Deliberators Must Be Imperfect

Derek Baker
derekbaker@ln.edu.hk

This is the penultimate draft of a paper forthcoming in *Philosophy and Phenomenological Research.* Please cite the published version.

Abstract

This paper argues that, with certain provisos, predicting one's future actions is incompatible with rationally deliberating about whether to perform those actions. It follows that fully rational omniscient agents are impossible, since an omniscient being could never rationally deliberate about what to do (omniscient beings, the paper argues, will always meet the relevant provisos). Consequently, theories that explain practical reasons in terms of the choices of a perfectly rational omniscient agent must fail. The paper considers several ways of defending the possibility of an omniscient agent, and concludes that while some of these may work, they are inconsistent with the aim of explaining practical normativity by appeal to such an agent.

1. The Problem

Foreknowledge and free choice are not always friends. "Deliberation crowds out prediction," in Isaac Levi's famous phrase (1997 :ix). Or, as Kant put it, practical reasoning must proceed "under the idea of freedom" (1785, 4:448). Even if one's actions are determined, one cannot think of them as determined and still sensibly pose to oneself the question of what to do. Jay Wallace elaborates, writing that without unpredictability "there would be no room for deliberation and for the related phenomenon of acting as the result of decision or choice, on the basis of reasons" (1994: 3). You cannot deliberate about whether to perform an action if you already know that you will—at least not rationally. What would be the point?

But notice that, to be omniscient, one must know every event that will occur. One's future actions are events. It would never make sense, then, for an omniscient being to reason about what to do. But an agent is a creature that reasons about what to do—which means that "omniscient agent," or at least "perfectly rational omniscient agent," must be an oxymoron.

Tomis Kapitan (1990) offers a similar argument against the intelligibility of an omniscient *and* omnipotent God. This may seem of minor interest to secularly minded philosophers, but the argument is also a threat to any theory of practical normativity that appeals to counterfactuals about how ideally rational, ideally informed agency would be exercised—that is, to *Ideal Agent theories*.[1]

[1] (Österberg 1998) also points out that these considerations cast doubt on the ability of an omniscient agent to deliberate; however, he thinks that these limitations would only undermine the coherence of an omniscient deliberator if we accept an *actualist* theory of moral obligation, and so concludes that actualism about obligation is incompatible with the action-guiding role of morality. See (Carlson 2003) for criticisms. For recent overview and discussion of the actualism-possibilism debate about obligations, see (Louise 2009; Portmore 2011; Baker 2012; Hedden 2012; Ross 2013; and Timmerman *forthcoming*). For recent discussion of the ways third-person omniscience (particularly God's) might be incompatible with free agency, see (Fischer 1992; Todd 2013; and Fischer and Tognazzi 2014).

Ideal Agents are supposed to be perfectly rational omniscient agents; in fact, they are typically supposed to be perfectly rational and omniscient counterparts to actual agents. For theorists such as Michael Smith (1994; 2004b; 2004c; 2004d; 2004e; 2004f; and 2013) these Ideal Agents explain practical reasons. You ought to *A* because you would *A* (or *advise* your benighted counterpart to *A*) if you were fully informed and all your rational dispositions and capacities were perfectly functioning; more formally:

> Agent *S* has most reason to *A* in situation *C* in virtue of the fact that *S's* omniscient, perfectly rational counterpart would perform/advise *A* in *C*.[2]

[2] There are important theoretical reasons why Smith is committed to the omniscience of these Ideal Agents (it is not simply a matter of overlooking weaker claims that would be equally serviceable). Even one's perfectly rational counterpart will make bad decisions if she is ignorant of some fact. The obvious solution to this problem is to stipulate that the counterpart know all the *relevant* facts (cf. Smith 1994 and 2004b). But since the aim is to explain *reasons* in terms of the choices of the Ideal Agent, the theories can't characterize the Ideal Agent's knowledge in terms of what's relevant, on pain of circularity. It should be noted that in Smith's original (1994) characterization of relevant truths, relevance is given a dispositional gloss—and is characterized in terms of the following quote from Williams (1981): a relevant fact is such that "if he [the agent] did know it he would, in virtue of some element in… [the agent's set of desires]… be disposed to Φ…" Since the argument of this paper is that knowledge of what one will do can affect an agent's dispositions to act (by making her insensitive to other considerations), to stipulate that knowledge of one's own action is not relevant is to beg the question. The concern, very simply, is that while it may be true that I, as I am, would ψ rather than ϕ if I were to learn to that *p,* it may also be the case that if I were already convinced that I would ϕ, this would render me insensitive to the truth of *p,* or any other fact I might learn. The

We just saw an argument that these explanations cannot be given—because the *explanans* is incoherent. And while this argument can seem arcane, it actually goes to the heart of what Ideal Agent theories are trying to accomplish. These theories attempt to explain practical normativity by calling to our attention the various component activities of deliberation—considering one's evidence, reasoning about the consequences, imaginative engagement with these consequences, correcting for biases in one's feelings, considering how one would assess the action if one were patient rather than agent, etc. Then we imagine an agent who performs all of these activities perfectly. What that agent would do after such an act of deliberation determines what the relevant normative facts are. The argument here brings to our attention an assumption of that explanatory project—that all of the component activities *can* be exercised to their maximal extent and still harmonized into a single act of deliberation—and tells us the assumption is false. Some procedures, upon reaching perfection, block others.[3]

knowledge that I will ϕ is thus potentially relevant in the only sense available to the Ideal Agent theorist.

Moreoever, in Smith's later work the Ideal Agent is explicitly described as "possessed of all true beliefs" (2004c: 44) "omniscient" (2004f: 300) and "maximally informed" (2004d: 129ff; and 2004e: 265-6). Readers should also note that (2004c and 2004e) are both replies to criticisms of (Smith 1994), indicating that omniscience was always implicitly an element of ideal agency within Smith's theory.

[3] Note that this objection is also different in substance from superficially similar ones that may be more familiar. One familiar objection is that full-information of the relevant sort is impossible: a fully-informed agent could not know what it was like to be surprised or to have a novel experience—

The case against Ideal Agents can be broken into premises:

1. Necessarily, if an agent knows that she will *A* and the agent is perfectly rational, then the agent cannot deliberate about whether to *A*.
2. Necessarily, for any perfectly rational agent, there is some possible *A* the agent can deliberate about performing.
3. Necessarily, if an agent is omniscient, then for every possible *A* the agent knows whether she will *A*.

C. Perfectly rational omniscient agents are impossible.

The remainder of this paper will argue that while there may be ways of resisting this argument, they are unavailable to Ideal Agent Theorists, at least those with globalist explanatory ambitions. This paper will not address attempts to use Ideal Agents, including Ideal *Contractors*, to explain some domain of practical normativity, such as one's personal good (Railton 1986) or moral obligation (Firth 1952; Railton 1986; Scanlon 1998; and Parfit 2011). These theories *may* be vulnerable to a variation on the argument above, but the

but then he is not fully informed (Sobel 1994; and Rosati 1995). Another is that the particular idealization we settle on is *ad hoc*, without independent motivation except that it makes the account of reasons extensionally adequate (Enoch 2005). Another is that idealized agents are too idealized—making them so alien to their actual counterparts that the normative import of their preferences is obscure; that I want something has some relevance to what I should do, but it seems irrelevant that I would want something else if my psychology were different in ways hard to even imagine (Rosati 1995; Hubin 1996; and Enoch 2005). The objection of this paper would stand even if the other objections were answered.

paper's dialectical structure imposes agnosticism on this issue.[4] I will show that the most obvious ways of resisting the above argument are unavailable to the globalist. Attempting to do the same for local Ideal Agency theories is beyond the scope of this paper.

