The Toxin and the Dogmatist

Bob Beddor

Forthcoming in *AJP*.

(Please cite published version.)

Abstract

According to the dogmatist, knowing p makes it rational to disregard future evidence against p. The standard response to the dogmatist holds that knowledge is defeasible: acquiring evidence against something you know undermines your knowledge. However, this response leaves a residual puzzle, according to which knowledge makes it rational to *intend* to disregard future counterevidence. I argue that we can resolve this residual puzzle by turning to an unlikely source: Kavka's toxin puzzle. One lesson of the toxin puzzle is that it is irrational to intend to do that which you know will be irrational. This yields a simple reply to the dogmatist: it is irrational to intend to disregard future evidence because you can know in advance that it will be irrational to do so.

1 Introduction

You know that you left your copy of Hume's *Treatise* on your desk; you saw it lying there just this morning. From this knowledge, you infer that if you were to receive evidence that you didn't leave the *Treatise* on your desk—if, for example, your colleague were to tell you that she just checked your desk and saw no sign of it—then this evidence would be misleading, since it would be evidence against something true. According to the dogmatist, this entitles you to disregard such evidence.[1] Harman offers an influential statement of the dogmatist's argument:

[1]The paradox was first formulated by Kripke in a 1972 lecture to the Moral Sciences Club at Cambridge. A revised version was eventually published as Kripke (2011).

> If I know that [p] is true, I know that any evidence against p is evidence against something that is true; so I know that such evidence is misleading. But I should disregard evidence that I know is misleading. So, once I know that p is true, I am in a position to disregard any future evidence that seems to tell against p. (Harman 1973: 148-149)

The standard response to the dogmatist, also due to Harman (1973), maintains that knowledge is defeasible. Even though you know p (*I left the* Treatise *on my desk*) at t_1, if you were to acquire sufficiently strong evidence e against p at some later time t_2, you would cease to know p. And since you would cease to know p, you would also cease to know that any evidence against p is misleading. Hence it would be irrational for you to disregard e at t_2.

While Harman's solution is widely accepted,[2] Kripke (2011) observes that it leaves a residual puzzle. Imagine that at t_1 you reason as follows: "I left the *Treatise* lying on my desk. Therefore, any evidence that I didn't do so is misleading. So I should not be influenced by any such evidence." This seems highly counterintuitive, but wherein lies your mistake? Harman's solution provides no answer. After all, your reasoning takes place at t_1, prior to acquiring any evidence against p, and hence prior to losing your knowledge. To put it another way: Harman's solution explains why someone who has already received counterevidence to something they know cannot rationally ignore this counterevidence. But it does not explain why it would be irrational for someone who has yet to receive this counterevidence to *intend* not to be influenced by it.

As Kripke notes, there are two ways of forming this intention. First, you could form a *disregarding intention*: an intention to disregard any evidence against p that you might receive. For example, you could form the intention: *Even if my colleague tells me that my copy of the* Treatise *isn't on my desk, I will ignore her testimony.* Alternatively, you could form an *avoidance intention*: an intention to take measures to prevent yourself from receiving any evidence against p. For example, you could intend to steer clear of colleagues who you suspect might cast doubt on p. You could even hire an epistemic bodyguard to

[2] Ginet (1980); Sorensen (1988); Conee (2001); and Ye (2016) all defend versions of Harman's solution.

ensure that evidence against p never crosses your path. The puzzle, then, is to explain why these two species of dogmatic intentions—disregarding intentions and avoidance intentions—are irrational. Call this "the dogmatic intentions puzzle."[3]

To solve this puzzle, I suggest that we inquire into the conditions under which an intention is rational or irrational. While this topic has not received much attention from epistemologists, it has been extensively discussed in the practical rationality literature—particularly in the literature on Kavka's toxin puzzle. One conclusion commonly drawn from the toxin puzzle is that it is irrational to intend to ϕ if you know in advance that ϕ-ing will be irrational. I argue that this principle gives us just what we need to explain the irrationality of dogmatic intentions.

2 A Closer Look at the Puzzle

2.1 Fleshing out Harman's Solution

Let us start by reviewing Harman's solution to the original puzzle in a bit more detail. Harman's statement of his solution is compressed, and leaves some questions unanswered. How does acquiring new evidence defeat one's knowledge? And why would it be irrational to disregard this new evidence, once one's knowledge is defeated? Let us take each question in turn.

While there are different ways of explaining the defeasibility of knowledge, one natural story goes like this.[4] Start by assuming fallibilism, the view that one can know something without being absolutely certain of it. On this view, while knowing p requires assigning a high credence to p, the credence need not be 1.

