that being wrong does not discredit you as a knower or agent by pointing out that it happens to the very best of us which, in the epistemic case, means *experts*. The strategy is therefore to note that experts often get things wrong; disagree with each other all the time; and that many of the things on which there is now expert-consensus will be rejected as false by future generations of experts. If even experts very often get things wrong, then surely there is no shame in the rest of us getting things wrong just as often. This should help alleviate our epistemic angst, making it easier for all of us to admit when we were wrong and, one might hope, make groups that have invested heavily in false views about vaccines, climate change, 5G, wind turbines or lizard-people less resistant to good evidence and reasoning. Why? Because a very significant impediment to constructive dialogue in such cases is the deep entrenchment of the views of both sides. The deeper the entrenchment, the deeper the epistemic sunk cost. But if we can reduce the *epistemic switching costs* associated with admitting you were wrong, we remove a very significant barrier to engagement. That is, we can ensure that failure not to change one's mind, given one's prior epistemic, emotional, evaluative and temporal investment really will be irrational, *qua* sunk cost fallacy, rather than, as present, looking close to *practically, subjectively* rational, given one's present position and what one stands to lose – in terms of epistemic self-esteem – in the current political epistemological climate.

But the catch is that promoting that very idea, *viz.* that experts get things wrong all the time too, can lead to the caustic interpersonal doubt against which promoting the idea of widespread fallibility was supposed to provide a bulwark. First, because it can lead to doubt concerning expertise. Second, because it threatens our faith in the good epistemic standing of our community. Third, because it devalues truth's already deflated currency. If there's nothing wrong with being wrong, then why care so much about getting things right?

The social ill of epistemic angst currently in focus manifests among groups who are distrustful of experts. So, promoting the idea that they are also highly fallible is just as likely

to be fuel to the fire of those already skeptical of expertise.[95] Using expert disagreement or error to ward off non-expert's angst concerning their own fallibility risks a form of *levelling-down* problem, by way of a dilemma. Either some groups in society – likely the very groups worst affected by the ills associated with epistemic angst – come to believe that experts really are, after all, no better placed than they are in assessing the evidence and making policy recommendations as a result, for example, *or* they come to understand that experts really are much better placed than they are, even though experts are still highly error-prone. This latter view is preferrable, but if epistemic angst is the problem, then demonstrating that ordinary people are in a far worse epistemic position than they thought they were previously, both absolutely and relatively, *vis-à-vis* experts, is unlikely to be a particularly effective solution. The risk is that the targets of the proposal now lose faith in themselves as well as experts.

This is the first of three cases that, I claim, demonstrate the difficulties associated with epistemic angst and the importance understanding it. This is not the sort of epistemic angst that grips an individual – likely a philosopher – concerning their inability to rule out radical global skeptical hypotheses, nor is it the sort of potentially helpful epistemic anxiety associated with events that seem to call for explanation or recognition of the importance of deliberating carefully on some high stakes question.[96] This is a social, interpersonal problem, arising because in public debate being right and wrong is an epistemic *zero-sum* game. There are winners and losers, and no one likes to think of themselves as a loser.

4. Common sense, alienation and epistemic paternalism

My hypothesis is that it is difficult to understand and solve the problem of mistrust of experts among, for example, climate change and vaccine skeptics without understanding the role played by epistemic angst. While these are particularly obvious cases where epistemic case causes trouble, they are, I suggest, public manifestations of a more general phenomenon. In this section, I provide a further illustration of this point, one which also demonstrates the risk posed by ignoring epistemic angst *and* the risks posed by acknowledging it. As with the previous section's point, the idea will be that apparent overconfidence on the part of non-experts is fuelled by underlying epistemic angst, but that confronting this angst risks precisely the catastrophic loss of epistemic self-esteem the behaviour in question helps to avoid.

[95] Cf. (Dellsén, F. 2018; 2020); (Estlund, 2008); (Hazlett, 2015); and (Roussos, 2021).

[96] Cf. (Wright, 2021).

One apparently strange aspect of some individuals' and groups' mistrust of experts on matters relevant to their expertise is that people often use expert's explanations of the phenomena in question "against them". That is, they will unpack the experts' explanations of what is happening and use the facts as the expert has presented them as the basis for an alternative hypothesis that accounts for them: whether it be a rise in temperatures, rapid spread of a virus, or an apparent increase in viral case numbers following the instigation of a lockdown policy. That is, it is common for non-experts to apply their own common-sense reasoning to highly complex scientific, social or political problems, *granting that the facts are as the experts claim*. The reason this seems strange is, of course, that if the groups in question agree that the experts are best-qualified to ascertain the facts, then it seems odd that their qualifications do not also justify a degree of deference to their explanations of them.

But the allure of common-sense reasoning applied to complex problems not suited to it is extremely widespread. Worse still, it is in part encouraged by the very experts whose application it sometimes undermines. Worse still, it is difficult to see how this situation can be avoided without risking deepening the epistemic angst at its source.

Every time you pick-up a newspaper and appraise the government's tax policy, or a proposal for emissions reduction, or a tweak to immigration law, or weight the competing claims of experts concerning the latest foreign policy posturing in the South-Pacific, you use your common sense to appraise matters that are very likely far too complex to be analysed by it. But this is an unavoidable aspect of what it is to be a responsible, autonomous epistemic agent, let alone engaged citizen. Without activity of this kind, we would have to live in a very circumscribed sort of epistocracy (Estlund, 2008). But following (Watts, 2012), common sense analyses of complex social, economic and policy issues tend to mislead us into thinking we understand far more about the world and human behaviour than we really do. Even among experts, attempts to predict, manage, or manipulate social and economic systems often go awry, especially when relying on common sense to analyse complex macro phenomena heavily influenced by compounding effects of system-wide small-scale interactions.