2. Premise 1

So why believe premise 1? Here is a tempting argument that unfortunately assumes too much. According to Kapitan, one can only deliberate about what one believes to be contingent in some appropriate sense, a sense compatible with exercising control over it. That seems hard to deny.[5] But then Kapitan goes on to argue that the appropriate sense of contingency is contingency "relative to all that he (the agent) then (at t) believes" (129). Levi makes a similar point: "If Sam is certain that he will not yield his wallet, paying is not possible as far as he is concerned" (1997: 28). In other words, I cannot believe that I will *A,* and still see *A* as answerable to my control.

[4] My hunch is that this argument is problem for some localist theories and not for others.

[5] Precisely characterizing this sense of contingency is difficult. It may be that it reduces to some other sort of contingency, but (Maier 2013) argues that *agentive modalities* should be defined in terms of *being an option*, an essentially agency-related property to be treated as primitive. This would make agentive modality a *sui generis* form of modality. I will assume that Maier is correct at least in so far as, reducible or not, agentive modality has not yet been successfully reduced to other types of modality, and it is ineliminable. For the purposes of this discussion, then, the sense of contingency will be treated as primitive, relying on our intuitions about cases to decide what an agent is free and not free to do.

David Hunt (1997: 277-9) points out this argument seems to commit the fatalist fallacy, or at least to insist that the perfectly rational agent reason like a fatalist. Hunt, using "$\Box_0 P$" to indicate that *P* is true at all worlds it is within the agent's power to bring about, starting from some arbitrary time t_0 (and holding the past up to t_0 fixed), states the objection thus: "…[I]f it is false that $P \supset \Box_0 P$, it is a total mystery why every rational agent should nevertheless believe $\Box_0 P$ upon believing *P,* and do so as a requirement of rationality" (279). The belief that I will *A* does not justify belief in *A's* inevitability, and so Kapitan has misidentified the sense of contingency appropriate to choice.[6]

In the next subsection, I will argue that a fatalist principle will hold for the Ideal Agent, and so Kapitan and Levi are correct that being epistemically possible (in some very weak sense at least) is a condition on being practically possible. But I will grant that simply assuming as much is question-begging. For now, I will simply treat premise 1 as provisionally plausible. There is difficulty in making sense of the state of mind of the person who knows what he will do and still goes about trying to decide what to do. Something seems incoherent about the psychology.

Still, 1 needs to be weakened. If we read on in our Kant about why we must regard ourselves as free, he explains that "one cannot possibly think of a reason that would consciously receive direction from any other quarter with respect to its judgments, since the subject would then attribute the determination of his judgment not to his reason, but to an impulse" (*ibid.*). The problem is not with seeing one's will as determined, but as determined

[6] See (Carlson 2002) for a similar objection.

by something other than its own deliberation. Rather than crowding out prediction, deliberation can serve as its basis.[7] So we have a reason to reject 1.

But there is a slogan in the neighborhood of Levi's that we should still accept. Let us call knowledge of what I will do based on my conclusion about what to do, *deliberative*. Let us call knowledge of our future actions based on anything besides deliberation—that is, perception, brain scans, psychological laws, oracles, or everyday inductive generalizations ("I know I will not finish this paper unless there is some sort of deadline")—*non-deliberative*. I can know what I will do in the future in both ways.

Our new slogan: non-deliberative prediction crowds out deliberation (cf. Carlson 2002: 80-1).[8] If I know, on the basis of total knowledge of the state of my brain and complete knowledge of the neuropsychological laws, that an electron is going to zoom down my spine once the proper neuron is triggered and as a result I will *A,* what could be the point of deliberating about whether to *A?*[9]

So we should replace 1 with

1′. Necessarily, if an agent *non-deliberatively* knows that she will *A,* then she cannot rationally deliberate about whether to *A*.

[7] This should not be read as an objection to Levi's view, however. Levi could likely agree to this, despite the slogan.

[8] Carlson's criticism is discussed in fn. 12.

[9] A caveat: these "what would be the point?" intuitions should be taken as merely making a provisional case for the impossibility of omniscient and perfectly rational deliberation. Again, an independent argument for why omniscient beings would, if rational, reason like fatalists will come in the next section.

But omniscient agents will know everything non-deliberatively.[10] So we can also replace 3 with

3′. Necessarily, if an agent is omniscient, then for all *A* she *non-deliberatively* knows whether she will *A*.

C still follows.

Have we overlooked the possibility of overdetermined knowledge—that is, knowledge that is fully justified both on the basis of deliberation and on the basis of non-deliberative methods (observation, inference, etc.)?[11] Such overdetermination may be possible—for example, if I know that I will mow the lawn first on the basis of deliberation and later on the basis of an infallible oracle's say so. But it is enough for this argument to note that once I have heard the oracle's prophecy and know it for true, *further* deliberation is irrational, at least if our intuitions are accurate (and there will be additional argument that they are). But this is enough for premise 1′. 3′ will be true, in turn, for all omniscient agents

[10] Even if one refuses to grant this, there are reasons why Ideal Agents, if they are to do theoretical work in explaining practical normativity, must have non-deliberative knowledge. The most obvious reason is presented in section 4.2. So the premises here could be replaced with even weaker ones and still lead to a conclusion that undermines Ideal Agent theories (though it would perhaps be weaker than C).

[11] Thanks to an anonymous referee for raising this question.

once they have become omniscient (who are presumably the agents we are interested in—a position defended in more detail in section 4.2).

Another way to object to 1′ is insist that we think about the consequences of fallibilism about knowledge. It may be that I non-deliberatively know that I will not finish a paper, given the absence of any firm deadline. Still, I should think about what the reasons favor, because, from my point of view, unlikely as it is, this time my assessing of the reasons might be the first step in the exercise of backbone.[12]

In other words, it is not fore*knowledge* but *certainty* that rules out deliberation.[13]

But omniscient agents would be certain of all the truths they know. They know they are omniscient, and so they know that any evidence contrary to their beliefs is misleading. They would have, then, no rational basis for any degree of doubt about their beliefs. They must, moreover, assign all true propositions a probability of 1, or else there will be some

[12] Thanks to Tristram McPherson for presenting this objection and the connection to fallibilism. Also see (Carlson 2002: 81-2), where it is pointed out that believing that I will *A* on the basis of non-deliberative evidence is compatible with the belief that my deliberation nonetheless causes me to *A*. This is correct, but as Carlson points out, I will deliberate in these cases knowing that my conclusion will be to *A*. It remains unclear, then, how such deliberation is *rational*. The next section will argue, moreover, that non-deliberative beliefs pose a special threat to freedom, assuming plausible coherence requirements on prediction and intention.

[13] Levi (1997: 32; and 76, fn. 5) argues, however, that a rational agent assigning probabilities to her actions *must* assign probabilities of 0 or 1 to (i.e., be certain about) each option in a given set of alternatives. (Rabinowicz 2002) elaborates and criticizes Levi's argument, as well as other similar arguments in favor of thinking that mere (credential) prediction, as opposed to certainty, could rule out deliberation. See (Levi 2007) for further discussion of the argument and a reply.

very long true conjunctions in which their confidence is too low to count as knowledge. So if we replace 1′ with

> 1″. Necessarily, if an agent is *non-deliberatively certain* that he will *A,* then he cannot rationally deliberate about whether to *A*.

Then we must replace 3′ with

> 3″. Necessarily, if an agent is omniscient, then for all *A* he is *non-deliberatively certain* about whether he will *A*.[14]

C still follows.

2.1. Fatalism[15]

[14] Even if one still denies that omniscience entails omnicertainty, it is still the case that Ideal Agents, to do the theoretical work they are supposed to do, must be omnicertain. In some cases, the simple presence of doubt (no matter how tiny) that a venture will succeed will lead a rational agent to refrain from it. But then there will be cases in which the Ideal Agent could have realized some goal, but doesn't, because she isn't completely certain of the outcome. 1″ and 3″ could potentially be replaced with even weaker premises, then, which would still rule out the Ideal Agent's possibility.