Next, assume a standard Bayesian picture, according to which rational agents update their credences by conditionalizing on their evidence. That is, letting $c_t(\cdot)$ be a rational agent's credence function at t, and letting e_t be an agent's total evidence at t:

[3] Borges (2015) calls this the 'synchronic dogmatism puzzle.' Given the nature of the solution that I advocate, I have chosen a name that emphasizes the fact that the puzzle fundamentally concerns the rationality of certain intentions.

[4] My reconstruction of Harman's solution draws on Lasonen-Aarnio (2014). See also Conee (2001) and Ye (2016) for similar reconstructions.

Update by Conditionalization: For all $t_1, t_2 : c_{t_2}(p) = c_{t_1}(p \mid e_{t_2})$.

Putting these pieces together yields a simple model of knowledge defeat. Suppose that at t_1 your credence in p is .96, and that the credal threshold for knowledge is .95. And suppose that at t_2 you receive some counterevidence to p. Rationality requires you to conditionalize on this evidence. If the counterevidence is strong enough, this will lead your credence in p to dip below the threshold for knowledge. More generally: new evidence can compel you to lower your credence to depths at which knowledge can no longer survive.[5]

This model also helps answer our second question: once some evidence comes along and destroys your knowledge, why is it irrational to disregard it? Consider what it means to *disregard* some evidence. Taking our cue from Lasonen-Aarnio (2014: 421-424), a plausible necessary condition on disregarding evidence is that it involves a failure to revise one's credences. That is, if you disregard some evidence e as it bears on p, then you fail to adjust your credence in p in light of e. But this would mean that you fail to update by conditionalizing on e, which—given Update by Conditionalization—is irrational.

While by no means uncontroversial, I think this way of developing a Harman-style solution to the original puzzle holds considerable appeal. In particular, it proceeds from independently motivated premises: fallibilism, together with a widely accepted updating rule.[6] Going forward, I propose to take this solution on board. What I wish to explore is whether we can extend this solution to account for the irrationality of dogmatic intentions.

[5] As Lasonen-Aarnio notes, this framework places constraints on one's theory of evidence (2014: 425-426). In particular, it is inconsistent with Williamson's (2000) identification of one's evidence with one's knowledge. (After all, if E = K, then Update by Conditionalization requires one to have credence 1 in everything one knows. But this contradicts our fallibilist assumption that knowledge does not require certainty.) A more congenial theory of evidence, for the purposes of the present framework, is that one's evidence consists in those propositions one knows with certainty (cf. Beddor (2016): chp.3).

[6] Lasonen-Aarnio (2014) criticizes this Harman-style solution on the grounds that it conflicts with an 'Entitlement' principle, which holds that if S knows, at t, that some evidence e is misleading with regard to p, then at t S is entitled to disregard e as it bears on p. A full discussion of Lasonen-Aarnio's challenge is beyond the scope of this paper. However, my own view is that the initial plausibility of Entitlement tends to dissipate once one recognizes that it conflicts with the conjunction of fallibilism and Update by Conditionalization. For criticism of Entitlement, and an attempt to debunk some of the intuitions motivating it, see Ye (2016).

2.2 Fleshing out the Dogmatic Intentions Argument

In order to diagnose the dogmatist's error, it will be useful to have a more precise formulation of the dogmatist's argument. Taking our cue from Kripke (2011: 43-44), here's one way of filling in the details. Start with a boilerplate single-premise closure principle:

Closure: For all p, q: if S knows p, S knows p entails q, and S competently infers q on the basis of this knowledge, then S knows q.

Next, add a claim about what it means for evidence to be 'misleading'. For evidence to be misleading is for it to count against a truth:

Misleading: For all p: if p, then any evidence against p is misleading.

Finally, add a general principle connecting the known consequences of an action with the rationality of intending that action:

Rationality of Intending the Good: Suppose S knows at time t_1 that ϕ-ing at time t_2 will lead to consequence c. If S knows at t_1 that c is desirable, it is rational for S to intend (at t_1) to ϕ at t_2.

A principle along these lines is *prima facie* plausible, at least if 'desirable' is read as meaning 'all-things-considered desirable'. As Kripke notes: "Suppose [someone] knows that if he opens the door, someone standing outside is going to shoot him. It would thus be a reasonable thing for him to resolve not to open the door" (2011: 44).

But if we accept these three principles, trouble is in store. Suppose you know p, and you also know Misleading. And suppose that on the basis of this knowledge, you competently infer that all evidence against p is misleading. By Closure, you know that all evidence against p is misleading. And suppose that you want, above all else, to avoid error about p. Finally, suppose you realize that if you disregard evidence against p, or avoid its acquisition in the first place, you will be able to avoid error about p. It follows—by Rationality of Intending the Good—that it is rational for you to intend either to disregard such evidence, or to avoid acquiring it.

Our *prima facie* plausible principles led to a paradoxical conclusion. And this time we cannot appeal to Harman's solution to escape the paradox. After all, your reasoning takes place prior to receiving any actual counterevidence to p, and hence prior to being rationally required to conditionalize on such evidence.