The use of common sense to understand complex, large-scale problems comes very naturally to us. This is not just because common sense reasoning serves us very well in our day-to-day lives, but because recognising that we are very poorly equipped to understand much beyond them would be, for many of us, not only deeply damaging to our epistemic self-esteem but potentially epistemically paralysing. How ought one vote in the next election if the technical matters concerning the effects of climate, immigration, monetary, taxation and foreign policy (to mention but a few) are beyond one's epistemic competence to assess?

Of course, to facilitate public discussion of the problems that face society, it is therefore *essential* that experts attempt to convey their knowledge and proposals in terms that are not only understandable but amenable to common sense reasoning (John, 2018). But communicating the facts and policy proposals resulting from them *sustains* the illusion that, for example, complex macroeconomic problems, and the relative merits of proposed solutions to them, are comprehensible without understanding the mathematics behind the relevant models. The basic problem is that common sense is not really *scalable* and most of our collective social problems are large in scale.

Once again, this gives rise to a dilemma: either experts provide common sense explanations of complex phenomena so that the public can feel they understand and therefore evaluate social, scientific, economic and scientific problems and the policy solutions to them (e.g. climate change, vaccination, monetary policy) or they don't, and we all just rely on experts' testimony about what to do instead. But if they do, they alleviate epistemic angst only by distorting the complexity of reality and risk licensing faulty common-sense inferential reasoning, of the sort prominent among those individuals and groups that develop their own theories to compete with the experts'. But if they do not, they and governments risk alienating the public by adopting a paternalistic stance towards them as knowers, with an inevitable increase in suspicion regarding both experts and the government, together with yet more epistemic angst. No one is likely to be satisfied by the answer that 'the ultimate question of life, the universe and everything' is 42 (Adams, 1979). But can the answer, and the reasoning behind it, be approximated in a way that is not misleading?

Angst concerning our own epistemic capabilities is a natural response to the bewildering complexity of the world. Whether this manifests in attempts to make the world more comprehensible by making it simpler than it really is or in significantly downgrading our level of confidence in our own epistemic standing to the point where we no longer feel equipped to engage in evaluation of experts' competing proposals, the risks associated with epistemic angst loom large well beyond a healthy dose of epistemic humility, reverberating through our practical, deliberative lives. Epistemic angst is a social problem: to live well together, we must both admit we know very little and are, individually, very cognitively limited appraisers of expert testimony, and yet we must work *together* to decide what to do with that testimony. Before offering the beginnings of a proposal, let me sketch one more, related problem, with epistemic angst at its foundations.

5. Natural Is/Ought detectors, novel problems and skepticism about expertise

Let me begin this section by sharing a hunch, related to the previous section's discussion of problems posed by using common sense to analyse complex problems. The hunch is that people have a very good intuitive grip of the is/ought gap; and are intuitively skeptical of moves from the descriptive to the modal. That is, one reason many people do not take experts to be better placed to pronounce on policy issues than they are, *especially* once the expert has done "their job," namely, ascertaining the facts, is that many people are skeptical that once those technical facts have been established, we require technical knowledge to work out what to do in light of them. This is, I suspect, a pernicious blend of our desire to believe our practical reasoning capabilities are more refined and of wider application than they really are, and a (largely correct) sense that there is a significant gap between working out what's going on and working out what to do about it, where the skills and knowledge required to perform the former task are not necessarily identical to those required for the latter. The problem seems to be the temptation to think that the epistemic division of labour therefore lies in experts possessing the former, and our possessing the latter, skills.

When novel problems confront society (e.g. climate change, Covid), many people are skeptical that experts' understanding of what has happened so far or in related circumstances qualifies them to predict what will happen or draw normative conclusions about what we *should* do. So, while they might agree that scientists are detecting unprecedented changes in temperature or other climactic conditions, they disagree that this gives scientists a superior position from which to deliberate about what to do as a result. Once we understand that the world is warming, say some portions of society, for example, many conservative governments, what to do about it and who to blame for it are policy questions, and policy questions are for the general public acting through their politicians, not scientists. This problem dovetails with the common-sense problem above. There is a tendency to deny that experts have privileged epistemic and normative standing once facts are established. One natural explanation for this phenomenon is that doing so threatens the status of our "un-trained" reasoning. That is, recognising experts' superior epistemic and normative standing would then involve admitting that experts are not just better informed but also *better at deciding what to do* considering the information available. The explanation for this is, of course, that expertise does not just consist in possession of "more facts" but in superior judgments about what to do, given those facts. There is, of course, a compounding problem**,** which is that there is plenty of evidence to suggest experts are not very reliable when it comes to predicting, managing, and manipulating complex

social systems in precisely the way making policy decisions about large scale societal problems requires, for example.[97]

In a now familiar theme, one obvious apparent "solution" to this problem is to convince the general public that the experts, despite their error-rate, are still doing vastly better than most of us would. But, of course, doing so will require employing further modal reasoning and the relevant counterfactuals – for example, explaining what would happen, were the alternatives proposed by lay-people adopted, for example – would need to be explained and analysed by experts. Note, moreover, that even if this claim could be defended in a satisfactory way, doing so would likely further entrench high degrees of epistemic angst. After all, most of us live in a world in which we either think we have a good handle on what's going on and what to do about it or believe that experts do. Finding out that both propositions are false will benefit very few, one might think. Once again, epistemic angst looks like it plays a significant role. Many people are reluctant to believe experts because it threatens their sense of their own good epistemic standing, but convincing them otherwise involves revealing to them how poor their epistemic standing really is. And if they were already resistant to expertise because of the implicit violence it did to their sense of themselves as reliable knowers, making it explicit is unlikely to make it more palatable.

PART III: SOLVING THE SOCIAL PROBLEM OF EPISTEMIC ANGST

6. Case study: expertise

It is sometimes said that the first step to solving a problem is admitting that you have got one. We have identified epistemic angst as a social problem. Have we, thereby, made headway on solving it? In this section I present the beginnings of a strategy. It is not a complete solution – since the problem is both theoretical and practical, its solution will also require practical elements far beyond my competence to examine – but it does lay the groundwork for the theoretical solution. The strategy comprises two components: one addresses angst arising from entrenchment, the other addresses angst arising from the erosion of competence.