[15] Thanks to Jack Woods and Jian Shen for double-checking the validity of the formal argument in this section. Any problems that remain with the argument are of course mine.

The previous section acquiesced to some of the objections to 1, and the result was 1″. The positive case for the premise, though, was simply the intuition that deliberation would be pointless if one already was certain what its conclusion would be. But there is reason for skepticism. Hunt's objection brought to our attention an unreflective tendency to assume the truth of fatalism when prediction is in play. Our intuitions may be corrupt.

A stronger case will be made for 1″ by meeting Hunt's challenge head on. A perfectly rational omniscient agent would have to reason like a fatalist, given some fairly weak and plausible assumptions about the rational constraints on intention and belief.

Following Hunt, let us use "$\Box_0$" to indicate what is *uniquely* possible for a rational agent at some arbitrary time t_0 (what would be true in all worlds in his power to bring about from t_0 on, what is necessary relative to what is under the agent's control, what is *inevitable* for that agent). Let us use "*p*" for the proposition that the agent performs some intentional action and "~*p*" for the proposition that the agent performs some alternative (any available action ruling out the truth of *p*). "*Bel (p)*" will refer to the agent's belief that *p,* and "*Dec (p)*" to the agent's decision that *p*.

So, at any t_0 when the agent is omniscient and perfectly rational, the following will be true:

i. $p \rightarrow Dec\ (p)$

ii. $p \rightarrow Bel\ (p)$

iii. $Bel\ (p \rightarrow Dec\ (p))$

i is true because *p* is an intentional action. **ii** and **iii** just follow from the agent's omniscience.

Now, because the agent's beliefs are non-deliberative certainties, the following will also be true:

iv. $Bel\ (p) \rightarrow \Box_0\, Bel\ (p)$

v. $Bel\ (p \rightarrow Dec\ (p)) \rightarrow \Box_0\, Bel\ (p \rightarrow Dec\ (p))$

iv and **v** (and notice that **iv** implies **v**) will be true of the omniscient agent for two reasons. First, certainties cannot be rationally revised, so it is not within the agent's power to come to a different conclusion by say, performing a voluntary mental act such as reconsidering the evidence or searching for new evidence. Second, "$\Box_0$" indicates that something is true in all possibilities under one's *voluntary* control from a certain point in time. Since voluntarism about beliefs is false, a belief will only be under voluntary control if one formed the belief as the result of the exercise of volitional capacities, which cannot be the case with non-deliberative beliefs.

One additional point should be made here, to avoid the charge that these premises are question-begging. The premises do not claim that if *p* is true, then there is no world in which the omniscient agent has the freedom to bring it about that she believes *~p*. *p* is not necessary, and so presumably there are worlds in which the agent is free to bring about *~p,* and given her superior epistemic circumstance she could then have made it the case that she believes *~p* (or so we should assume to avoid begging the question). Remember "$\Box_0$" indicates that event is inevitable for the agent from some arbitrary point in time. The claim then is that once the agent is omniscient and perfectly rational in a *p*-world—once she *knows* that *p*—it will no longer be possible for her to *revise* her belief about *p* through the exercise of voluntary capacities: the knowledge is a certainty, and it is based on non-deliberative evidence.[16]

Now we can introduce **vi**:

vi. $\Box_0 ((Bel\ (p)\ \&\ Bel(p \rightarrow Dec\ (p))) \rightarrow {\sim}Dec\ ({\sim}p))$

vi. is the claim that an omniscient rational agent who believes that *p* depends on his decision that *p*, and also believes that *p*, cannot maintain those beliefs and decide to perform an alternative to *p*. Note that this is weaker than the claim that one must believe that one will

[16] David Faraci offers the following objection: at t_0 the agent might believe that she will deliberate at t_1, and the conclusion of this deliberation will be her performance of *p* at t_2. In this case, she non-deliberatively believes *p* at t_0, even though the belief could still be revised via deliberation. This is to imagine that the agent's belief is non-deliberative at t_0, but will cease to be non-deliberative and turn into a purely deliberative belief at some later point (the agent comes to 'own' the deliberation, rather than simply predicting it, by going through it). The problem is that the agent's reasons for believing that *p* are massively overdetermined (which is presumably true of all of her beliefs). She knows that *p* because she knows how she will deliberate, but also because she knows with certainty the natural laws of her world and its current state, because she knows with certainty her own psychological states, because she knows with certainty which events will follow upon *p,* and so what must have happened to cause those events. She will always have overwhelming non-deliberative evidence of what she will do. Consequently the belief will never shift into a deliberative belief, at least so long as the agent is epistemically rational, even if in normal agents such shifts can occur.

do what one intends. It only says that one cannot intend contrary to one's beliefs *without* revising those beliefs.[17] In fact, given what we've already said about omniscient agents, that all of their beliefs are certainties, the claim is even weaker than that: one cannot rationally intend contrary to one's *certainty* without giving up that certainty.

Two additional points should be made about **vi.** First, it does not entail on its own that the belief that *p* leaves the agent with no alternatives to *p,* because normally the agent can revise her beliefs (though readers can probably see what's coming).

Second, it is stated as a claim on what a rational agent *can* do. This may raise worries: believing *p* while deciding *~p* is irrational, not impossible; the perfectly rational agent *will not* do it, but that does not imply that she cannot. But keep in mind here that the rational requirement is a requirement of coherence, a requirement of coherent states of mind. We cannot, in general, simply decide to violate a norm of coherence. Our attitudes adjust towards coherence through automatic, non-voluntary processes. The "ability" to have incoherent states of mind depends, then, on non-voluntary dispositions: on prior habits or tendencies such as wishful thinking, bias, disordered appetites, inferior levels of self-control, tendencies to fly into rages or panics, or other imperfections in the agent's non-voluntary adjustment. But if the agent's non-voluntary dispositions are so disordered, the agent is not perfectly rational. In short, it is correct that the perfectly rational psychology is such that she *cannot* violate these norms.

[17] See Holton (2009 :40-52) for stronger and weaker consistency requirements matching beliefs with intentions. While I do not follow his exact terminology, the requirement proposed here is close to his "Very weak consistency for partial intentions" (42), which requires only that one not believe contrary to one's intention.

Now, from the above premises, it follows that[18]

vii. $p \rightarrow (\Box_0 {\sim}Dec\ ({\sim}p))$

But if an agent exists in a *p*-world, and it is impossible for her to decide *~p,* then alternatives to *p* will be outside of her power. After all, she cannot choose any of those alternatives, she cannot do anything to bring them about. The only way they could come about is through luck or good fortune—that is, precisely through the things that are outside of her control. And, being in a *p*-world she is in a world where good fortune does not bring those things about. To put this formally:

viii. $(p\ \&\ \Box_0 {\sim}Dec\ ({\sim}p)) \rightarrow \Box_0 p$

By combining **vii** and **viii** we get

ix. $p \rightarrow \Box_0 p$

That's why the omniscient agent would reason like a fatalist.

3. Premise 2

[18] Or at least it follows if we allow ourselves to assume an axiom of all normal modal logics: $\Box(p \rightarrow q) \rightarrow (\Box p \rightarrow \Box q)$.

Premise 2 claims that *necessarily, for any perfectly rational agent, there is some possible action the agent can deliberate about performing.* But maybe deliberation is not essential to agency. Maybe agency only requires action—that is, bodily movements caused in the right way by beliefs and motives. Peter Railton (2004) has argued compellingly that the exercise of rationality (including rational agency) depends on unreflective brute dispositions of our psychology to conform to what the reasons favor. Reflective, conscious reasoning (including conscious deliberation) is a failsafe, correcting the unreflective when it goes off the rails.[19] The Ideal Agent does not make mistakes and knows she does not. Such a being has no need to deliberate. So the Ideal Agent theorist can reject 2 (or so the argument might go).