2.3 Looking Forward

In searching for a solution, a natural strategy is to take a closer look at the conditions under which intentions are rational or irrational. Fortunately this is a topic that has received considerable scrutiny in the practical rationality literature. In the next section, I'll argue that a well-known case from this literature—the toxin puzzle—supports the following constraint: *It is irrational to intend to ϕ if you can know in advance that ϕ-ing is irrational.* I go on to use this constraint (§§4-5) to provide a solution to the dogmatic intentions puzzle.

3 Toxic Intentions

Here is Kavka's toxin puzzle:

> *Toxin:*
> You have just been approached by an eccentric billionaire who has offered you the following deal. He places before you a vial of toxin that, if you drink it, will make you painfully ill for a day, but will not threaten your life…The billionaire will pay you one million dollars tomorrow morning if, at midnight tonight, you intend to drink the toxin tomorrow afternoon. (Kavka 1983: 33)

Many have drawn the conclusion that you cannot rationally intend to drink the toxin. After all, you know that by tomorrow morning you will either have the million dollars in your bank account or you won't. And so you know that, come tomorrow afternoon, you will have no reason to imbibe the toxin and every reason not to do so.[7]

[7]Perhaps you can intend to take steps aimed to induce your future self to drink the toxin—e.g., going to a hypnotist, or paying a hitman to kill you unless you drink the toxin. But intending to take such steps seems importantly different from simply intending to drink the toxin. Moreover, following Kavka, we can stipulate that the billionaire's offer forbids such measures.

This suggests a principle connecting the irrationality of performing an act with the irrationality of intending to perform that act:

Irrationality of Intending the Irrational: It is irrational for S to intend at t_1 to ϕ at t_2 if at t_1 S is in a position to know that ϕ-ing at t_2 will be irrational.

Of course, this principle also stands in need of explanation: why does an intention to perform an action you know to be irrational inherit that action's irrationality? Discussions of the toxin puzzle frequently locate the explanation in the functional role of intention. By now, at least two explanations have emerged in the literature. The first is suggested by Kavka:

> [Intentions are] dispositions to act which are based on *reasons to act*... Thus we can explain your difficulty in earning a fortune: you cannot intend to act as you have no reason to act, at least when you have substantial reasons not to act. (1983: 35)

Kavka's explanation is compressed, but here is one way of reconstructing the basic idea. An intention to ϕ is a commitment to ϕ on the basis of sufficiently strong reasons—reasons that one will possess at the time of action. But it would be incoherent to commit oneself, at t_1, to ϕ-ing on the basis of sufficiently strong reasons at t_2 if one knows at t_1 that one will lack such reasons at t_2. Hence the irrationality of intending the irrational.

An alternative explanation of Irrationality of Intending the Irrational is provided by a 'cognitivist' theory of intentions, which holds that intentions require beliefs. According to a strong version of this requirement, intending to ϕ requires believing that one will ϕ (Grice 1971; Harman 1976). Some have questioned this strong belief requirement, arguing that I might intend to submit a paper by Thursday, even though I know that I often fail to budget enough time for proofreading, hence there's a good chance I won't get the paper off until Friday. But even if counterexamples along these lines undermine the strong belief requirement, a weaker belief requirement survives. According to this weaker requirement, intending to ϕ requires believing that it's sufficiently likely that one

will ϕ, or, at the very least, that there's some chance that one will ϕ (Audi 1973; Wallace 2001). This weaker requirement seems quite plausible: I cannot rationally intend to submit the paper by Thursday if I know that there's no chance I'll submit the paper by then. Moreover, this weaker requirement suffices to explain why it is irrational to intend to act in ways that you know will be irrational. Presumably, in *Toxin* you know that tomorrow afternoon you will behave rationally. And so you know that tomorrow afternoon you will decline the toxin. By the weak belief requirement on intention, a rational agent cannot intend today to drink the toxin tomorrow.

We thus have two candidate explanations for why it is irrational to intend to do that which you know will be irrational. For present purposes, we need not choose between the two explanations. Rather, what's important is the principle itself, Irrationality of Intending the Irrational. The fact that there are two promising explanations for this principle is further reason to find it plausible.

Equipped with this principle, we now return to the dogmatic intentions puzzle.

4 Disregarding Intentions Defeated

4.1 The Dogmatist's Error

If we accept Irrationality of Intending the Irrational, then we must reject—or at least qualify—the dogmatist's principle, Rationality of Intending the Good (§2.2). Suppose you know today that ϕ-ing tomorrow will lead to a desirable outcome, but will also be irrational. Can you rationally intend, today, to ϕ tomorrow? Rationality of Intending the Good says 'Yes.' Irrationality of Intending the Irrational says 'No.'

