Akrasia and Traitors

Anonymous

2025-06-19

Bar Luzon argues that akrasia is irrational because it leads to violating a principle called **Avoid Treachery**. In response, I argue that Avoid Treachery is insufficiently motivated, that it presupposes a picture of rational inference that defenders of akrasia have independent reason to reject, and that there are models in which Avoid Treachery is false.

1 The Debate

A prominent debate in recent epistemology has been whether it can ever be rational to believe propositions of the form of **SA**, or of some similar forms.

SA q and it is irrational for me to believe q.

The *enkratic* philosopher says all beliefs of that form are irrational. The *anti-enkratic* philosopher says that they are sometimes rational.

The debate here isn't always about **SA** (Simple Akrasia); some philosophers focus on **LA** (Likely Akrasia).

LA *q* and it is probably irrational for me to believe *q*.

The difference between these will be important, especially in Section 5, because Timothy Williamson (2000, 2011, 2014) has offered the most influential arguments for the anti-enkrasia position about **LA**, but agrees with the enkratic philosopher about **SA**.

Recently in this journal, Bar Luzon (Forthcoming) has argued on the side of the enkratic philosophers about **SA**. Rather than start with a mere appeal to intuitions, as many in this debate do, she starts with a principle she calls **Avoid Treachery**.

Avoid Treachery (AT) For every proposition *p* and for every positive epistemic status *E*, if one knows that [*p* has *E* for one only if *p* is false], then one ought not believe *p*.

In this principle, E ranges over the statuses epistemic justification, epistemic rationality, evidential support and epistemic permissibility, and the conditional is a material conditional. The 'ought' is purely epistemic; if one thought belief in God was justified on Pascalian grounds one wouldn't be moved by an argument from **AT**. So I'll take 'one ought not believe *p*' to just be that it's (epistemically) irrational to believe that *p*. So we can formalise **AT** as **AT Formalised**. In it, KA is that Hero knows A, RA is that Hero rationally beliefs A, and E picks out one of the four statuses from the start of the paragraph. Whichever one E picks out, EA is that *p* has that status for Hero.

AT Formalised $K\neg(p \land Ep) \rightarrow \neg Rp$

The argument for the irrationality of **SA** follows pretty quickly. Let p be $(q \land \neg Rq)$. Assume $E(A \land B)$ implies EA, that RA implies EA, and that Hero knows anything that can be proven in a few lines of logic. Then it's easy to show $K\neg(p \land Ep)$, and hence $\neg Rp$, which just is the enkratic view.

The point of this note is to argue that the anti-enkratic philosopher mostly has good reasons to reject **AT**. I say mostly because there is one argument for **AT** that might work, but only if like Williamson one treats **SA** and **LA** differently.

It would be too easy to say that the anti-enkratic view view implies **AT** is false. Of course it does, since Luzon's argument against from **AT** to the enkratic view is valid! What I want to argue is that the reasons behind the anti-enkratic view give us somewhat independent reason to reject **AT**. I'm going to offer the following arguments against **AT** in sections 3 to 6.

1. **AT** fails for other nearby values of E, and this undermines the motivation for believing it holds for these values.

2. The main argument for **AT** turns on an understanding of what it means to say evidence is a guide to truth that the anti-enkratic philosopher rejects.

3. An argument for **AT** from the idea that beliefs violating **AT** would be 'self-undermining' at most supports the enkratic view about **SA**, not **LA**.

4. There are plausible models for evidence and belief where **AT** fails.

But first it helps to rehearse the arguments for the anti-enkratic view, to see how these objections flow from them.

2 The Arguments

Simplifying greatly, the anti-enkratic view relies on one presupposition, followed by one of two (independent) arguments. The presupposition is easiest to see with an example.

Hero has a faculty meeting today, but they have forgotten about it. Fortunately, they just got a reminder email from their chair saying there's a meeting today. Now they believe, indeed know, there's a meeting today.

The presupposition is that the following three things are in principle distinct.

1. Hero's reason for believing that there is a meeting today, i.e., the email they got from the chair.

2. The facts that make the email from the chair a reason to believe there is a meeting today. Just what those are turns on the full theory of testimony, but presumably they include things like the chair's reliability, the frequency of emails being faked, and so on.