Entrenchment first: at the outset, we noted that a significant source of epistemic angst can be the entrenchment of beliefs core to a valued social or political identity and definitive of membership in groups that are very important to us. So much is uncontroversial. But an underappreciated aspect of the entrenchment of these beliefs, I argued, is the damage that

[97] Cf. (Watts, 2012) and (Kahneman 2011).

would be done to our sense of ourselves as reliable *knowers* were we to admit to ourselves or others that we were badly mistaken about the views we held and that had come to be self-defining. After all, if we were so badly wrong about these views in which we were deeply invested, our sense that we can rely on our own judgment is now damaged. Implicit in this pattern of reasoning – and sometimes explicit in less-than-helpful critiques of "deplorable" worldviews and beliefs – are inferences conforming to the following pattern:

If you do not know *that*, then you do not know anything.

Where the 'that' in question stands for some fact or facts upon which there is scientific or establishment consensus that is denied or doubted by the person being criticised. As noted earlier, inferences conforming to that pattern are commonplace in defences of radical global skepticism, where the 'that' in question substitutes for propositions concerning brains in vats; whether your hands are in front of you; whether they disappear when you stop looking at them; evil demons; new evil demons; and new new evil demons (Lord, 2018).

The social problem of epistemic angst also arises where one's alleged failure to know something challenges one's epistemic credentials *tout court*. But rather than the proposition in question being a foundational one concerning your common sense or everyday experience, the propositions in question tend to be more complex propositions of scientific or otherwise technical nature. While this might have seemed to make such an inference illicit from the outset, the reason the criticism bites as it does is because it impugns the use of everyday common-sense reasoning or experience in attempts to disagree with scientific consensus. That is, the criticism bites because its target does not understand the limits of their own reasoning and perceptual powers. In not understanding the uses to which these powers can legitimately be put, the object of criticism exposes a much deeper misunderstanding about the nature and limits of their understanding more generally, which in turn calls into question the reliability of their understanding generally. Their failure to know when common-sense reasoning is not applicable calls into question their ability to engage in common-sense reasoning generally.

My proposed strategy for blocking this inference is inspired by Prichard's use of Wittgenstein's response to skepticism to defeat epistemic angst of the more traditional, radical and global cartesian variety. The key to solving that problem was to understand the limits of rational evaluation. By understanding that all rational evaluation was necessarily local and that, therefore, moving from local to global rational evaluations was an impermissible move, Prichard argues we can block the skeptical inference from 'if you do not know *that*, then you do not know *anything*.' What is the equivalent for epistemic angst as a social problem?

The basic idea is to block the inference from the fact that someone misapplies their common sense to the conclusion that they are incapable of using it correctly in the spheres in which it does apply. The structure of the solution is also in some ways similar to acknowledging the inherent locality of rational evaluation: we note that common-sense reasoning also has its limits. It cannot be used globally but is rather appropriately confined to the everyday topics in which it has proven successful. This limitation also insulates it from the failures of its application in domains to which it is ill-suited. That our common-sense misfires when applied to technical or scientific issues is not evidence that it misfires generally, it is simply part and parcel of the fact that these domains of inquiry are not amenable to common-sense reasoning. The final piece of the puzzle is therefore how to handle our failures to recognise these limits.

My proposal is straightforward: we acknowledge that the question as to the limits of the application of one's common sense is not itself a common-sense question. It is a technical, meta-cognitive one that requires specialist education and training to answer. By and large what we now recognise as its limits were not intuited by or divined using common sense, rather, the limits were established by experts who showed us where our common-sense reasoning went awry. But that just means that the inference from the failure to recognise where the application of one's common sense is inapt is not itself a failure of common sense, which is enough to block the relevant inference. Our capacity to recognise *that* some issue is actually one requiring expertise does not diminish our standing as an everyday knower, any more than our inability to determine which of two or more disagreeing experts does. The boundaries of common sense are a matter of *expertise*. There is therefore no shame in a novice failing to recognise that some matter is not one their common sense can treat *in circumstances where merchants of doubt are suggesting otherwise*. My strategy for solving this problem therefore ties it to our efforts to solve the novice/2-expert problem without reducing the former to the latter. We need experts to help us ascertain the boundary between common sense and expertise. This makes sense. Physicists consistently tell us things about the universe which fly in the face of common sense and everyday perceptual experience. The conclusion to draw from this is *not* that we cannot rely on common-sense generally, but rather that physics is not the sort of topic that can be investigated using our common sense or everyday perceptual experience as evidence.

That is my theoretical solution to the first half of epistemic angst as a social problem. Of course, practically speaking, one would have to convince both critics and their disputants of these theoretical propositions in a way that really *hit home*. Angst is not a theoretical problem (though theoretical problems can generate and underpin it). It is an emotional and practical one.

The second half epistemic angst as a social problem is the phenomenon I labelled 'the erosion of competence.' I distinguished between List 1 and List 2, where the former list contained topics like how to raise your child, resolve conflicts, eat or sleep, that do not *require* expertise, and the latter list (e.g. law, accounting, physics, mechanics) that does. A significant source of epistemic angst, I argued, was the lengthening of List 2 and the shortening of List 1: in the modern world, it can seem that we are no longer competent to do *anything* without being told by experts either how to do it better or that we are doing it wrong. That is, List 1 has collapsed into List 2, and our sphere of competence has collapsed with it.