The upshot, then, of our foray into toxin territory is that it is not always rational to do that which you know will bring about desirable consequences. It also has to be consistent with your knowledge that so acting will be rational at the time of action. I submit that this is where the dogmatist's argument goes astray. In particular, by combining Irrationality of Intending the Irrational with our Harman-style solution to the

original dogmatism puzzle, we get a story about why it is irrational to intend to disregard counterevidence to that which you know.

Suppose that a rational agent S knows p at t_1. Harman's solution tells us that if S were to gain sufficiently strong evidence e at t_2, it would be irrational for S to disregard e as it bears on p. Now suppose that at t_1 S also knows—or is in a position to know—all of the foregoing; that is, she knows that if she were to gain e at t_2, it would be irrational for her to disregard it. (Perhaps she knows this because she knows our Harman-style solution to the original dogmatism puzzle.) By Irrationality of Intending the Irrational, it is irrational for S to intend, at t_1, to disregard e at t_2.

This provides a simple account of why disregarding intentions are irrational.[8] Our account relied on just two ingredients: a widely accepted Harman-style solution to the original dogmatism puzzle, and a principle connecting up the foreseeable irrationality of an act with the irrationality of intending to perform that act. Let me now consider two natural objections that arise for this explanation.

4.2 The Wrong Kind of Irrationality Objection

A first objection is that my solution mixes up epistemic and practical irrationality. According to this objection, intending to ϕ when you know that ϕ-ing will be irrational is *practically* irrational. But, intuitively, dogmatic intentions are *epistemically* irrational.

However, my solution can accommodate this intuition. Plausibly, if you intend to do something you know to be irrational, your intention inherits whatever species of irrationality infects the object of your intention. Drinking the toxin is practically irrational, so intending to drink the toxin is practically irrational. Disregarding counterevidence is epistemically irrational, so intending to disregard counterevidence is epistemically irrational.

[8] This does not yet tell us why avoidance intentions are irrational. Stay tuned.

4.3 Can We Sometimes Intend the Irrational?

A more worrisome objection is that, in order for it to do the work I require, Irrationality of Intending the Irrational needs to be a very strong principle. In particular, in order for it to be inconsistent with Rationality of Intending the Good, it needs to be understood as claiming that it is *always* irrational to intend to act in ways that you know will be irrational, even if you also know that so acting will bring about a desirable consequence. However, it is not clear that the toxin puzzle supports this strong claim. In the toxin puzzle, you not only know, at midnight, that quaffing the toxin tomorrow will be irrational. You also know that toxin-quaffing will not lead to any desirable outcome. (Come tomorrow, you will either have the million dollars already, or you won't.)

While this is an important challenge, I think it can be met. Consider a variant toxin scenario where you know in advance that drinking the toxin would be both desirable and irrational:

> *Toxin with Amnesia:*
> Another day, another eccentric billionaire. This one offers to give you a million dollars if you drink the toxin tomorrow afternoon. As usual, there's a catch: tomorrow morning he will erase your memories of the details of the offer. And so when you find a noxious toxin left on your doorstep tomorrow afternoon, you will have no idea that drinking it will make you rich, only that it will make you retch.

Can you rationally intend, now, to drink the toxin tomorrow? It seems to me that you cannot. After all, while you now know that there will *be* a reason for you to drink tomorrow, you also know this is not a reason you will *have*. Come tomorrow afternoon, all of your *possessed* reasons will counsel against consumption.[9]

This shows that even when you know that an act will lead to a desirable consequence, you still cannot rationally intend to perform this act if you also know that it will be ir-

[9] Perhaps you can intend now to take steps aimed to induce your future self to drink—e.g., tattooing your body with *Memento*-like messages explaining the benefits of drinking. But, as in the original toxin scenario, intending to take such steps seems importantly different from simply intending to drink. And, as in the original, we can stipulate that the fine print of the billionaire's offer forbids such shenanigans.

rational at the time of action. Our strong interpretation of Irrationality of Intending the Irrational is thereby vindicated.

The strong interpretation is also supported by theoretical considerations. As we saw in §3, we would like to explain our intuitions about *Toxin* by appealing to the functional role of intentions. There we reviewed two promising explanations: the view that intentions are commitments to act on the basis of reasons, and the cognitivist view of intention. Both of these explanations support the strong version of Irrationality of Intending the Irrational. Start with the view that intentions are commitments to act on the basis of reasons that you will *possess* at the time of action. This predicts that you cannot intend, at t_1, to ϕ at t_2 if you know at t_1 that none of the reasons you will possess at t_2 support ϕ-ing. Next, consider the cognitivist proposal that intending to ϕ requires believing that one will ϕ, or that there's a chance one will ϕ. If you're certain that your future self will act rationally, then you'll think there's no chance that your future self will perform an irrational act that leads to a desirable outcome. Thus, both the reasons-based theory of intentions and the cognitivist theory support the view that it is *always* irrational to intend to behave in ways you know will be irrational.