3. The reasons Hero has for believing that the email is a reason to believe there's a meeting today.

The presupposition is that 1 and 3 are distinct. The reason that they are distinct is that 1 and 2 are distinct, and 3 requires Hero to have thoughts about (or at least sensitive to) 2, while 1 does not.

With that in place, the first argument for the anti-enkrasia view starts with anti-exceptionalism about epistemology.[1] Just like Hero might not know descriptive facts

---

[1]For anti-exceptionalism about logic, see Martin and Thomassen Hjortland (2024). This kind of argu-

like when the meeting is, she might not know epistemological facts like just why the email is a reason to believe its content. If Hero can reasonably have false beliefs about descriptive facts, she can have false beliefs about what makes something a reason to believe.[2] If those beliefs are *false*, she could reasonably believe that the meeting is today, while reasonably believing that 2 fails to obtain.

The second argument relies on formal models, like the model of Williamson's unmarked clock, in which the formal translations of **SA** (or at least **LA**) are rationally believed. What's distinctive about these formal models is that while agents know the epistemic facts, they know what is rational to believe in what situation, they don't know what situation they are in. It makes the discussion clearer to have a concrete theory about what is rational in a situation, so I'll work with a very crude form of evidentialism. (Everything I say could, with some work, be repurposed for an argument that makes different assumptions about what facts about a situation are relevant, but this is an easy one to work with.) In particular I'll assume:

- What's rational to believe supervenes on one's evidnce;

- One's evidence is all and only what one knows.

- It is rational for a person whose evidence is E to believe p iff $\Pr(p \mid E) \geqslant 0.9$, where Pr is the evidential probability function.

Again, I'm not saying this theory is true; in fact it's completely implausible. What

---

ment is particularly prominent in Lasonen-Aarnio (2020).

[2]If, like Williamson, one denies that false beliefs can be reasonable, one will treat **SA** and **LA** differently. As noted earlier, I'm mostly ignoring that distinction here.

matters is that (a) what's rational to believe varies from one situation to another, and (b) someone might not know precisely what situation they are in, just like they might be ignorant of any empirical fact.

Assuming evidentialism lets us distinguish two ways in which one might be ignorant about one's situation.

- One might know p, but not know one knows it.
- One might not know p, but not know one doesn't know it.

Williamson's models typically assume the first kind of ignorance, and this has been rather controversial. But as I'll discuss in Section 6, we can get the problem going with just the second kind of ignorance.

So anti-enkratic philosophers have employed two kinds of strategies: argue that people can believe **SA (LA)** because they can rationally have false beliefs (lack true beliefs) about what is rational in a situation, or because they can rationally have false beliefs (lack true beliefs) about what situation they are in. These strategies seem to exhaust the options for the anti-enkratic philosopher. If agents always know which situation they are in, and know what's rational in every situation, they'll know what's rational for them. So they can't rationally believe p and not know they rationally believe it. But both strategies seem promising.

3 Other Statuses

The first reason to be sceptical of **AT** is that it doesn't hold for some nearby statuses a proposition might have. A simple case is that since one can rationally believe p without having Cartesian certainty that it's true, if we took E to be Cartesian certainty the principle, **AT** would be clearly false.

More interestingly, consider the case where E is *is provable in Peano Arithmetic*. That's not really an *epistemic* status, since it doesn't refer to an agent. But it's interesting to note how **AT** fails for this value of E. If p is that Peano Arithmetic is consistent, then Hero knows that p is E iff p is false. But that's no reason to reject p; indeed, Hero should believe p.

The point here is not that epistemic rationality is so analogous to Cartesian certainty, or provability in Peano arithmetic that we can simply argue by analogy that since **AT** fails when E is one of the latter statuses, it fails when E is epistemic rationality. That's a weak analogy; there are too many differences between them.[3] A better argument is that noting **AT** fails for these latter statuses puts a constraint on what any argument for an instance of **AT** must look like. An argument that **AT** holds when E is epistemic rationality better not generalise to an argument that **AT** holds for these two latter statuses. That would be a clear case of overgeneration.