Remember, though, that 2 does not actually hold that a perfectly rational agent necessarily deliberates about her options, merely that she must be *able* to do so. The argument in favor of this is simply that deliberation seems like a rational capacity, and perfect rationality requires the ability to exercise all rational capacities. To deny 2, then, one must hold, first, that having unreflectively rational desires is enough—that possession of the failsafe of deliberation is not necessary in an Ideal Agent; and second, that deliberative capacities are not rational capacities.

This, however, raises a second problem for the argument. 'Able' (or 'can') is ambiguous. The omniscient agent may be able to deliberate in the sense of having the capacity, but unable in the sense that the capacity is *masked.* Thus, the argument may equivocate.[20] I will ignore that problem for now. It will emerge that the considerations in favor of 2 are also reasons for rejecting the charge of equivocation. I will present reasons

[19] Similar arguments are advanced in (Arpaly and Schroeder 2012).

[20] This objection was pressed independently by John Maier and by an anonymous referee of an earlier version.

why a theorist such as Smith, who aims at explaining practical reasons in general by appeal to an Ideal Agent, must accept 2. Then I will return to the worry about equivocation.

3.1. Why the Ideal Agent Theorist Is Committed to 2

The first thing to note is that 2 has *overwhelming* intuitive plausibility: deliberation seems like a rational capacity if anything does. But there are additional reasons why theories that aim to explain all of practical reason in terms of an Ideal Agent must allow that powers of reflective deliberation are essential to perfect rationality.

Denying 2 is inconsistent with a standard explanatory motive of the theory. Smith, for example, writes that it is an advantage of his Ideal Agent Theory that it allows us "to explain why it is rational to desire in accordance with the beliefs about the reasons we have" (2004b: 37). But if we deny 2 then deliberation is not a rational capacity, and so failing to control one's motives on the basis of an assessment of reasons cannot be a rational failing. In other words, this defense only blocks the objection by sacrificing one of the core reasons for holding the theory in the first place.

Second, in order to deny 2, we must maintain that unreflective rationality of motives is enough for perfect rationality. But then we need some way of characterizing what it is for motives to be rational.

Given his explanatory ambitions, Smith cannot explain unreflective rationality of motives in terms of what there is reason to desire, what it's appropriate to desire, what one ought to desire. Objects are worthy of desire *because* the Ideal Agent would desire them, not the other way around. This leaves four basic options for how to identify an agent's motives as rational.

The first option is purely Humean. The agent's desires are rational if they display means-end coherence. This is plausible enough. But it makes the Ideal Agent into a theoretical fifth wheel. After all, the only real point of looking at the agent's omniscient counterpart is so that we can correct for ignorance when deciding what the agent ought to do. But we could massively simplify here. The agent ought to do whatever would lead to the satisfaction of her intrinsic desires—not what she mistakenly thinks would lead to their satisfaction.

In short, the Humean instrumentalist can explain reasons in terms of desires *plus* truth. The appeal to a counterpart who takes the appropriate attitude to that truth is an idle embellishment. Such a counterpart may be useful for *modeling* what one has reason to do, but it is not explanatory.

The second option is to hold that an agent's motives are rational if they are instrumentally coherent *and* the intrinsic desires would survive being fully informed (cf. Brandt 1972). I will acknowledge that this sort of ideal agent theory might be able to avoid the objection of this paper. The theorist who endorses this *and* uses it as a defense against the objection must of course deny that deliberation is a rational activity (or perhaps insist that an agent can be relevantly ideal despite suffering incapacities unusual in ordinary agents—though this seems a harder position to maintain); the theorist cannot appeal to a rational connection between deliberation and choice to motivate the theory either. What we should also note, however, is that no process of critical self-reflection plays a role in determining the rational status of motives on this account. This leads to the concern that the impact of information on is purely causal—that there is no obvious way in which it counts as a rational improvement (Hubin 1996).[21]

[21] Also see (Gibbard 1990: 20-1).

In contrast, we can consider the third option—that intrinsic desires are rational if they would survive well-informed critical self-reflection (Smith 1994 and 2004b). This allows us to put aside the worry that the impact of knowledge on intrinsic desires is purely causal. The impact, on this view, is mediated by rational activity, and in virtue of that mediation the changes to intrinsic motives are rational improvements. But the focus of this mediating activity, critical self-reflection is procedural rationality—and being procedural it is successful not based on the verdicts it delivers, but when it treats like cases alike, when it achieves a certain level of generality and uniformity in its verdicts. So it will presumably involve asking the questions such as "Is the motive upon which I am about to act inconsistent with my other motives?" "Have I imagined one situation more vividly than a rival?" "Am I biased to the near?" "Have I given more weight to my point of view than to my neighbor's?" and shifting one's motives when the answer is yes (although exactly which of these questions are genuinely part of procedural rationality is up for debate). But this is effectively to ask whether to be motivated to *A*. So we have the problem, does it make sense to deliberate about whether to be motivated to *A,* if one already knows that one will be motivated to *A*?

In short, the motives of the Ideal Agent (if she is to be genuinely ideal) must have survived a process of critical self-reflection that is fully informed. But the process of critical self-reflection seems to require the exercise of one's deliberative capacities. It seems largely to just *be* deliberation, except it includes questions of what to do in various hypothetical circumstances as well as actual ones, since the aim is determining motivational *dispositions* to act, and not just individual actions. So if this is our view of what it is for motives to be rational, we cannot deny 2.

Then there is our fourth option. In his more recent (2013), Smith proposes the following criticism of Humean pictures of ideal agency:

> According to Hume, an ideal agent is one who fully and robustly possesses and exercises the capacities to do two things: to have knowledge of the world in which he lives, and to realize his desires in it. The main problem with this account, according to the Constitutivist, is that in a wide range of circumstances their exercise pulls in opposite directions. The full and robust exercise of the one capacity does not fully cohere with the full and robust exercise of the other. To the extent that this is so, the ideal agent's psychology is therefore not maximally coherent. … Hume's account of an ideal psychology must therefore be mistaken.

(13-4)

It is not enough, on this picture, that an agent happens to have omniscience and happens to be instrumentally rational, if that agent is to be ideal. Smith argues that epistemic and practical perfection must also be *robust*; the Ideal Agent must be able to exercise her epistemic perfection over a wide range of counterfactuals and to perfectly pursue her desires over a wide range of counterfactuals. He argues, however, that the agent's ability to exercise these powers will be even greater if she possesses certain "*coherence-inducing desires*" (16), defined as desires to promote and protect the capacities already agreed to be part of rational agency, the capacities for knowing and acting. To be rational, then, a set of desires must meet two necessary conditions: the set must be instrumentally coherent, and it must include the intrinsic coherence-inducing desires.

While this account of the rationality of desires is not itself inconsistent with denying 2, the motivation Smith offers for the account is. His view, remember, is that Ideal Agency cannot simply be perfectly informed and perfectly means-end coherent. Additionally, any psychological states "which would ensure that an ideal agent's psychology is much more coherent and unified than it is according to the Humean's conception" are rationally required as well (15). Smith proceeds to argue that possession of a variety of coherence-inducing desires is a requirement of rationality, by engaging in pairwise comparisons: the otherwise ideal agent who possesses the desire will enjoy more secure possession of her rationality than the otherwise ideal agent who does not (16-7). Given this, it seems obvious that even if deliberative capacities are only failsafe devices, being failsafes they would make the agent's rational coherence more secure, and so should themselves be components of ideal rationality on the same grounds as the coherence-inducing desires.