But now note that no one feels angst at the prospect of epistemic outsourcing to lawyers, accountants or mechanics on the topics relevant to their expertise. These are simply not topics on which we are expected to know everything all by ourselves. Part of the reason we have experts on these topics is that the body of human knowledge has now grown enough that *no one* knows it all. Lawyers need help from mechanics who need help from accountants who need help from relationship consultants and there is nothing the least bit surprising about this compartmentalisation of knowledge: it is a trend that has been developing for the last 1000 or more years. While it can sometimes feel like there are just two epistemic classes in the world – experts and novices – the reality is that everyone is a novice on most topics and an expert on very few. The erosion of competence is, or at least, should be, both liberating and affirming. Once you are no longer *expected* to know these things, there is no shame in not knowing them. That is, the more expertise encroaches into our day-to-day lives, the less you are expected to "know best" without seeking assistance. Intellectual humility turns out to be the perfect salve to epistemic angst; one that does not diminish your standing *qua* knower at all.

Once you are freed of the burden of having to know all this stuff all by yourself, life gets easier. And once the mark of wisdom and being a capable of knower is admitting that you do not know and are wrong about much of what you thought you knew – the bar for being an exemplary knower is significantly *lowered*! That is a big win, both theoretically and practically.

Note, moreover, the effect widespread acceptance of this approach could have on the pernicious expert/anti-expert dialectic we encountered earlier. The demagogue, merchant of doubt's power lies in preying on people's fear that they really should know certain things and then convincing them that they do! If people instead know that there is nothing wrong with not knowing, then the power of those telling them they should and do is significantly reduced. Now it is the merchants of doubt telling us we fail some standard imposed from outside, instead of the liberal elites. That seems like progress. And that is my solution to the theoretical second

half of epistemic angst as a social problem. Convincing everyone that this is true is another, practical problem. I leave it to the experts.

Problems will, of course, remain. Not least, the novice/2-expert problem. But approaching the novice/2-expert problem with a clear sense of the role played by epistemic angst, the limits of common-sense reasoning, and the humility to accept that recognising one's general, relative incompetence is a virtue, will help.

Conclusion

The aim of this paper was to investigate a phenomenon I called the 'the social problem of epistemic angst'. Its main purpose was exploratory: to establish that there is a unified phenomenon worth further investigation. In Part I, I provided a working definition of epistemic angst and explained how further investigation into the phenomenon might help us better understand a variety of philosophical and social problems. Part II comprised a worked example of epistemic angst's utility in deepening our understanding of existing philosophical and social problems. I demonstrated the role epistemic angst plays in mistrust of expertise and used it to diagnose a novel aspect of this problem fuelled by what I labelled 'the erosion of competence.' In Part III, I sketched a partial solution to the problem of epistemic angst as it manifests in mistrust of expertise.

References

Abramson, K. 2014. "Turning Up the Lights on Gaslighting." *Philosophical Perspectives*, 28, pp. 1–30.

Adams, D. 1979. *The Hitchhiker's Guide to the Galaxy*. London: Pan Books.

Adler, J. 2005. "William James and What Cannot Be Believed." *Harvard Review of Philosophy*, 13, pp. 65–79.

Anderson, E. 2000. "Beyond homo economicus: new developments in theories of social norms." *Philosophy and Public Affairs,* 29, pp. 170–200

Anscombe, G. E. M. 1963. *Intention*. Second edition. Oxford: Blackwell.

Appiah, K. 2018. *The Lies That Bind: Rethinking Identity*. London: Profile Books.

Audi, R. 2001. "Doxastic Voluntarism and the Ethics of Belief." In M. Steup (ed.), *Knowledge, Truth, and Duty*. New York: Oxford, pp. 93–111.

_____. 2008. "The Ethics of Belief: Doxastic Self-control and Intellectual Virtue." *Synthese*, 161, pp. 403–418.

Arpaly, N. 2003. *Unprincipled Virtue: An Inquiry into Moral Agency*. New York: Oxford University Press.

Arrow, K. 1963. *Social Choice and Individual Values*. 2nd edition. New Haven: Yale University Press.

Baker, D. 2018. "Skepticism about ought simpliciter." In R. Shafer-Landau (Ed.), *Oxford Studies in Metaethics* (Vol.13, pp. 230–251). New York: Oxford University Press.

Benton, M. 2016. "Expert Opinion and Second-Hand Knowledge." *Philosophy and Phenomenological Research*, 92, pp. 492–508.

Berman, M. 2019. "Of law and other artificial normative systems." In D. Plunkett, S. Shapiro & K. Toh (Eds.), *Dimensions of Normativity: Essays on Metaethics and Jurisprudence* (pp. 137–162). New York: Oxford University Press.

Bicchieri, C. 2006. *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge: Cambridge University Press.

_____. 2017. *Norms in the Wild: How to Diagnose, Measure, and Change Social Norms*. New York: Oxford University Press

Black, D. 1948. "On the Rationale of Group Decision-making." *Journal of Political Economy*, 56, pp. 23–34.

Blackburn, S. 1993. *Essays in Quasi-Realism*. Oxford: Oxford University Press

Bokros, S. 2021. "A deference model of epistemic authority." *Synthese*, 198, pp. 12041–12069.

Borda, J.-C. de. 1784. “Mémoire sur les élections au scrutin par M. de Borda” in *Mémoires de l’Académie Royale des Sciences année 1781*, Paris: l’Imprimerie Royale, pp. 657–665

Bradford, G. 2015. *Achievement*. Oxford: Oxford University Press.

Brandt, R. 1979. *A Theory of the Good and the Right*. Oxford: Oxford University Press.

Bratman, M. 1987. *Intentions, Plans and Practical Reason*. Cambridge, MA: Harvard University Press.

_____. 2000. “Valuing and the Will.” *Philosophical Perspectives,* 14, pp. 249–265.

_____. 2003. “Autonomy and Hierarchy.” *Social Philosophy and Policy*, 20, pp. 156–176.

Brennan, G. & Pettit, P. 2006. *The Economy of Esteem: An Essay on Civil and Political Society*. New York: Oxford University Press.

Brennan, G., Eriksson, L., Goodin, R. E., & Southwood, N. 2013. *Explaining Norms*. New York: Oxford University Press.