That last claim is what I'll argue for in the rest of this paper. The arguments that Luzon offers for **AT** are mostly arguments that do overgenerate.

---

[3] I'm grateful here to a reviewer for steering me away from a not very plausible argument.

4 Guide

The example of provability in Peano Arithmetic is relevant to the main argument Luzon gives for **AT**. She argues that **AT** must be true for the values of E she presents because if it fails, E can't be a good guide to truth. Since justification, rationality, etc are guides to truth, **AT** must be true.

The simplest response is that this claim about **AT** can't be right in general because provability in Peano Arithmetic is a good guide to truth when discussing the natural numbers, but **AT** fails when E is provability in Peano Arithmetic. Provability is a good guide to arithmetic truth in general, even if there are cases where it is not in fact a good guide.

What would be implausible is to say that provability is the *only* guide to truth, but **AT** fails for it. Assuming that any reason is a guide to truth, then if provability were the only guide, we'd have no reason to believe that arithmetic is consistent. Assuming also that it is irrational to believe something we have no reason to believe, it would follow that we're irrational to believe that arithmetic is consistent, and we would not in fact have a counterexample to **AT**.

The argument from the last paragraph generalises. In general, it seems incoherent to say of any E that it is the only guide to truth, but **AT** fails for it. It would be nice here to have a theory about what it takes for something to be a guide, but we don't need one for the argument to work. As long as being a reason is sufficient to being a guide, the argument goes through. But the key point here is that this argument only goes through on

the assumption that E is the only guide. If there is another guide, the argument fails. Just what that other guide is in arithmetic, whether it is Gödelian intuition, or diagrammatic reasoning, or something else, is controversial.[4] But all that matters is that there is some other guide, which there must be if we rationally believe Peano arithmetic is consistent.

At this point you might think it matters that Luzon restricted E to things like evidential support. Surely the evidentialist does think that evidence is the only guide to truth. Here the presupposition I noted in Section 2 is important. When Hero believes that there's a meeting today, her guide is not that she has evidence for this: it's the email. She's guided by the fact that she received this email, not by the fact that it's evidence. If she checks her computer and sees snow is forecast, her belief that it will snow is guided by something different. That's so even though there are a few descriptions we can give which make it look like she is guided by the same thing. In both cases, for instance, she is guided by words on her computer screen. In the same sense, she is guided in both cases by her evidence. But in the most important sense, the email and the weather forecast are different guides.

This I suspect is ultimately the biggest difference between the enkratic and the anti-enkratic philosopher. The enkratic philosopher thinks that all beliefs are guided by the same thing: one's evidence. The anti-enkratic philosopher thinks different beliefs are guided by different things: the different pieces of evidence. **AT** is not, I say, a good constraint on E being a guide, but it is a good constraint on E being the only guide. So the anti-enkratic philosopher who distinguishes facts which constitute evidence from the

---

[4] The discussion of "proof chauvinism" about explanation in D'Alessandro (2020) is helpful here, though his focus is on explanation not knowledge.

fact that that fact is a piece of evidence, has good reason to reject AT.

5 Undermining

The other big argument for **AT** is that if **AT** holds, a belief that p would be self-undermining, and hence irrational. Presumably this means that the belief couldn't achieve it's aim or goal. What it is to undermine someone is to stop them achieving their aim or goal, so to be self-undermining is to do this to yourself.

Whether **AT** implies this depends on what one thinks the aim of belief is. If it's truth, then **AT** doesn't have this implication. It could be that **AT** is true with E something like evidential support, and still p is true. Indeed, part of what's puzzling about enkratic arguments is that beliefs like *p and I'm irrational to believe p* could well be true. I often have irrational beliefs, and sometimes I'm lucky and they're true!

So the argument must assume a stronger view about the aim of belief. A natural thought is that the aim of belief is knowledge. Here I think the argument for **AT** does go through. If the aim of belief is knowledge, then if p satisfies the antecedent of **AT** it can't possibly satisfy it's aim, and it's irrational to do something that can't satisfy it's aim. So given that aim, Luzon's argument goes through, and the enktratic philosopher is right to say that any belief of the form **SA** is irrational.