It may seem that Smith would have grounds, however, for treating a deliberative capacity differently from coherence-inducing desires: the argument of this paper so far. Deliberation is a capacity that cannot be enjoyed by an omniscient agent without incoherence, thus it could not be a capacity of the ideal agent. But in fact this simply raises the specter that no possible agent could meet Smith's specifications of perfect rationality.

Imagine two possible agents. One is omniscient, has the correct sort of desires, and lacks the ability to deliberate. The second is extremely knowledgeable, but just ignorant enough to make room for deliberation, has correct desires, and can deliberate of course. The first may have greater epistemic powers, but her possession of these powers is more fragile and more a matter of luck. She has, for example, fewer abilities to resist the charms of advertisements extolling the value of ignorance. So what we have here is one agent who is less than completely perfect in one respect (her epistemic capacities are slightly defective) and another agent who is less than completely perfect in another (her rationality will degrade faster and in a wider range of counterfactuals). In other words, given Smith's account of rationality, tradeoffs must be made.

This by itself is arguably enough to undermine any agent's claim to being ideal: for any such agent there will always be some other agent who is her rational superior *in some respect*. To resist this conclusion, there must be a uniquely optimal set of tradeoffs, such that one agent can claim to be most rational *all things considered*. The difficulties with this strategy will be discussed in section 4.5. For now I will simply note that we would need some argument, then, for regarding some combination of tradeoffs as uniquely optimal. So far as I know, no one has advanced such an argument. Without such an argument, it is unclear how Smith could deny 2, given that a capacity of deliberation, even conceived of as a mere failsafe, improves an agent's standing on one of Smith's metrics of rationality.

In summary, there are two pictures of Ideal Agency which may be consistent with denying 2. But on one of these the Ideal Agent is explanatorily otiose. On the other theory the Ideal Agent would be explanatory, but the theory requires a number of counterintuitive commitments which presumably call for defense, most notably that the ideal agent cannot control her motives in anything resembling the way that any normal adult human being can. There are two other theories in which the Ideal Agent is explanatory and has some obvious normative standing, but these views appear committed to 2.

3.2. Equivocation and the Limits of Dialectical Space

With these points on the table, we can return to the problem of equivocation. It is unclear whether we should think of omniscience as eliminating or simply masking the ability to deliberate—or whether there always is a sharp distinction between masking a capacity and eliminating it. But this is not a pressing concern for the argument of this paper. What we need to know is whether the sense of 'cannot' in 1″ is incompatible with the sense of 'can' in 2. This is not obviously the case: there is a sense in which I can speak Chinese, and another sense in which I cannot (Maier 2013).

There may be viable interpretations of the paper's argument on which it does equivocate. But the previous discussion should make clear that these interpretations are not available to the Ideal Agent theorist.

As we saw, they need the motives of the Ideal Agent to result from the process of fully-informed critical self-reflection, or they must embrace a condition of counterfactual robustness as a condition on rationality. On the first option, it is not enough that the Ideal Agent could deliberate in some very weak sense, one compatible with the capacity being masked or otherwise blocked by omniscience. Again, her desires must be the output of deliberation that takes place under the guidance of omniscience. But our arguments tell us that is impossible. Similarly, if we hold that ideal rationality requires robust rationality, then the Ideal Agent must be able to deliberate in the sense that, were she to acquire an irrational desire, she could recognize it as such and this recognition would cause her motives to shift back to a more coherent state. But omniscience, according to the arguments in section 2, would prevent such a shift.

For 2 to be true, the perfectly rational agent must stand in some relation to the capacity to deliberate—perhaps mere possession, perhaps unmasked possession, or perhaps some even stronger relation. There may be reasons to insist on a very weak relation, like mere possession, which might be compatible with omniscience. But the Ideal Agent

Theorist, given her theoretical ambitions, must accept a stronger relation, one that omniscience rules out.[22]

Given the lack of equivocation, C still follows.

4. Some Expedients that Probably Do Not Work

There are a number of moves that *might* save omniscient agents. But these expedients, much like the charge of equivocation, are inconsistent with the aims of Ideal Agent Theory.

4.1. Bracketing

Perhaps the incompatibility of deliberation and prediction can be overcome, if we simply allow that the offending knowledge is bracketed while one reasons about what to do (Rabinowicz 2002: 92-3; and possibly Levi 1997: 33-6). Omniscience and perfect rationality would be compatible, because the perfectly rational deliberator will simply bracket her knowledge about what she is going to do while deliberating.

For this defense to work, we need some account of what bracketing is. The Ideal Agent theorist cannot simply stipulate that there is some mental operation that takes an irrational combination of psychological states and makes them no longer irrational, or else

[22] Note, however, that as strong as the Ideal Agent theorist must interpret 2, nothing here claims that it need be so strong as to rule out the possibility of an ideal agent who is asleep, or about to die in the next half-second. Thanks to Barry Maguire for asking me to clarify these points.

she is merely stipulating away the objection that omniscient deliberators cannot be ideally rational.

What forms of bracketing are we familiar with? There are the various psychological partitions necessary for self-deceptive and wishful thinking. But those undermine rationality. Cases of bracketing which seem compatible with rationality are those in which I avoid dwelling on a piece of information that is useless or likely to bias my decision. When I grade my students' papers, I bracket my knowledge about their musical tastes. This is the sort of bracketing Ideal Agents must engage in, if they engage in any. After all, bracketing information that I don't believe might lead me astray is irrational.

The first thing to notice is that even this sort of bracketing plausibly involves deviation from *perfect* rationality, and the fact that it is sometimes reasonable for us to bracket in this way fails to show otherwise. There may be cases in which I have very good reason to engage in self-deception. The self-deception is still an irrationality. Even if I have good reason to block out certain facts while I deliberate, this plausibly involves a partitioning of my psychology, of the sort that counts against rational perfection.

But let's assume that bracketing isn't inherently irrational. Notice that it still only makes sense to bracket knowledge because that knowledge is irrelevant or potentially misleading. If our goal, however, is to explain practical reasons in terms of how an Ideal Agent would choose, there is no prior notion of relevance or accuracy to which we can appeal. It makes sense to ignore information or skip processes of reasoning if doing so makes it more likely that I will get the right answer. But that cannot provide a rationale for bracketing if full-information and processes of reasoning are what determine the right answer. It might make sense to shoot against the wind to hit the bull's-eye, but not if the location of the bull's-eye is defined as wherever, if I were shooting perfectly straight, the arrow would land. Rational bracketing is rationalized by considerations to which Ideal Agent theorists cannot appeal; it is precluded by the same commitments that precluded Ideal Agent theorists from restricting the Ideal Agent's knowledge to knowledge of what is relevant.

As an aside, less ambitious theorists, who only wish to explain some local domain of practical normativity, such as morality, in terms of an Ideal Agent, could potentially appeal to some independent notion of a reason to justify bracketing. Such theorists may, then, have a principled way of resisting the objection of this paper, because they could offer a rationale for ignoring certain pieces of evidence—ignoring such evidence makes one substantively more rational (that is, more likely to arrive at the correct answer).

We might try, then, another strategy. We could see Ideal Agent's role as that of an *advisor* (cf. Smith 2004b). Such Ideal Advisors could bracket knowledge about their *own* future choices, while remaining omniscient about the choices of their advisees. Like bracketing, this would effectively block-off unwanted forms of information, but without raising questions about the rationality of reasoning as though ignorant of something you know to be true.

