Akrasia and Traitors

Anonymous

2025-04-10

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.

In "Epistemic Akrasia and Treacherous Propositions", Bar Luzon (Forthcoming) argues that akratic doxastic states[1] are irrational because they violate a principle called **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.

Many arguments for the irrationality of akrasia simply appeal to an intuition that akratic states are irrational, or in some way weird, so it's very useful to see a theoretical

[1] For current purposes, an akratic doxastic state is a pair of beliefs in p and *It is irrational for me to believe p*, or some other claim about the belief in p failing to meet a similar standard

argument for the anti-akrasia view put forward. This note responds to those arguments on behalf of the kind of pro-akrasia, or at least anti-ant-akrasia, position put forward by philosophers such as Maria Lasonen-Aarnio (2020).

Luzon's principle AT has two important clarifications that should be on the table from the start. First, the quantifier over statuses has a very restricted range; it just covers "epistemic justification, epistemic rationality, evidential support and epistemic permissibility" (Luzon Forthcoming, 1). Second, the conditionals in it are material conditionals.

The second clarification makes it clear why the first is needed. If I know that my belief in *p* is not *E*, **AT** implies that I ought not believe *p*. This can't be right if E ranges over all positive epistemic statues, and not just these four.

Let *E* be a very strong epistemic status, like Cartesian certainty. Let *p* be the proposition that the cup of coffee I left in my office an hour ago has gone cold. I can rationally believe *p*, and hence that I should make more coffee, while knowing that Cartesian certainty in *p* is uncalled for. After all, cooling is a somewhat random process, and there is some miniscule chance that the molecules have buzzed around in just the right way to keep the coffee warm. Or let E be sensitive knowledge, and let *p* be any of the counterexamples that Saul Kripke (2011) gives to the sensitivity theory of knowledge put forward by Robert Nozick (1981). In either case, I'll know that my belief is not *E*, and hence that it is *E* only if $\neg p$, so unrestricted **AT** implies, implausibly, I shouldn't believe it.

Now these aren't counterexamples to **AT** as stated, because it is restricted. But they do raise a problem for any argument for **AT**. The main argument is that if *E* violates

AT, then *E* is not a good guide to truth. This can't be right in general, because Cartesian certainty and sensitive knowledge are excellent guides to truth, and they don't satisfy **AT**.

In general, it isn't clear why a universal claim like **AT** is true of a status for it to be a guide. If the properties of *being rationally believed* and *being true* are sufficiently positively correlated, then each is a guide to the other in one perfectly good sense of 'guide'. That can happen even if there are clear cases of one property obtaining without the other.

There is a stronger sense of 'guide' that Luzon might have in mind. (The discussion on page 7 comparing epistemic status to testifiers suggests this reading.) Say that feature *F* is a guide to truth just in case the inference from *Belief in p is F* to *p* is a reasonable inference, and a reasonable way to form for the first time a belief in *p*. In this sense, it won't matter if there are counterexamples to **AT** that turn on one knowing that the belief is not *F*, because they won't raise problems for this inference.

Again, we might quibble here over whether a good guide has to be an infallible guide; this inference might be reasonable even if in rare cases it leads from truth to falsity. But there is a deeper issue that is important here. A common pro-akrasia position, starting with Nomy Arpaly (2003), denies that this kind from normative status to a first-order belief is reasonable.

When all goes well, one's beliefs are supported by evidence, and they are rational because they are supported by evidence. But that does not mean that one infers that one's evidence supports *p*, and hence *p*. Rather, one simply infers *p* from that evidence. Perhaps one goes on to infer that one's evidence supports *p*, or one does not; it matters little.

There is an important parallel between ethics and epistemology here. In the moral case, it is often a vice and not a virtue to be motivated by normative considerations. A good friend is motivated by their friend's interests, not by the fact that doing something for their friend would be a manifestation of the virtue of friendship.

The anti-akrasia position follows naturally from a position where everything I said in the last apragraph is mistaken, and good inferences do go via normative considerations. A nice example of such a view is the approach to inductive reasoning defended by John Bigelow and Robert Pargetter (1997). They say that a typical instance of inductive reasoning goes like this.

1. I've seen lots of penguins, in lots of situations, and none of them have been able to fly.
2. That's all my evidence.
3. Therefore, it's rational for me to believe the next penguin I see won't be able to fly.
4. Therefore, the next penguin I see won't be able to fly.

They say that the inference from 1 and 2 to 3 is valid[2], and the inference from 3 to 4 has the nice feature that whenever the premise is true, the conclusion is rational. If this is how inductive reasoning works, then failures of **AT** will make the move from 3 to 4 somewhat questionable. So this kind of general reflection on the nature of induction could explain why **AT** is important for rationality to be a guide to truth in the sense that it is the thing that comes between 1/2 and 4 in this chain of reasoning.

[2] More carefully, the implication corresponding to this inference is valid.