Note here that the overgeneralisation worries that I've been leaning on simply don't apply. No one thinks the aim of belief is Cartesian certainty, or provability in Peano arithmetic. Violations of **AT** for the case where E is one of those statuses are not self-

10

undermining because belief does not aim for those statuses. This is another reason to not think the examples in Section 3 don't work as arguments by analogy; one important feature of the statuses Luzon stresses is that they are all plausible aims (or entailments of aims) of belief.

The argument for **AT** from the aim of belief generalises in one important direction, but not in another. The important direction is if the aim of belief is such that beliefs that achieve their aim are (a) true and (b) satisfy E, then any belief which satisfies the antecedent of **AT** will be self-defeating, and hence irrational. So any such hypothesis about the aim of belief will imply, via **AT** that any belief of the form **SA** is irrational. The direction in which it does not generalise is that there isn't an argument here for **LA**. This can be seen from the fact that Williamson endorses knowledge as the aim of belief (and says that beliefs like **SA** are irrational), but also says that beliefs like **LA** can be rational. Assuming his position is even coherent, which I think we have good reason to believe is true, the anti-enkratic philosopher can coherently say that there is no argument against their position about **LA** from considerations about the aim of belief.

We could at this point debate about whether the core debate here concerns **SA** or **LA**, or about what akrasia/enkrasia really is. This does not strike me as a productive line of inquiry. It's better, I think, to note the space of possibilities here.

1. On weak theories of the aim of belief, e.g., where the aim is merely truth, the argument for **AT** does not go through.

2. On strong theories of the aim of belief, where satisfaction of the aim implies truth

11

and rationality, the argument for **AT** does work, and the anti-akratic philosopher is correct, but only about **SA** not **LA**.

3. Either way, there isn't yet an argument here for the anti-akratic view about **LA**. Such an argument would require an extra premise that it is irrational to have a belief which is probably self-undermining.

So to wrap up, I'll make a small note about the status of that missing premise, that it is irrational to have a belief that probably doesn't achieve its aim.

6 Formal Models

In several places, Williamson has used formal models to show that **LA** is compatible with the knowledge aim of belief. I think these arguments are perfectly sound, but they have been criticised in a number of ways. The following four stand out.

1. The epistemic accessibility relation in the model is intransitive, so the the KK principle fails (Das and Salow (2019)).

2. The epistemic accessibility relation in the model is not nested, so intuitive principles about the value of evidence fail (Geanakoplos ([1989] 2021), Dorst et al. (2021)).

3. The models are cases where the probability of the target proposition is not a good guide to its truth. (Horowitz (2014))

4. The models assume that updating is by conditionalisation on one's evidence, even when isn't sure precisely what one's evidence is. (Gallow (2021))

I'm going to present a model where **AT** fails[5] (and so does **LA**) even though the model is modified to avoid the first three objections. That is, I'll present a model where epistemic accessibility is transitive and nested, and in a sense I'll make precise probability is a good guide to truth, but where **AT** is still false. I won't have anything to say about the fourth objection; I think there are good defences of that assumption, but it would be a massive digression to present them here.

Onto the model. There is a random variable X whose prior probability is a uniform distribution over $[0, 1]$. If $X = x$, Hero will learn $X \leqslant x$. That is, from the world $X = x$, all worlds $X = y$ are possible, as long as $y \leqslant x$. This accessibility relation is clearly transitive and nested.

Hero will update on what they learn, i.e. $X \leqslant x$. I'll use Pr for the initial probability of some proposition, and Cr for Hero's credence after learning $X \leqslant x$.

I'm going to focus primarily on propositions of the form $X \in (a, b)$, where $0 < a < b < 1$. Call this proposition i, for interval. There are three interesting possibilities for $Cr(i)$.

1. If $X \leqslant a$, then $Cr(i) = 0$, and p is false, so that's all good.

2. If $X = b$, then $Cr(i)$ is at its highest value, $(b - a)/b$. That's not great since p is false, but it's just one point.

---

3. Otherwise $Cr(i)$ is in $((b - a)/b, b - a)$.

In the third case, there's a striking result we can prove about Cr.

For any threshold $t \in ((b - a)/b, b - a)$, $\Pr(i \mid Cr(i) \geqslant t) = t$.