Unfortunately, this strategy relies on an overly literally understanding of the idea of an Ideal Advisor: it is explicit in Smith's version of the theory that the advisor is not a distinct agent giving advice, but instead a more rational version of the advisee, who can be treated *as though she were giving advice*. As Smith puts it:

> The internalism requirement tells us that the desirability of an agent's ϕ-ing in certain circumstances C depends on whether *she* would desire that *she* ϕs in C if *she* were fully rational. … (2004b: 18) [emphasis added]

This principle is explicated in terms of an ideal advisor, not to deny the identity of the agent who might ϕ with the fully rational agent who might desire her to ϕ, but rather to properly represent circumstances C:

> We are to imagine two possible worlds: the *evaluated* world in which we find the agent in the circumstances she faces, and the *evaluating* world in which we find the agent's fully rational self. …[W]e are to imagine the agent's fully rational self in the evaluating world looking across at herself in the evaluated world (so to speak) and forming a desire about what her less than fully rational self is to do in the circumstances she faces in that evaluated world. We might imagine that the self in the evaluating world is giving the self in the evaluated world advice about what to do. Accordingly, this is what I call the "advice" model of the requirement. (ibid)

In other words, talk of an advisor and advice is a heuristic to help us better understand the relation between the counterfactuals. What the advice really is, on this picture, is what the advisee would choose or want for her currently limited self, were she ideal.

Now in order to be able to give advice the Ideal Agent cannot simply know all the events that do occur in the less-than-ideal advisee's world. She must also be omniscient with respect to all of the counterfactuals in that world (how else would she know which action would be best?). But counterfactual omniscience must include knowledge of the answer to "What would my advisee do if she were ideally rational and omniscient (or had the restricted omniscience I do)?" And crucially, since the Ideal Advisor is an idealized version of the advisee, that question is the equivalent of "What would I do?" So, for the advisor, omniscience about the advisee entails omniscience about oneself.

But couldn't the Ideal Agent be ignorant that the advisee was a more limited version of herself? She couldn't, unless the theory is abandoning central motivations. Remember, once again, Ideal Agent theories aim to explain some domain of reasons in terms of the results of practical deliberation. But deliberating is always deliberating about one's own actions. I can judge that someone else ought to do something, but when I reason about whether to *A,* it must always be about whether *I* will *A*.

Moreover, theories like Smith's are compelling because they say that what I ought to do is what *I would do* or *would want my suddenly limited self to do* if I were deliberating perfectly. If the advisor does not regard the advisee as a counterpart, however, the advice merely expresses what *I would want someone else in circumstances like mine to do* if I were reasoning perfectly. This has some normative significance, but not the same significance. I can think myself reasonable and justified in wanting the car dealer to offer me a lower price, while thinking that if I were in his shoes I would have no reason to show such charity.

Facing these problems, we might propose that instead of being ignorant of the advisee's identity, the advisor doesn't know all counterfactuals about her advisee. Specifically, she doesn't know what her advisee would do were she to become ideal. It's not like that counterfactual is particularly relevant to the choices any actual agent faces.

But imagine that the only way to defeat an army trans-dimensional brain-melting baby-torturing demons is to drink the potion that will make one ideal. The theory must say that in situations like this there is no fact of the matter about what one ought to do. The advisor will now be ignorant about whether the advisee will use her newfound ideality to battle the demons or to take their side; otherwise she would know whether she would battle the demons or take their side, which we have stipulated she must not. But since what one ought to do is a function of one's advisor's preferences, and one's advisor, being in the dark, will have no preferences about this matter, what one ought to do will be indeterminate, even if one choice prevents the never-ending torture of all sentient life and the other does not. This is a bad prediction for a theory to make.

4.2. In the Beginning Was the Deed

Recall the earlier concession that deliberation can serve as a basis for knowledge. Perhaps this means omniscience is compatible with deliberation because it is based on deliberation. Imagine an agent who initially starts with enough uncertainty that she can deliberate, making choices about his future actions. As her system of plans for what she will do in the future gets filled in, more knowledge about the future is given. Eventually she is made fully omniscient, after she has planned out what she will do for each point of decision she will reach in her life. Such an agent would be omniscient but would have exercised deliberative powers.

The obvious problem here is that even if the agent is omniscient, her choices are not the product of that omniscience. If the advice, for example, of the Ideal Agent looks especially good, so that it is not only *evidence* of what I ought to do, but I ought to do it *because* of that advice, that advice must be fully informed. It had better be the case that I never have to worry that she only advised me to do *this* because she was ignorant of the consequences. And if it turns out that she was ignorant when she decided to advise me to do *this,* the fact that she was made omniscient afterwards is not at all reassuring—especially if you go on to add that that very same omniscience prevents her from changing her mind once the plans are in place.

4.3. Indeterminism

One might think that omniscience leads to fatalism only in deterministic worlds.[23] Perhaps in indeterministic worlds there is no fact of the matter about what will happen in the future, and so an agent who was certain of all the facts could still deliberate about her future conduct.[24] Maybe—but an Ideal Agent theorist who says this now faces a dilemma. She can offer a disjunctive account of reasons, so that what it is for an agent in an indeterministic world to have most reason to *A* is for her ideal counterpart to perform/advise *A;* but, in deterministic worlds, reasons are explained by something else, have a different nature. The other option is to embrace practical nihilism for deterministic worlds. Both options look like reductios.[25]

[23] For another problem deterministic worlds may generate for Ideal Agent theories, see (Hare 2011).

[24] We should be careful here about what we mean by "no fact of the matter." If the future is completely random, it is unclear how any rational action could be possible. Presumably, what we must have in mind is that the future is probabilistic. However, if we take this position, it is unclear whether indeterminism would really buy us the possibility of an omniscient deliberator. First, (Levi 1997) argues that even assigning probabilities to one's future actions is incompatible with deliberating about them. Second, the omniscient agent's omniscience will presumably include certainty about the objective chances of all possible future events. However, if I am non-deliberatively certain that I am 70% likely to *A,* it is still unclear whether I can deliberate about whether to *A*, for the same reason that certainty that I am 100% likely to *A* rules out deliberation. Thanks to Jiji Zhang for calling these points to my attention.

[25] Keep in mind that the second option is not just the denial that agents in deterministic worlds are responsible, which is a reasonable enough position, but the much stronger view that there is never any reason for such agents to act in one way or another, a view with considerably less intuitive

Could we take the agent in the deterministic world and go to the nearest indeterministic world where she has an omniscient and rational counterpart, and use her as the Ideal Advisor?[26] Maybe, but this view will have to answer difficult questions before it can be taken seriously as a way of saving Ideal Agent theory.

We need to know that counterfactuals such as "You would *A* if you were omniscient and perfectly rational and the world were indeterministic" are evaluable when uttered in deterministic worlds. But it is not clear that anything could have my essential properties in a world with radically different physical laws. Let us say the actual world is deterministic. Can any of the animals in an indeterministic world be of the same species as human beings, given that the workings of their bodies must be governed by different processes? Or, if the identity in question is the identity of a Lockean person, as opposed to the human animal, is enough psychological similarity retained if we go to a world where significantly different laws govern the evolution of that psychology? If the relevant essential properties or psychology similarities cannot be co-instantiated given significantly different natural laws, then the Ideal Agent theorist is still invoking impossibilities in her explanations.

It might be tempting to think that the Ideal Agent theorist could evade this potential quagmire by taking a view of *de re* modality similar to Lewis's (1986). Whether someone (or something) qualifies as my counterpart is largely a matter of conversational context. As he puts it:

plausibility. Robert Kane (1998, chapter 6) does argue that without incompatibilist free will, certain values, such as creativity, love, and friendship may fail to obtain—but even this falls short of maintaining that there are no reasons for action at all in a deterministic world. See (Arpaly 2006, chapter 2) for criticisms of Kane's position.

[26] Thanks to the referee of an earlier version of this paper for asking me to consider this objection.