But the pro-akrasia side has a simple reply to all this. Step 3 isn't a step that good reasoners must, or typically do, go through. The inference from 1 and 2 to 4 is at least as good, arguably better, than the one that goes via 3. If one thinks that the direct inference from 1 and 2 to 4 is good, then one doesn't need rational belief to satisfy anything like **AT**. It would be bad if rational belief came apart from truth too often, but it need not satisfy anything nearly as strong as the universal claim in **AT**.

This can feel like an impasse. There are two natural combinations of views.

Anti-Akrasia

- Akratic doxastic states are irrational.
- Principles like **AT** have to hold universally for rationality to be properly connected to truth.
- Good reasoning can, and perhaps should, go through normative steps like 3.

Pro-Akrasia

- Akratic doxastic states can be rational.
- Principles like **AT** should hold for most propositions, but they can fail occasionally without rationality coming too far apart from truth.
- Good reasoning need not, and perhaps should not, go through normative steps like 3.

Both pictures strike me as internally coherent, and in both cases different parts of the picture can be well motivated by looking at its other parts. But we shouldn't hold out too much hope for a conclusive refutation of either picture.

I prefer the second picture, largely because of arguments like Arpaly's against the idea that morally good actions are based in beliefs about the morality of that very action.[3] By analogy, it seems we should be able to reason directly from beliefs about observed penguins to beliefs about unobserved penguins. If that's right, we don't need rationality to be a guide in the sense that we are directly guided by it in individual bits of reasoning, and the reason to accept **AT** doesn't work.

Every step of the reasoning in that last paragraph seems far from watertight, and it's easy to see the appeal of the anti-akrasia picture. As I noted at the start, one of the great virtues of Luzan's paper is that it makes clear what's motivating the anti-akrasia picture, and how it holds together.

Still, it would be disappointing to end with just an impasse. So I'll end with a new reason to prefer the pro-akrasia picture. As Luzon notes, one prominent recent motivation for the pro-akrasia picture has been that anti-akrasia principles fail in formal models like Timothy Williamson's example of the unmarked clock (Williamson 2014). That model has a couple of contentious features, and recent supporters of the anti-akrasia picture have rejected it because of those features. I'll present a model without the features that are thought to its problematic features.

[3] The analogy with ethics might cut both ways. If one was convinced by Zoë Johnson King (2020) that Arpaly's view is wrong in ethics, that would be a reason to prefer the anti-akrasia side in epistemology.

The formal model I'll use here is similar to the kind of model developed by Kevin Dorst et al. (2021). Start with a set W of worlds, an a priori probability function π defined over W, and an epistemic accessibility relation R. Intuitively, xRy means that if x is actual, y is epistemically possible. That is, a person in x might, for all they know, be in y. Our hero will find themselves in some world $w \in W$, and learn the proposition $\{w\colon xRw\}$. They update on this, so their posterior probability is $\Pr(p) = \pi(p \mid \{w\colon xRw\})$. We'll work with a simple probabilistic model of evidential support. A proposition p is evidentially supported at w iff its probability, conditional on the evidence available at w is greater than a threshold t. To make life easier, I'll set t at 8/9, but it could be any value less than 1 and the following analysis will go through.

Whether this kind of model supports the pro-akrasia or the anti-akrasia picture depends in large part on what restrictions we put on R. If we say R must be an equivalence relation, so the epistemic logic we're using is S5, then we get a raft of anti-akrasia results.[4] But that's a rather implausible epistemic logic. As Lloyd Humberstone (2016, 380–402) argues, in that epistemic logic we can't distinguish knowledge from belief. We need a weaker epistemic logic. If we say that R need only be reflexive, so the epistemic logic is KT, then as Williamson showed (in Williamson (2014) and works cited therein) that we get very pro-akrasia results.[5]

In response to Williamson, many authors have argued for stronger constraints on R

[4] A lot of these results can be found in, or easily derived from, work by David Blackwell (1953).

[5] This is a slightly misleading way to describe Williamson's own views, which avoid some akratic results by the use of knowledge norms. It's better to describe Williamson as a level-splitter than a pro-akrasia person; but in terms of the two big pictures I described earlier, level-splitting is close enough to the pro-akrasia picture.

that rule out the models he uses. In Dorst et al. (2021), they get some anti-akrasia results from adding two constraints. One is that R is transitive, so the epistemic logic becomes at least as strong as S4. The second constraint is that R is 'nested', an idea that goes back to John Geanakoplos ([1989] 2021). I'll use the formulation of this idea in Williamson (2019); he calls the relevant principle **quasi-nestedness**.