5 Avoidance Intentions Defeated

5.1 First Steps

Thus far I've told a tale about why disregarding intentions are irrational. But, on the face of it, this story does not generalize to explain the irrationality of avoidance intentions—intentions to avoid getting the counterevidence in the first place. And so the resilient dogmatist may reason as follows: 'I concede that if I receive counterevidence to p, I will be rationally compelled to abandon my (true) belief in p. But if I never receive any counterevidence, I will be permitted to keep my belief. All the more reason to ensure that no counterevidence falls in my hands!'

This is a serious worry—indeed, I think it is the most troublesome version of the dogmatist's argument. But let us not despair quite yet. Note that it's not just irrational to

intend to avoid counterevidence to something you know. Intuitively, the very *act* of avoiding counterevidence is irrational.[10] For example, it's not just irrational to intend to hire an epistemic bodyguard to prevent counterevidence from falling in your hands. Rather, the very act of hiring the bodyguard is irrational.

Here too, it's natural to explain the irrationality of the intention in terms of the irrationality of the intended act. This suggests the following two-step strategy for explaining the irrationality of avoidance intentions:

> *Step One:* Explain why the act of avoiding counterevidence is irrational (and foreseeably so).
>
> *Step Two:* Appeal to Irrationality of Intending the Irrational to explain why the intention to avoid counterevidence is also irrational.

In the rest of this section, I'll offer what strikes me as the most promising way of executing Step One. (Note that readers who prefer some alternative account of the irrationality of avoiding counterevidence are free to plug it into my explanatory schema.)

5.2 A Good Result

The question of why it is irrational to avoid counterevidence is closely related to another question: why is it ever a good idea to seek out new evidence, rather than sticking with what one already knows? In the philosophy of science literature, the most influential answer to the latter question is due to I.J. Good (1967). Good's answer, in brief, is that gathering evidence will help you make better decisions. More precisely, Good showed that, given certain assumptions, it always maximizes expected value to gather and use further evidence before making a decision, provided that the cost of gathering evidence is negligible.

To illustrate, suppose that based on your current evidence your are very confident that your flight departs from gate G30. Now suppose you could either go directly to G30 or you could—at no cost to yourself—first glance at the departures monitor. Suppose you also

[10] There may be some exceptions to this, as I discuss below.

know with certainty that if you glance at the monitor, you will update by conditionalizing on your new evidence,[11] and perform whatever action maximizes expected value relative to this more informed body of evidence. Good shows that the expected value of refusing to look at the monitor is always less than or equal to the expected value of looking, and strictly less when it's possible that you'll prefer a different act—say, going to gate A20—after the glance.[12]

Applying this to the dogmatism puzzle: counterevidence is a type of evidence. By Good's Theorem, even if e is counterevidence to something you know, updating and conditionalizing on e will still maximize expected value. Now, combine this with the orthodox assumption that rationality requires maximizing expected value. This gives us the explanation we sought: it is irrational to turn down free counterevidence to that which you know, because doing so conflicts with the requirement to maximize expected value.

This is, I think, an attractive explanation. But there are a number of ways one might challenge it. Let me address three concerns.

5.3 Dominance Reasoning

First, some might worry that my use of a decision theoretic argument sits uncomfortably with my endorsement of a Harman-style solution to the original dogmatism puzzle. Here's why. Given fallibilism, knowledge does not require certainty. Suppose, then, you know that your flight is at G30, even though you are not quite certain of it. By misleading and Closure, you know that if the monitor displays a gate other than G30, it will be misleading. You also know that if the monitor misleads you, then this would lead to a worse outcome (going to the wrong gate) than simply ignoring the monitor.

Those attracted to the idea that knowledge plays an important role in practical reasoning have sometimes proposed that if S knows p, then we can omit any $\neg p$ states from

[11] Perhaps you know this because you know that you will respond to the evidence rationally, and you know that rationality requires you to update by conditionalization.

[12] For the details of the proof, see Good (1967). It is important to stress that Good's Theorem is concerned with *expected value* rather than *value*. Your glance at the monitor *could* mislead; it could cause you to head towards gate A20, when your flight is indeed at G30 (perhaps a glitch caused it to display the wrong flight information). What Good shows is not that gaining evidence guarantees success in action, but that it guarantees *expected* success in action.

S's decision table (Weatherson 2012). If this is right, we should be able to represent your decision problem along the following lines. Letting 'G30' abbreviate the proposition that your flight is at G30:

	G30 & monitor says, 'G30'	G30 & monitor doesn't say, 'G30'
Look	+10	-10
Don't look	+10	+10

But if this is what your decision problem looks like, then it seems any plausible decision theory recommends averting your eyes from the monitor! At least this follows if we assume a weak dominance constraint on rationality:

Weak Dominance: If in all possible outcomes ϕ-ing is at least as good at ψ-ing, and there is at least one possible outcome where ϕ-ing leads to a better result than ψ-ing, it would be irrational to ψ rather than ϕ.