That is, conditional on Hero, who is inside the model, having credence at least t in i, the probability that we, who are outside the model, should have in i is t. That is, I think, a good sign that in this case Hero's credence in i, the evidential probability of i inside the model, is correlated with the truth of i. The correlation isn't perfect, the edge case in point 2 will become important, but in general the posterior probability of i is correlated with its truth.

I won't go over the proof of this result here. It's a trivial but somewhat tedious bit of algebra. What's more interesting is to see how this affects **AT**.

Consider the case where the interval is $(0.03, 0.3)$. So i is that X is in that interval. And consider in particular the case where X is 0.3. In that case $Cr(i)$ is 0.9, which we earlier assumed was the threshold for rational belief. Indeed, this is the only point where $Cr(i)$ reaches that threshold. But in that case i is false. So the only case where it is rational to believe i, i is false. Since Hero knows the model, they are certain before receiving any evidence that if they rationally believe i, it will be false. But still, this is case satisfies all the constraints that anti-akratic philosophers have argued were missing from Williamson's earlier models. Accessibility is transitive and nested, and in a good sense the evidential probability is a good (if not perfect) guide to truth.

14

So if the anti-enkratic philosopher was moved in the first place by models like Williamson, then even taking on board the criticisms of those models, they have good reason to reject **AT**: it fails in cases like these.

7 Conclusion

So much of the literature on enkrasia consists of raw appeals to the unintuitiveness of **SA** and **LA**. So I think it's great to see actual arguments from principles like **AT** for enkratic principles. But I don't think the anti-enkratic philosophers should be changing their minds over this.

If one's initial motivation for anti-enkrasia was based in the metaphysical distinction between reasons and what makes something a reason, then there are good grounds for rejecting the idea that rationality can only be guiding if **AT** is true. And if one's initial motivation was based in the kinds of models that Williamson developed, then even taking on board the recent criticisms of those models, there are variants of his models that falsify **AT**.

D'Alessandro, William. 2020. "Mathematical Explanation Beyond Explanatory Proof." *British Journal for the Philosophy of Science* 71 (2): 581–603. https://doi.org/10.1093/bjps/axy009.

Das, Nilanjan, and Bernard Salow. 2019. "Transparency and the KK Principle." *Noûs* 52 (1): 3–23. https://doi.org/10.1111/nous.12158.

Dorst, Kevin, Benjamin A. Levinstein, Bernhard Salow, Brooke E. Husic, and Branden Fitelson. 2021. "Deference Done Better." *Philosophical Perspectives* 35 (1): 99–150. https://doi.org/10.1111/phpe.12156.

Gallow, J. Dmitri. 2021. "Updating for Externalists." *Noûs* 55 (3): 487–516. https://doi.org/10.1111/nous.12307.

Geanakoplos, John. (1989) 2021. "Game Theory Without Partitions, and Applications to Speculation and Consensus." *The B.E. Journal of Theoretical Economics* 21 (2): 361–94. https://doi.org/https://doi.org/10.1515/bejte-2019-0010.

Horowitz, Sophie. 2014. "Immoderately Rational." *Philosohical Studies* 167 (1): 41–56. https://doi.org/10.1007/s11098-013-0231-6.

Lasonen-Aarnio, Maria. 2020. "Enkrasia or Evidentialism? Learning to Love Mismatch." *Philosophical Studies* 177 (3): 597–632. https://doi.org/10.1007/s11098-018-1196-2.

Luzon, Bar. Forthcoming. "Epistemic Akrasia and Treacherous Propositions." *Philosophical Quarterly*, Forthcoming.

Martin, Ben, and Ole Thomassen Hjortland. 2024. "Anti-Exceptionalism about Logic (Part i): From Naturalism to Anti-Exceptionalism." *Philosophy Compass* 19 (8): e13014. https://doi.org/https://doi.org/10.1111/phc3.13014.

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

———. 2011. "Improbable Knowing." In *Evidentialism and Its Discontents*, edited by Trent Dougherty, 147–64. Oxford: Oxford University Press.

———. 2014. "Very Improbable Knowing." *Erkenntnis* 79 (5): 971–99. https://doi. org/10.1007/s10670-013-9590-9.