Broome, J. 2001. “Are Intentions Reasons? And How Should We Cope With Incommensurable Values?” In C. W. Morris and A. Ripstein (Eds.), *Practical Rationality and Preference: Essays for David Gauthier*, (pp. 98-120). Cambridge: Cambridge University Press.

_____. (2013). *Rationality Through Reasoning*. Oxford: Oxford University Press

Brownlee, K. 2012. *Conscience and Conviction: The Case for Civil Disobedience*. Oxford: Oxford University Press.

Burnett, B. & Evans, D. 2016. *Designing Your Life: How to Build a Well-Lived, Joyful Life*. New York: Knopf Publishing.

Cabrera, F. 2021. “Is epistemic anxiety an intellectual virtue?” *Synthese*, 199, pp. 13471–13495.

Callard, A. 2016. “Proleptic Reasons.” In Russ Shafer-Landau (ed.), *Oxford Studies in Metaethics*, Volume 11, New York: Oxford University Press, pp. 129–154.

_____. 2017. “Liberal Education and the Possibility of Valuational Progress.” *Social Philosophy and Policy*, 34, pp. 1–22.

_____. 2018. *Aspiration: The Agency of Becoming*. New York: Oxford University Press.

Cappelen, H. 2018. *Fixing Language: An Essay on Conceptual Engineering*. New York: Oxford University Press.

Center for Applied Rationality. https://rationality.org, accessed 27 August 2022

Chang, R. 2001. *Making Comparisons Count*. New York: Routledge.

_____. 2002. “The Possibility of Parity.” *Ethics,* 112, pp. 659–88.

_____. 2009. “Voluntarist Reasons and the Sources of Normativity.” In D. Sobel and S. Wall (Eds.), *Practical Reason and Action*, (pp. 243-71). Cambridge: Cambridge University Press.

_____. 2013. “Grounding Practical Normativity: Going Hybrid.” *Philosophical Studies* 164, pp. 163–87.

_____. 2021. “What is it to be a Rational Agent?” In R. Chang and K. Sylvan (Eds.), *The Routledge Handbook of Practical Reason*, (pp. 95-109). New York: Routledge.

Chignell, Andrew, “The Ethics of Belief”, *The Stanford Encyclopedia of Philosophy* (Spring 2018 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2018/entries/ethics-belief/>.

Clifford, W.K. 1877 [1999]. “The Ethics of Belief.” In T. Madigan, (ed.), *The Ethics of Belief and Other Essays*, Amherst, MA: Prometheus, pp. 70–96.

Condorcet, M.J.A.N. de C., Marque de. 1785. *Essai sur l’application de l’analyse à la probabilitié des décisions rendues à la pluralité des voix*. Paris: l’Imprimerie Royale.

Conee, E. & Feldman, R. 1985. “Evidentialism.” *Philosophical Studies*, 48, pp. 15–34.

Constantin, J. & Grundmann T. 2020. “Epistemic authority: preemption through source sensitive defeat.” *Synthese*, 197, pp. 4109–4130.

Copp, D. 2007. *Morality in a Natural World*. Cambridge: Cambridge University Press

Cukor, G. 1944. *Gaslight*. Metro-Goldwyn-Mayer.

Dandelet, S. 2021a. “Doxastic Wronging and Evidentialism.” *Australasian Journal of Philosophy*, (ahead of print), pp. 1–14.

_____. 2021b. “Epistemic Coercion.” *Ethics*, 131, pp. 489–510.

Darwall, S. 1983. *Impartial Reason*. Ithaca, NY: Cornell University Press.

Dellsén, F. 2018. “When Expert Disagreement Supports the Consensus.” *Australasian Journal of Philosophy*, 96, pp. 142–156.

_____. 2020. “The epistemic value of expert autonomy.” *Philosophy and Phenomenological Research*, 100, pp. 344–361.

Dorst, K. 2020. Evidence: a guide for the uncertain, *Philosophy and Phenomenological Research,* 100, pp. 586–632

Dorsey, D. (2017). “Idealization and the Heart of Subjectivism.” *Noûs,* pp. 196-217.

Dovi, S. 2018. Political Representation. In E. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. <https://plato.stanford.edu/archives/fall2018/entries/political-representation/>.

Duijf, H. 2021. “Should one trust experts?” *Synthese*, 199, pp. 9289–9312.

Dworkin, R. 1988. *Law’s Empire*. Boston: Belknap Press.

Dyke, M. 2020. “Could our epistemic reasons be collective practical reasons?” *Noûs*, pp. 1–21.

Elster, J. 1982. “Sour Grapes: Utilitarianism and the Genesis of Wants.” In Amartya Sen and Bernard Williams, eds., *Utilitarianism and Beyond*. Cambridge: Cambridge University Press, pp. 219–238.

_____. 1983. *Sour Grapes: Studies in the Subversion of Rationality*. Cambridge: Cambridge University Press.

Enoch, D. 2005. “Why Idealize?” *Ethics*, 115, pp. 759–787.

_____. 2011a. “Giving practical reasons.” *Philosophers’ Imprint*, 11, pp. 1–22

_____. 2011b. “Reason-giving and the law.” In L. Green and B. Leiter (Eds.), *Oxford Studies in Philosophy of Law* (Vol.1, pp. 1–38). Oxford: Oxford University Press

_____. 2014. “Authority and reason-giving.” *Philosophy and Phenomenological Research,* 89*,* pp. 296–332

_____. 2019. “Is general jurisprudence interesting?” In D. Plunkett, S. Shapiro & K. Toh (Eds.), *Dimensions of Normativity: Essays on Metaethics and Jurisprudence* (pp. 65–84). New York: Oxford University Press.

Eriksson, L. & Southwood, N. 2011. “Norms and conventions.” *Philosophical Explorations,* 14, pp. 195–217.

Feldman, R. 2000. “The Ethics of Belief.” *Philosophy and Phenomenological Research*, 60, pp. 667–695.