This is puzzling. We have one decision theoretic argument that recommends looking (courtesy of Good's Theorem), but another that counsels against looking (courtesy of Weak Dominance). Where did we go wrong?

I think the mistake was in how we represented your decision problem. We stipulated that while you know that your flight is at G30, you are not certain of it—and rationally so. So by your lights there is *some* possibility that the flight is not at G30.[13] Presumably you should take this possibility into account in your practical reasoning. And so your decision problem actually looks more like this:

	G30 & monitor says, 'G30'	G30 & monitor doesn't say, 'G30'	¬G30 & monitor says, 'G30'	¬G30 & monitor doesn't say, 'G30'
Look	+10	-10	-10	+10
Don't look	+10	+10	-10	-10

[13]Some might question this. Aren't the possibilities in question *epistemic* possibilities? And aren't the epistemic possibilities those compatible with what's known? While this may be a perfectly legitimate sense of 'epistemic possibility', there's an equally legitimate alternative: the epistemic possibilities are those compatible with your *certainties*.

This requires rejecting Weatherson's proposal that we can omit a possibility from a decision table whenever the decision maker knows that it does not obtain. But I think that Weatherson's proposal should be unattractive to anyone who fully embraces the fallibilist picture. If an agent is rationally required to assign some non-zero credence to a possibility incompatible with their knowledge, presumably they should take that possibility into account in practical reasoning.

Given this alternative representation of your decision problem, we cannot use Weak Dominance to derive the result that it would be irrational to look. The apparent conflict between our application of Good's Theorem and dominance-based reasoning is resolved.

5.4 The Wrong Kind of Irrationality Objection (Round Two)

Another concern is that the 'Wrong Kind of Irrationality' objection rears its head again. According to this objection, Good's Theorem only shows that avoiding counterevidence is *practically* irrational. By the principle that intentions inherit whatever species of irrationality infects their objects (§4.2), I have only shown that avoidance intentions are practically irrational. But, intuitively, the defect is epistemic.

Here is one way of making this objection vivid: suppose Rene enjoys reading about obscure topics in the library, even though she realizes she will never put this knowledge to use. In the course of her reading, she comes to know some proposition p. She also knows that the next tome on the shelf contains further information bearing on p. Intuitively, it is irrational for her to avoid reading this book merely so as to protect her knowledge of p. On the face of it, Good's Theorem is of no help in capturing this intuition. After all, we have stipulated that Rene will never face a decision that hinges on whether p is true, and that she realizes as much.[14]

While this is a natural concern, it has a natural solution, first recognized by Oddie (1997). Oddie showed that we can transpose Good's proof into the framework of *epistemic decision theory*. To illustrate, it will be helpful to briefly introduce some basic concepts from epistemic decision theory.

[14]Thanks to an anonymous referee for raising this issue.

Standard decision theory offers a framework for calculating the expected practical value of actions. Epistemic decision theory co-opts this machinery to provide a framework for calculating the expected *epistemic* value of credal states. How should we understand this notion of epistemic value? A natural thought is that epistemic value is closely connected with *accuracy*: the closer a credence is to the truth, the more epistemically valuable it is. This can be formally captured via a *scoring rule*—a measure of how accurate an agent's credence function is. Equipped with a scoring rule, we can then use an agent's credence function c to calculate the expected epistemic value of c, as well as the expected epistemic value of any other credence function c'. Using this framework, Oddie proves an epistemic analogue of Good's Theorem. Assuming a plausible constraint on the scoring rule, the expected epistemic value of an agent's current credal state is always less than or equal to the expected epistemic value of updating this credal state with cost-free evidence, and strictly less when updating on the new evidence might change the agent's credences.[15]

Let us see how this applies to Rene. Either Rene knows p with absolute certainty or she does not. If she does, then no matter what she reads in the next volume on the shelf, her credence in p will remain 1. Hence the expected epistemic value of reading is the same as the expected epistemic value of not reading. In this case, while she is not required to read the book, it would be irrational to avoid reading merely to protect her knowledge of p, since nothing she reads could jeopardize that knowledge.