Quasi-Nestedness

- $(aRb \wedge aRc \wedge bRd \wedge cRd) \rightarrow (bRc \vee cRb)$

- i.e., $\neg(aRb \wedge aRc \wedge bRd \wedge cRd \wedge \neg bRc \wedge \neg cRb)$

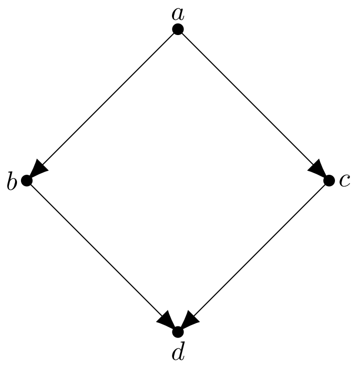

Figure 1: What a counterexample to quasi-nestedness looks like.

This principle rules out the situation depicted in Figure 1, where two worlds accessible from a can access a common point, but cannot access each other.

As Williamson notes, adding quasi-nestedness on its own as a frame condition does not change the epistemic logic. Any non-theorem of S4 is false on some quasi-nested frame. But quasi-nestedness is clearly implied by the characteristic condition on frames for S4.3, i.e., $(aRb \wedge aRc) \rightarrow (bRc \vee cRb)$. So if we do everything in S4.3, and in frames for it, we shouldn't beg any questions against the anti-akratic side. Such frames satisfy the two criteria, transitivity and quasi-nestedness, which have been put forward to avoid results friendly to the Pro-Akrasia picture.

So consider a model where W is the set of reals in [0, 1], with $w_@$ rigidly denoting the actual world, the prior probability function is the flat distribution over measurable subsets of [0, 1], and R is such that xRy iff $x \leqslant y$. So at x, the hero knows that the true value is at least x, and they know that they know this, but they don't know that that's the strongest thing they know. So in some sense they don't know what their evidence is; for all they know they may have more evidence than they in fact do. This kind of ignorance is unavoidable once we agree that S5 is not the correct epistemic logic.

Let p be that $\{w_@ \in (0.1,0.9)\}$; note that this excludes the end-points of the interval. If $w_@ \geqslant 0.9$, hero knows for certain that p is false. If $w_@ \leqslant 0.1$, then the posterior probability of p is $(0.2 - w_@) / (1 - w_@)$. And if p is true, the posterior probability of p is $(0.9 - w_@) / (1 - w_@)$. Note that the maximal value for this posterior probability is 8/9, and it gets to that value at one point only: when $w_@ = 0.1$. That's to say, p is evidentially supported iff it is false. Since hero can be assumed to know the model, hero can know that p is evidentially supported iff it is false. So **AT** fails in this model.

Now this is an artificial model. And **AT** fails at quite literally an edge-case. But it is important for the anti-akrasia argument that **AT** holds everywhere. Even the pro-akrasia theorist thinks that **AT** typically holds. (Indeed, that fact does some work in Lasonen-Aarnio's broader theory.) So if **AT** fails in formal models like this one, we should be sceptical of it.

So the pro-akrasia side has a few moves available here. **AT** can't be motivated by general constraints on epistemic statuses, because it doesn't hold for very strong statues like certainty. If one thinks that rational belief formation requires going through a step like *This belief is rational*, then **AT** is very plausible, but pro-akratics have long rejected steps like that. And **AT** has counterexamples even in the kinds of models that have usually been taken to be friendly to the anti-akratic view. Here, I think, there is plenty of interesting work to be done. If there is a good motivation for an epistemic logic between S4.3 and S5 (and Humberstone is very good on what some of the possibilities are here), that could be a new way to motivate anti-akratic epistemology.

References

Arpaly, Nomy. 2003. *Unprincipled Virtue*. Oxford: Oxford University Press.

Bigelow, John, and Robert Pargetter. 1997. "The Validation of Induction." *Australasian Journal of Philosophy* 75 (1): 62–76. https://doi.org/10.1080/00048409712347671.

Blackwell, David. 1951. "Comparison of Experiments." *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability* 2 (1): 93–102.

———. 1953. "Equivalent Comparisons of Experiments." *The Annals of Mathematical*

Statistics 24 (2): 265–72.

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.

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.

Humberstone, Lloyd. 2016. *Philsophical Applications of Modal Logic*. Milton Keynes: College Publications.

Johnson King, Zoë. 2020. “Praiseworthy Motivations.” *Noûs* 54 (2): 408–30. https://doi.org/10.1111/nous.12276.

Kripke, Saul. 2011. “Nozick on Knowledge.” In *Philosophical Troubles: Collected Papers, Volume 1*, 161–224. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199730155.003.0007.

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.

Nozick, Robert. 1981. *Philosophical Explorations*. Cambridge, MA: Harvard University Press.

Williamson, Timothy. 2014. “Very Improbable Knowing.” *Erkenntnis* 79 (5): 971–99. https://doi.org/10.1007/s10670-013-9590-9.

———. 2019. “Evidence of Evidence in Epistemic Logic.” In *Higher-Order Evidence: New Essays*, edited by Mattias Skipper and Asbjørn Steglich-Petersen, 265–97. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780198829775.003.0013.