Finlay, S. 2019. “Defining normativity.” In D. Plunkett, S. Shapiro & K. Toh (Eds.), *Dimensions of Normativity: Essays on Metaethics and Jurisprudence* (pp. 187–216). New York: Oxford University Press.

Fischer, J. 2006. “The Cards That Are Dealt You.” *Ethics*, 10, pp. 107–129.

Fogal, D. 2020. “Rational Requirements and the Primacy of Pressure.” *Mind*, 129/516, pp. 1033–1070.

Frankfurt, H. 1971. “Freedom of the Will and the Concept of a Person.” *Journal of Philosophy*, 68, pp. 5–20.

_____. 1988. “Identification and Wholeheartedness.” In his *The Importance of What We Care About*. Cambridge: Cambridge University Press, pp. 159–176.

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

Galef, J. 2021. *The Scout Mindset: Why Some People See Things Clearly and Others Don’t*. London: Portfolio.

Gert, J. 2018. “Underdetermination by reasons.” In D. Star (Ed.), *The Oxford Handbook of Reasons and Normativity* (pp. 443-460). New York: Oxford University Press.

Gibbard, A. 1990. *Wise Choices, Apt Feelings*. Cambridge: Harvard University Press.

Gigerenzer, G. 2010. *Rationality for Mortals: How People Cope with Uncertainty*. Oxford: Oxford University Press.

Goldman, A. 2011. “Experts: Which ones should you trust?” In Goldman, A., & Whitcomb, D. (Eds.). *Social epistemology: Essential readings*. Oxford: Oxford University Press, pp. 109–133.

Hampton, J. 1993. “Selflessness and the Loss of Self.” *Social Philosophy & Policy*, 10, pp. 135–165.

Hart, H. L. A. 1961. *The Concept of Law*. Oxford: Clarendon Press.

Hazlett, A. 2015. “The Social Value of Non-Deferential Belief.” *Australasian Journal of Philosophy*, 94, pp. 131–151.

Heathwood, C. 2005. The Problem of Defective Desires, *Australasian Journal of Philosophy,* 83, pp. 487-504.

Henrich, N. & Henrich, J. 2007. *Why Humans Cooperate: A Cultural and Evolutionary Explanation*. Oxford: Oxford University Press.

Hieronymi, P. 2006. “Controlling Attitudes.” *Pacific Philosophical Quarterly*, 87, pp. 45–74.

_____. 2008. “Responsibility for Believing.” *Synthese*, 161, pp. 357–373.

Holton, R. 2004. “Rational Resolve.” *Philosophical Review*, 113, pp. 507–535.

_____. The Act of Choice, *Philosophers’ Imprint 6,* pp. 1–15.

Horne, A. 2022. “Too Many Cooks.” *Synthese,* 200, p. 311.

Horowitz, S. 2019. “The Truth Problem for Permissivism.” *The Journal of Philosophy*, 116, pp. 237–262.

Horty, J. & Nair, S. 2018. “The logic of reasons.” In D. Star (Ed.), *The Oxford Handbook of Reasons and Normativity* (pp. 67–83). New York: Oxford University Press.

Hubin, D. 1996. “Hypothetical Motivation.” *Noûs,* 30, pp. 31–54.

James, William. 1896 [1956]. *The Will to Believe and Other Essays in Popular Philosophy*. New York: Dover Publications.

John, G. 2011. “Expert Testimony and Epistemological Free-Riding: The MMR Controversy.” *The Philosophical Quarterly*, 61, pp. 496–517.

_____. 2018. “Epistemic trust and the ethics of science communication: against transparency, openness, sincerity and honesty.” *Social Epistemology*, 32, pp. 75–87.

Joyce, R. 2001. *The Myth of Morality*. Cambridge: Cambridge University Press

Kahan, Dan M. 2017. “Misconceptions, Misinformation, and the Logic of Identity-Protective Cognition.” *Cultural Cognition Project Working Paper Series No. 164*, Yale Law School, *Public Law Research Paper No. 605*, *Yale Law & Economics Research Paper No. 575*, Available at SSRN: https://ssrn.com/abstract=2973067 or http://dx.doi.org/10.2139/ssrn.2973067

Kahane, G. 2021. “Is the Universe Indifferent? Should We Care?” *Philosophy and Phenomenological Research,* 104, pp. 676–695.

Kahane, G. & Savulescu, J. 2015. “Normal Human Variation: Refocussing the Enhancement Debate.” *Bioethics*, 29, pp. 133–143.

Kahneman, D. 2011. *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.

Kahneman, D., Sibony, O., & Sunstein, Cass R. 2021. *Noise: A Flaw in Human Judgment*. New York: Little, Brown Spark.

Kelly, T. 2011. “Peer Disagreement and higher order evidence” In Goldman, A., & Whitcomb, D. (Eds.). *Social epistemology: Essential readings*. Oxford: Oxford University Press, pp. 183–217.

Kiesewetter, B. 2017. *The Normativity of Rationality*. Oxford: Oxford University Press.

Kolodny, N. 2003. “Love as Valuing a Relationship.” *Philosophical Review*, 112, pp. 135–189.

_____. 2005. Why Be Rational?, *Mind,* 114, pp. 509–563

_____. (2011). Aims as Reasons. In S. Freeman, R. Kumar and R. J. Wallace (Eds.), *Reasons and Recognition: Essays on the Philosophy of T. M. Scanlon*, (pp. 43-78). Oxford: Oxford University Press.

Kornblith, H. (2002). *Knowledge and its Place in Nature*. New York: Oxford University Press.

Korsgaard, C. 1986. “Skepticism about Practical Reason.” *Journal of Philosophy*, 83, pp. 5–25.

_____. 1996. *The Sources of Normativity*. Cambridge: Cambridge University Press.

_____. 2009. *Self-Constitution: Agency, Identity and Integrity*. Oxford: Oxford University Press.

Langton, R. 2007. “Objective and Unconditioned Value.” *Philosophical Review,* 116, pp. 157–185.