Assume, then, that she does not know p with certainty. Then reading the next volume might well bring her credence in p closer to the truth. There are two ways this could happen. If p is true, she could get some further evidence in favor of p, which might boost her credence in p even higher—thereby making it more accurate. And if p is false, she

[15]The plausible constraint is that the scoring rule is 'strictly proper': the expected epistemic value of a credence function c, when calculated using c itself, must exceed the expected epistemic value of any other credence function. This corresponds to the idea that rational agents are 'immodest': they should regard their credence function as having a better shot of getting at the truth than any other credence function. One motivation for this idea is that if rationality did not require immodesty, there would be no reason for someone to stick with their credence function rather than switching to some other, equally accurate credence function. For discussion, see Lewis (1971); Oddie (1997); Greaves and Wallace (2006). (Note that while strict propriety suffices for an epistemic analogue of Good's Theorem to hold, it is not necessary. See, for example, the scoring rule in Horwich (1982).)

could get some evidence against p, thereby lowering her credence in p—again making it more accurate. Of course, it's also possible that reading the book will make her credence less accurate: it might give her further evidence that p is true, when p is in fact false, or evidence that p is not true, when in fact it is. Here too, Good's Theorem does not tell us that reading is guaranteed to maximize epistemic value, only that it is guaranteed to maximize *expected* epistemic value.[16]

Recasting Good's Theorem in the framework of epistemic decision theory gives us a story about why avoiding evidence is epistemically—not just practically—irrational. Plausibly, just as practical rationality requires us to maximize expected practical value, so epistemic rationality requires us to maximize expected epistemic value. This explains why it is epistemically irrational for Rene to refuse to peruse the volume. Given the assumption that an irrational intention inherits whatever species of irrationality infects its object, it follows that it would be epistemically irrational for Rene to *intend* to avoid reading it.

5.5 Idealization Failure

A different objection is that Good's Theorem relies on certain idealizing assumptions. First, it assumes that agents will update by conditionalizing on their evidence. Second, it assumes that they will act in ways that maximize expected value. These assumptions may not always hold in practice. When they fail, what are we to say about avoidance intentions?

It seems to me that when these assumptions fail, avoiding counterevidence can be

[16] To get a sense of how the proof works, suppose that Rene knows with certainty that if she reads the next book she will learn one of $e_1 \ldots e_n$, where these propositions are mutually exclusive and jointly exhaustive. If she learns some e she will conditionalize on it, resulting in a credence function c_e. If she doesn't read, she will stick with her current credence function (c). Thus the expected epistemic value of not reading, by the lights of c, is simply the expected epistemic value of sticking with c:

$$EV_c(\text{don't read}) = EV_c(c) = \sum_e EV_{c_e}(c) \cdot c(e). \tag{1}$$

The expected epistemic value of reading, again by the lights of c, is the expected epistemic value of updating Rene's credence function on whatever she learns from reading:

$$EV_c(\text{read}) = \sum_e EV_{c_e}(c_e) \cdot c(e). \tag{2}$$

If epistemic value is calculated using a strictly proper scoring rule, then $EV_{c_e}(c_e) \geq EV_{c_e}(c)$, with inequality if $c_e \neq c$. And so if reading the book might change Rene's credence function, $EV_c(\text{read})$ exceeds $EV_c(\text{don't read})$.

perfectly rational. Start by considering a case where the first assumption fails. Suppose you know from experience that you tend to give disproportionate credence to anything published on BuzzFeed. As a result, you avoid reading BuzzFeed, on the grounds that if you were to read it, you might respond to it irrationally. Here it seems that the act of avoiding BuzzFeed is perfectly rational, as is the intention to avoid it.

Next, consider cases where the second assumption fails. Ordinary people are not always indifferent between two gambles with equal expected value. They sometimes prefer gambles that have a better worst case outcome. Such agents are *risk-avoidant*, in the technical sense that they are unwilling to accept the possibility of a loss for an equivalently-sized possibility of a gain. Buchak (2010) develops a risk-weighted decision theory to model such agents. She then shows that if we replace standard decision theory with risk-weighted decision theory, Good's Theorem does not always hold. To see why, consider again our airport example. While the expected value of looking at the monitor is higher than the expected value of refusing to look, there is still a chance that looking at the monitor will mislead you, causing you to go to the wrong gate. If you are risk-avoidant, then this worst case outcome is not necessarily balanced by the expected gains in the best case outcome. Hence the risk-weighted expected value of looking might well be lower than the risk-weighted expected value of averting your gaze.

While it is uncontroversial that some people *are* risk-avoidant, it is controversial whether risk-avoidance is rationally permissible. If it is not, then our Good-inspired solution still applies to such agents: they *ought* to maximize expected value, so they ought not avoid free counterevidence to their knowledge. But suppose we set this response aside; suppose we follow Buchak in allowing that there a variety of rationally permissible attitudes towards risk, and that risk-avoidance is one of them. Once we grant this, then I think it becomes much less clear that it would be irrational for risk-avoidant agents to avoid counterevidence that they think has a high probability of being misleading. And it also becomes much less clear that it would be irrational for them to intend to do so.[17]