> Could Hubert Humphrey have been an angel? A human born to different parents? A human born to different parents in ancient Egypt? A robot? A clever donkey that talks? An ordinary donkey? A poached egg? …
>
> You could do worse than plunge for the first answer to come into your head, and defend that strenuously. If you did, your answer would be right. For your answer itself would create a context… such as to make your answer true. … I suggest those philosophers who preach that origins are essential are absolutely right—in the context of their own preaching. They make themselves right: their preaching constitutes a context in which *de re* modality is governed by ways of representing… that requires match of origins. But if I ask how things would be if Saul Kripke had come from no sperm and egg but had been brought by a stork, that makes equally good sense. I create a context that makes my question make sense…
>
> (251-2)

But if the Ideal Agent Theory takes this position on the counterpart relation, then the question of whether I have any reasons at all is also a matter of conversational context. If someone tells me I have most reason to read more and watch less TV, that may well be true. But if someone tells me that there is not any reason for me to do anything, that is also correct. He is creating a context in which I cannot be represented *de re* by an omniscient being. But that means that if I disagree, if I tell him that no, some things really do matter, I am just being conversationally uncooperative. Yet disagreements about nihilism seem genuine.

Can we give up on the idea that the Ideal Agent is a counterpart? The costs of this were discussed in section 4.1.

4.4. Counterpossible Explanations

Can the Ideal Agent theorist simply ignore these problems by allowing that she might be offering a counterpossible explanation of reasons, rather than a counterfactual explanation?[27] This move is not straightforwardly inconsistent, unless one thinks counterpossible claims are always non-evaluable—a position that might be hard to square with seemingly valid reductios. Nonetheless, it is theoretically awkward.

If we allow that the *explanans* is a counterpossible, we seem to give up any sense that practical reasons could be normatively authoritative, at least for us. I have some sense of why it is more rational for me to do what my omniscient, perfectly rational counterpart would do (or advise). But if you tell me that my omniscient counterpart is perfectly rational only so long as that counterpart exists in a world with different rational norms, or a world in which one can suffer a rational failing and still be perfectly rational, or something similar, I lose all grip of why I should regard the counterpart as authoritative. I, after all, live somewhere where none of these things are true—where they make you less rational, and hence less authoritative.

The Ideal Agent is supposed to determine the correct application of an *ought* that is inescapable. Inescapability means, in part, that it applies to me whatever my circumstances might be—I cannot get out of the requirement by caring about something else, or by living in a culture where these requirements are not enforced. This aspect of inescapability makes

[27] Thanks to Jamin Asay for presenting this objection.

sense, so long as we restrict the circumstances we consider to the normatively possible. But now we have opened up modal space to consider scenarios on the other side of what is possible, where different, local rationalities are in effect. If an agent addresses me from across that border, inescapability is lost; her authority is for a foreign jurisdiction.

4.5. Almost Perfect[28]

Must my ideal counterpart be *perfectly* rational? Why can't I settle for *as rational as possible*? The motivation behind Ideal Agency theories—of the globalist variety especially—was the idea that what I have most reason to do is constituted by what I would do, were I exercising all the capacities I engage in familiar cases of deliberation to their maximal extent. Don't the arguments in this paper simply show that the maximal is slightly more limited than we might have initially assumed?

The problem is, once we allow that epistemic powers and powers of self-control can clash, there are *three* obvious ways in which to scale down the perfection of the ideal counterpart. We can limit the agent's epistemic capacities, so that she no longer knows her own future actions (at least insofar as such knowledge blocks deliberation). We can deny her the power of self-control, insisting that her practical attitudes update unreflectively in the direction of rational coherence. Finally, we can allow that she deliberates despite the knowledge, even if doing so involves reasoning with a kind of double-mindedness, or wishful thinking, or insensitivity to what one knows. Our problem is one of harmonizing two rational powers; so we can scale back the first *or* the second, or we can give up on harmony.

[28] Thanks to an anonymous referee for pushing this line of objection.

We have, then, at least three nearly perfect agents, who will potentially give conflicting advice (or otherwise act differently) for different situations; so it must be that one of the three is optimal—that her irrationalities are the least severe—otherwise all three will have equal claim to explaining practical reasons, and so selecting any of them as the *explanans* will be *ad hoc*. But it may well be that all three are equally irrational, or that their comparative degrees of irrationality are indeterminate, so that it is mistake to expect optimality at all. After all, each agent is guilty of a very different kind of rational failing (ignorance versus lack of self-control versus internal incoherence), and it is unclear how to assign values to these failings on some scale of all-things-considered irrationality. But unless there is such a scale, and unless it does tell us that one of these agents is optimal, picking out any particular agent as the one whose actions or advice constitute the facts about practical reasons will be arbitrary.

Intuitively, it may seem that ignorance is the least significant of the three rational failings. Some of the difficulty here is that treating ignorance as a form of irrationality at all is counterintuitive; nonetheless, it is a requirement of Ideal Agent theories that we do so. Partly, this is for simple extensional adequacy: the ignorant will fail to take into account certain facts that intuitively count as reasons. But more importantly, full omniscience seems to involve, as we noted before, the fullest exercise of the evidence-gathering and reasoning powers we make use of when deliberating, and our account of practical reasons must be in terms of the full exercise of those powers. This casts some doubt on the evidential value of intuitions—the theory was already committed to something counterintuitive on exactly this point—but there are two additional reasons to be skeptical of intuitions that ignorance is the least significant failing.

First, ignorance of one's own future actions may free-up powers of deliberative self-control. But, as we saw in section 4.1, it still leads to incredible predictions in any scenarios in which how one would act while ideal is intuitively a reason for or against the action. Consider again the case in which drinking a potion that will make one ideal is the only way to fend off a universe-wide invasion by torture demons. Unless one's unreflective motives are extremely perverse, it seems the advice of the counterpart who is fully omniscient but lacks powers of critical self-reflection and control will be a much better guide to action than the counterpart who is ignorant of how she would act—and hence ignorant of how her advisee would act upon drinking the potion. The latter agent will have no advice to give, after all. (Of course, there will be many situations in which the slightly ignorant but self-reflective agent gives the superior advice; but this just emphasizes the problem that the severity of rational failings may depend on context.)

Second, the agent in question cannot simply be ignorant of her own future actions—because her performance of those actions is entailed by a great many other facts, such as the physical laws and the current physical state of her brain; or her own character, beliefs, and motives. The agent must either be ignorant of these facts as well, expanding the cases for which the Ideal Agent theorist must make counterintuitive predictions, or else she must fail to draw appropriate inferences from her beliefs. Keep in mind, her knowledge is extensive enough that what she knows will often *deductively* entail that she performs a given action. What's more, being nearly omniscient, she will generally know what her beliefs are and the rules governing deductive consequence (and to the extent that she doesn't, again the problem of counterintuitive predictions will loom). So she will know that she has failed to draw the conclusion that deductively follows from her beliefs, and yet still fail to draw that conclusion. This looks like a more substantial form of irrationality than simple failure to know certain facts about oneself.

In short, there is no principled argument against lowering the standards of rationality an ideal agent must meet. This may be the best way for the theory to develop. But there are several different ways in which rationality could be sacrificed, and so unless we have some way of identifying the optimal tradeoffs, we are left with a series of distinct, nearly perfect agents, each of whom has equal claim on being my advisor or model.

5. Conclusion

The Ideal Agent's acts or advice are authoritative on account of both her knowledge and her perfectly functioning rational capacities. But introducing her as an explanatory device assumes that all of our rational capacities could function perfectly at the same time. The conflict between deliberation and prediction suggests otherwise: one capacity exercised to its maximal extent will prevent the functioning of another.