_____. (2009). Esteem in the moral economy of oppression, *Philosophical Perspectives,* 23, pp. 273–291

Leefmann, J. & Lesle, S. 2021. “Knowledge from scientific expert testimony without epistemic trust.” *Synthese*, 197, pp. 3611–3641.

Leiter, B. 2011. "The Paradox of Fatalism and Self-Creation in Nietzsche." In John Richardson and Brian Leiter, eds., *Nietzsche*. Oxford: Oxford University Press, pp. 281–321.

Leon, M. 2002. "Responsible Believers." *Monist*, 85, pp. 421–435.

LessWrong, https://www.lesswrong.com, accessed 27 August 2022.

Levy, N. 2019. "Due deference to denialism: explaining ordinary people's rejection of established scientific findings." *Synthese*, 196, pp. 313–327.

Lewis, D. (1989). Dispositional Theories of Value, *Supplementary Volume, Proceedings of the Aristotelian Society, 63,* pp. 89-174.

List, C and Puppe, C. (2009). Judgment Aggregation: A Survey. In P. Anand, C. Puppe, and P. Pattanaik (Eds.), *The Handbook of Rational and Social Choice*, (pp. 457-482). Oxford: Oxford University Press.

Lord, E. 2018. *The Importance of Being Rational*. Oxford: Oxford University Press.

Manne, K. 2013. On Being Social in Metaethics. In R. Shafer-Landau (Ed.), *Oxford Studies in Metaethics* (Vol.8, pp. 50-73). New York: Oxford University Press.

McAdams, R. 1995. "Cooperation and Conflict: The Economics of Group Status Production and Race." *Harvard Law Review,* 108, pp. 1003–1084.

McPherson, T. 2012. "Mark Schroeder's hypotheticalism: agent-neutrality, moral epistemology, and methodology." *Philosophical Studies,* 157, pp. 445–453.

_____. 2018. "Authoritatively normative concepts." In R. Shafer-Landau (Ed.), *Oxford Studies in Metaethics* (Vol.13, pp. 253-277). New York: Oxford University Press.

Mill, J. S. 1957. *Utilitarianism*. New York: Bobbs Merrill.

Murphy, M. 1999. "The Simple Desire-Fulfillment Theory." *Noûs,* 33, pp. 247–272.

Nagel, T. 1970. *The Possibility of Altruism*. Oxford: Clarendon.

Nagel, J. 2010. "Epistemic Anxiety and Adaptive Invariantism." *Philosophical Perspectives*, 24, pp. 407–435.

Newton, L. 2022. "Epistemic anxiety and epistemic risk." *Synthese*, 200, pp. 324–324.

Nguyen, C. T. 2020. "Cognitive islands and runaway echo chambers: problems for epistemic dependence on experts." *Synthese*, 197, pp. 2803–2821.

Nietzsche, F. 1966. *Beyond Good and Evil.* Translated by Walter Kaufmann. New York: Random House.

Noggle, R. 2005. "Autonomy and the Paradox of Self-Creation." In James Taylor, ed., *Personal Autonomy*. Cambridge: Cambridge University Press, pp. 87–108.

Nozick, R. 1974. *Anarchy, State, and Utopia*. Oxford: Blackwell.

Parfit, D. 2011. *On What Matters.* Oxford: Oxford University Press.

Pascal, B. 1670 [1995], *Pensées*, A. Kreilsheimer (trans. and ed.), New York: Penguin.

Paul, L. A. 2014. *Transformative Experience*. Oxford: Oxford University Press.

Pettit, P. 1990. "Virtus normativa: rational choice perspectives." *Ethics*, 100, pp. 725–755.

_____. 2019. "Social norms and the internal point of view: an elaboration of Hart's genealogy of law." *Oxford Journal of Legal Studies,* 39, pp. 229–258.

Pickard, H. 2021. "Addiction and the Self." *Noûs,* 55, pp. 737–761.

Pinker, S. 2021. *Rationality: What It Is, Why It Seems Scarce, Why It Matters*. London: Allen Lane.

Podgorski, A. 2020. "Normative uncertainty and the dependence problem." *Mind*, 129, pp. 43–70.

Portmore, D. 2011. *Commonsense Consequentialism: Wherein Morality Meets Rationality*. New York: Oxford University Press.

Portmore, D. 2013. "Perform your Best Option." *Journal of Philosophy*, 110, pp. 436–59.

Prichard, D. 2016. *Epistemic Angst: Radical Skepticism and the Groundlessness of Our Believing*. Princeton, N.J: Princeton University Press.

_____. 2018. "Epistemic Angst." *Philosophy and Phenomenological Research*, 96, pp. 70–90.

Quine, W. 1976. "Posits and Reality", reprinted in *The Ways of Paradox and Other Essays* (pp. 246-254). Cambridge, MA: Harvard University Press.

_____. 1980. "Two Dogmas of Empiricism." reprinted in *From a Logical Point of View* (pp. 20-46). Cambridge, MA: Harvard University Press.

Railton, P. 1986a. "Facts and Values." *Philosophical Topics*, 14, pp. 5–31.

_____. 1986b. "Moral Realism." *The Philosophical Review,* 95, pp. 163–207.

Rawls, J. 1971. *A Theory of Justice*. Cambridge: Harvard University Press

Raz, J. 1988. *The Morality of Freedom*. Oxford: Oxford University Press.

_____. 1999. *Engaging Reason*. New York: Oxford University Press.

_____. 2012. *From Normativity to Responsibility*. New York: Oxford University Press.

Rosati, C. 1995. "Persons, Perspectives, and Full Information Accounts of the Good." *Ethics,* 105, pp. 296–325.

Roussos, J. 2021. "Expert deference as a belief revision schema." *Synthese*, 199, pp. 3457–3484.

Sampson, E. 2021. "What if ideal advice conflicts? A dilemma for idealizing accounts of normative practical reasons." *Philosophical Studies,* 179, pp. 1091–1111.