[17]Buchak's official (2010) position is that it would be practically rational but epistemically irrational for risk-avoidant agents to turn down cost-free evidence. However, she notes in passing (p.118) that if we use epistemic decision theory to derive the constraints on epistemic rationality (as advocated here), her arguments support a different conclusion: evidence-avoidance can be both practically and epistemically ratio-

This suggests that when the assumptions underpinning Good's Theorem do not apply, it *is* rational to avoid counterevidence, and to intend to do so. If this is right, it has two important upshots. First, far from being a problem for the Good-inspired solution offered here, these cases actually count in favor of such a solution. Second, these cases reveal an important difference between disregarding intentions and avoidance intentions. Whereas the former are always irrational, the latter are sometimes rational.[18] The account offered here explains this difference. Disregarding intentions are always irrational, because disregarding counterevidence is always irrational (since rationality always requires one to update by conditionalization). By contrast, avoidance intentions are sometimes rational, because avoiding counterevidence is sometimes rational (as when one foresees that one may not respond to counterevidence rationally, or when one is rationally risk-avoidant).

6 Conclusion

I have defended a solution to the dogmatic intentions puzzle that draws on work in the practical rationality literature. One moral of the toxin puzzle is that it is always irrational to intend to do that which you know in advance will be irrational. But a rational agent can know in advance that it is irrational to disregard counterevidence to something she knows. In most circumstances, she can also know in advance that it is irrational to avoid receiving such counterevidence. Hence the irrationality of dogmatic intentions.[19]

References

Robert Audi. Intending. *Journal of Philosophy*, 70(13):387–403, 1973.

Bob Beddor. *Reduction in Epistemology*. PhD thesis, Rutgers University, 2016.

Rodrigo Borges. On Synchronic Dogmatism. *Synthese*, 192(11):3677–3693, 2015.

nal for risk-avoidant agents, at least in certain circumstances. For extended defense of this conclusion, see Campbell-Moore and Salow (manuscript).

[18]Here I agree with Kripke (2011: 49), who also acknowledges that avoidance intentions are sometimes rational.

[19]Many thanks to David Black, Cameron Boult, Simon Goldstein, Chris Kelp, Ben Levinstein, Mona Simion, John Williams, and two anonymous referees at AJP for insightful comments. I am also grateful to participants in reading groups at KU Leuven and the National University of Singapore for very helpful discussion.

Lara Buchak. Instrumental Rationality, Epistemic Rationality, and Evidence-Gathering. *Philosophical Perspectives*, 24(1):85–120, 2010.

Catrin Campbell-Moore and Bernhard Salow. Avoiding Risk and Avoiding Evidence, manuscript.

Earl Conee. Heeding Misleading Evidence. *Philosophical Studies*, 103(2):99–120, 2001.

Carl Ginet. Knowing Less by Knowing More. *Midwest Studies in Philosophy*, 5:151–161, 1980.

I.J. Good. On the Principle of Total Evidence. *British Journal for the Philosophy of Science*, 17(4):319–321, 1967.

Hilary Greaves and David Wallace. Justifying Conditionalization: Conditionalization Maximizes Expected Epistemic Utility. *Mind*, 115(459):607–632, 2006.

Paul Grice. Intentions and Uncertainty. *Proceedings of the British Academy*, 57:263–279, 1971.

Gilbert Harman. *Thought*. Princeton University Press, Princeton, 1973.

Gilbert Harman. Practical Reasoning. *Review of Metaphysics*, 29(3):431–463, 1976.

Paul Horwich. *Probability and Evidence*. Cambridge University Press, Cambridge, 1982.

Gregory Kavka. The Toxin Puzzle. *Analysis*, 43(1):33–36, 1983.

Saul Kripke. Two Paradoxes of Knowledge. In *Philosophical Troubles*. Oxford University Press, Oxford, 2011.

Maria Lasonen-Aarnio. The Dogmatism Puzzle. *Australasian Journal of Philosophy*, 92(3):417–432, 2014.

David Lewis. Immodest Inductive Methods. *Philosophy of Science*, 38(1):54–63, 1971.

Graham Oddie. Conditionalization, Cogency, and Cognitive Value. *British Journal for the Philosophy of Science*, 48:533–541, 1997.

Roy Sorensen. Dogmatism, Junk Knowledge, and Conditionals. *The Philosophical Quarterly*, 38(153):433–454, 1988.

Jay Wallace. Normativity, Commitment, and Instrumental Reason. *Philosophers' Imprint*, 1(3):1–26, 2001.

Brian Weatherson. Knowledge, Bets, and Interests. In Brown and Gerken, editors, *Knowledge Ascriptions*, pages 75–103. Oxford University Press, New York, 2012.

Timothy Williamson. *Knowledge and its Limits*. Oxford University Press, Oxford, 2000.

Ru Ye. Misleading Evidence and the Dogmatism Puzzle. *Australasian Journal of Philosophy*, 94(3):563–575, 2016.