A problem with the divine nature occurred to Kant when considering the proofs of his rationalist predecessors for the existence of God. God was the most real being, which the rationalists took to mean that He possessed all positive (or real) properties to their maximal extent. But, Kant pointed out, having some properties might preclude having others; the manifestation of one power might undermine or cancel out the manifestation of another. If this were so, a being that contained all realities, or possessed all positive properties, would be impossible (1763, 2:85-7).[29]

[29] This is of course a simplified summary of Kant's argument. Thanks to Nick Stang for informing me of the argument and explaining to me how it works.

He was on to something. And the problem applies just as much to the seemingly less metaphysically fraught God-surrogates of Ideal Agent theories, the merely hypothetical and godlike. For if rational powers fight for turf, the godlike are not even hypothetical, but impossible.[30]

Works Cited

Arpaly, Nomy (2006) *Merit, Meaning, and Human Bondage: An Essay on Free Will,* Princeton University Press.

Arpaly, Nomy, and Timothy Schroeder (2012) "Deliberation and Acting for Reasons," *The Philosophical Review,* 121/2: 209-39.

Baker, Derek (2012) "Knowing Yourself—and Giving up on Your Own Agency in the Process," *Australasian Journal of Philosophy,* 90/4: 641-56.

Brandt, Richard (1972) "Rationality, Egoism, and Morality," *The Journal of Philosophy,* 69/20.

Carlson, Erik (2002) "Deliberation, Foreknowledge, and Morality as a Guide to Action," *Erkenntnis* 57: 71-89.

Enoch, David (2005) "Why Idealize?" *Ethics* 115/4: 759-87.

Firth, Roderick (1952) "Ethical Absolutism and the Ideal Observer," *Philosophy and Phenomenological Research* 12/3: 317-45.

Fischer, John Martin (1992) "Recent Work on God and Freedom," *American Philosophical Quarterly,* 92/2: 91-109.

[30] Thanks to Jamin Asay, David Faraci, Barry Maguire, John Maier, Dan Marshall, Tristram McPherson, Kelly Trogdon, Jian Shen, Nick Stang, Jack Woods, Jiji Zhang, the audiences at the Chinese University of Hong Kong and Hong Kong University, the anonymous referee at *Philosophy and Phenomenological Research,* and referees of earlier versions of this paper for criticisms, feedback, questions, and ideas. The research appearing in this paper was substantially funded by a research grant from the Research Grants Council of Hong Kong Special Administrative Region, China (LU342612).

Fischer, John Martin, and Neal A. Tognazzini (2014) "Omniscience, Freedom, and Dependence," *Philosophy and Phenomenological Research* 88/2: 346-67.

Gibbard, Alan (1990) *Wise Choices, Apt Feelings,* Harvard University Press.

Hare, Caspar (2011) "Obligation and Regret when There Is No Fact of the Matter about What Would Have Happened if You Had Not Done What You Did," *Nous* 45/1: 190-206.

Hedden, Brian (2012) "Options and Subjective Ought," *Philosophical Studies* 158/2: 343-60.

Holton, Richard (2009) *Willing, Wanting, Waiting,* Oxford University Press.

Hubin, Donald (1996) "Hypothetical Motivation," *Nous* 30/1: 31-54.

Hunt, David P. (1997) "Two Problems with Knowing the Future," *American Philosophical Quarterly,* 34/2: 273-85.

Kane, Robert (1998) *The Significance of Free Will,* Oxford University Press.

Kant, Immanuel (1763) *The Only Possible Argument in Support of a Demonstration of the Existence of God;* translated and printed in *The Cambridge Edition Works of Immanuel Kant: Theoretical Philosophy 1755-1770,* ed. and trans. David Walford in collaboration with Ralf Meerbote, Cambridge University Press, 1992.

___________ (1785) *The Groundwork to the Metaphysics of Morals;* translated and printed in *The Cambridge Edition of the Works of Immanuel Kant: Practical Philosophy,* ed. and trans. Mary Gregor, Cambridge University Press, 1996.

Kapitan, Tomis (1990) "Action, Uncertainty, and Divine Impotence," *Analysis* 50/2: 127-33.

Levi, Isaac (1997) *The Covenant of Reason,* Oxford University Press.

___________ (2007) "Deliberation *Does* Crowd out Prediction," in *Hommage á Wlodek: Philosophical Papers Dedicated to Wlodek Rabinowicz,* eds. T. Ronnow-Rasmussen, B. Petersson, J. Josefsson, and D. Egonsson; www.fil.lu.se/hommageawlodek.

Lord, Errol (*forthcoming*) "Abilities and Obligation," *Oxford Studies in Metaethics vol. 10,* Oxford University Press.

Louise, Jennifer (2009) "I Won't Do It! Self Prediction, Moral Obligation, and Moral Deliberation," *Philosophical Studies* 146/3: 327-48.

Maier, John (2013) "The Agentive Modalities," *Philosophy and Phenomenological Research* 87/3: 113-34.

Österberg, Jan (1998) “A Problem for Consequentialism,” in *Not Without Cause: Philosophical Essays Dedicated to Paul Needham on the Occasion of his Fiftieth Birthday,* L. Lindahl, J. Odelstad, and R. Sliwinski (eds.), Uppsala University, Uppsala.

Parfit, Derek (2011) *On What Matters, Volume One,* Oxford University Press.

Portmore, Douglas W. (2011) *Commonsense Consequentialism: Wherein Morality Meets Rationality,* Oxford University Press.

Rabinowicz, Wlodek (2002) “Does Practical Deliberation Crowd out Self- Prediction,” *Erkenntnis,* 57/1: 91-122.

Railton, Peter (1986) “Moral Realism” *The Philosophical Review,* 95/2: 163-207.

___________ (2004) “How to Engage Reason: the Problem of Regress” in *Reason and Value: Themes from the Moral Philosophy of Joseph Raz,* eds. J. Wallace, P. Pettit, S. Scheffler, and M. Smith, Oxford University Press.

Rawls, John (1971) *A Theory of Justice*, revised and reprinted 1999, Harvard University Press.

Rosati, Connie (1995) “Persons, Perspectives, and Full Information Accounts of the Good,” *Ethics* 105/2: 296-325.

Ross, Jacob (2013) “Actualism, Possibilism, and Beyond,” *Oxford Studies in Normative Ethics vol. 2,* ed. M. Timmons, Oxford University Press.

Scanlon, T. M. (1998) *What We Owe to Each Other,* Belknap Harvard University Press.

Smith, Michael (1994) *The Moral Problem,* Blackwell.

___________ (2004a) *Ethics and the A Priori: Selected Essays in Ethics and Moral Psychology,* Cambridge University Press.

___________ (2004b) “Internal Reasons” in (Smith 2004a).

___________ (2004c) “The Incoherence Argument: A Reply to Shafer-Landau” in (Smith 2004a).

___________ (2004d) “Rational Capacities,” (Smith 2004a).

___________ (2004e) “In Defense of *The Moral Problem:* A reply to Brink, Copp, and Sayre-McCord,” in (2004a).

___________ (2004f) “Exploring the Implications of the Dispositional Theory of Value,” in (Smith 2004a).

___________ (2013) "A Constitutivist Theory of Reasons: Its Promise and Parts," *Law, Ethics, and Philosophy,* 1: 9-30.

Sobel, David (1994) "Full Information Accounts of Well-being," *Ethics* 104/4: 784-810.

Timmerman, Travis (*forthcoming*) "Does Scrupulous Securitism Stand-up to Scrutiny? Two Problems for Moral Securitism and How We Might Fix Them," *Philosophical Studies.*

Todd, Patrick (2013) "Prepunishment and Explanatory Dependence: A New Argument for Incompatibilism about Foreknowledge and Freedom," *The Philosophical Review,* 122/4: 619-39.

Wallace, R. Jay (1994) *Responsibility and the Moral Sentiments,* Harvard University Press.

Williams, Bernard (1981) "Internal and External Reasons" in his *Moral Luck,* Cambridge University Press.