Scheffler, S. 2010. “Valuing.” In his *Equality and Tradition*. Oxford: Oxford University Press, pp. 15–40.

Schroeder, M. 2004. “The scope of instrumental reason.” *Philosophical Perspectives*, 18, pp. 337–364.

_____. 2005. “Cudworth and normative explanations.” *Journal of Ethics and Social Philosophy,* 1, pp. 1–27.

_____. 2007. *Slaves of the Passions*. New York: Oxford University Press.

_____. 2012. “Reply to Shafer-Landau, Mcpherson, and Dancy.” *Philosophical Studies,* 157, pp. 463–474.

Sen, A. 2004. “Incompleteness and Reasoned Choice.” *Synthese*, 140, pp. 43–59.

Setiya, K. 2017. *Midlife: A Philosophical Guide*. Princeton, N.J.: Princeton University Press.

Shafer-Landau, R. 2012. “Three problems for Schroeder’s hypotheticalism.” *Philosophical Studies,* 157, pp. 435–443.

Silverstein, M. 2012. “Inescapability and normativity.” *Journal of Ethics and Social Philosophy*, 6, pp. 1–26.

Sklar, L. 1975. “Methodological Conservatism.” *The Philosophical Review*, 84, pp. 384–400.

Smith, M. 1994. *The Moral Problem*. Oxford: Blackwell.

Smith, N. 2016. “Infinite decisions and rationally negligible probabilities.” *Mind,* 125, pp. 1199–1212.

Snedegar, J. 2017. *Contrastive Reasons*. New York: Oxford University Press.

Sobel, D. 2016. *From Valuing to Value: Towards a Defence of Subjectivism*. New York: Oxford University Press.

Sosa, E. 1991. *Knowledge in Perspective: Selected Essays in Epistemology*. Cambridge: Cambridge University Press.

Southwood, N. 2008. “Vindicating the Normativity of Rationality.” *Ethics*, 119, pp. 9–30.

_____. 2011a. “The authority of social norms.” In M. Brady (Ed.), *New Waves in Metaethics* (pp. 234–248). Hampshire: Palgrave Macmillan.

_____. 2011b. “The moral/conventional distinction.” *Mind,* 120, pp. 761–802.

_____. 2019. “Laws as conventional norms.” In D. Plunkett, S. Shapiro & K. Toh (Eds.), *Dimensions of Normativity: Essays on Metaethics and Jurisprudence* (pp. 23–44). New York: Oxford University Press.

_____. 2019. “Contractualism for Us As We Are.” *Philosophy and Phenomenlogical Research,* 99, pp. 529–547.

Strawson, G. 1986. *Freedom and Belief*. Oxford: Oxford University Press.

_____. 1994. “The Impossibility of Moral Responsibility.” *Philosophical Studies*, 75, pp. 5–24.

Street, S. 2009. “In Defense of Future Tuesday Indifference: Ideally Coherent Eccentrics and the Contingency of What Matters.” *Philosophical Issues*, 19, pp. 273–298.

Taleb, N. 2001. *Fooled by Randomness: The Hidden Role of Chance in Life and the Markets.* New York: Random House.

Tucker, C. (2021). “The dual scale model of weighing reasons.” *Noûs*, 56, pp. 366–392.

Ullmann-Margalit, E. & Morgenbesser, S. 1977. “Picking and Choosing.” *Social Research,* 44, pp. 757–85.

Valentini, L. 2018. “The content-independence of political obligation: what it is and how to test it.” *Legal Theory*, 24*,* pp. 135–157.

_____. 2021. “Respect for persons and the moral force of socially constructed norms.” *Noûs*, 55, pp. 385–408.

Vazard, J. 2021. “(Un)reasonable doubt as affective experience: obsessive–compulsive disorder, epistemic anxiety and the feeling of uncertainty.” *Synthese*, 198, pp. 6917–6934.

Walden, K. 2015. “The Euthyphro Dilemma.” *Philosophy and Phenomenological Research,* 90, pp. 612–639.

Wallace, J. 2013. *The View from Here: On Affirmation, Attachment, and the Limits of Regret*. New York: Oxford University Press.

Watson, G. 1975. “Free Agency.” *Journal of Philosophy*, 72, pp. 205–220.

_____. 1993. “Responsibility and the Limits of Evil: Variations on a Strawsonian Theme.” In John Martin Fischer, ed., *Perspectives on Moral Responsibility*. Ithaca, NY: Cornell University Press, pp. 119–148.

Watts, D. 2011. *Everything is Obvious (Once You Know the Answer): How Common Sense Fails*. London: Atlantic Books.

White, S. 2021. “Self-prediction in practical reasoning: its role and limits.” *Noûs*, 55, pp. 825–841.

Williams, B. [1976] 1981. “Persons, Character and Morality.” In his *Moral Luck*. Cambridge: Cambridge University Press, pp. 1–19.

_____. [1980] 1981. “Internal and External Reasons.” In his *Moral Luck*. Cambridge: Cambridge University Press, pp. 101–111.

_____. 1995. “Internal Reasons and the Obscurity of Blame.” In his *Making Sense of Humanity*. Cambridge: Cambridge University Press, pp. 35–45.

Wolf, S. 1982. “Moral Saints.” *The Journal of Philosophy*, 79, pp. 419–439.

_____. 1987. “Sanity and the Metaphysics of Responsibility.” In Ferdinand Schoeman, ed., *Responsibility, Character, and the Emotions*. Cambridge: Cambridge University Press, pp. 46–62.

Woods, J. 2018. “The Authority of Formality.” In R. Shafer-Landau (Ed.), *Oxford Studies in Metaethics* (Vol.13, pp. 207–228). New York: Oxford University Press.

Wright, S. 2021. “Epistemic harm and virtues of self-evaluation.” *Synthese*, 198 (Suppl 7), pp. S1691–S1709.

Yee, G. 2002. “Desiring to Believe.” *Monist*, 85, pp. 446–455.