Are You a Sim?

Brian Weatherson

Abstract

Nick Bostrom (2003) argues that if we accept some plausible assumptions about how the future will unfold, we should believe we are probably not humans. The argument appeals crucially to an indifference principle whose precise content is a little unclear. I set out five possible interpretations of the principle, none of which can be used to support Bostrom's argument. On the first two interpretations the principle is false, on the next two it does not entail the conclusion, and on the fifth it only entails the conclusion given an auxiliary hypothesis that we have no reason to believe.

In Will Wright's delightful game *The Sims*, the player controls a neighbourhood full of people, affectionately called sims. The game has no scoring system, or winning conditions. It just allows players to create, and to some extent participate in, an interesting mini-world. Right now the sims have fairly primitive psychologies, but we can imagine this will be improved as the game evolves. The game is very popular now, and it seems plausible that it, and the inevitable imitators, will become even more popular as its psychological engine becomes more realistic. Since each human player creates a neighbourhood with many, many sims in it, in time the number of sims in the world will vastly outstrip the number of humans.

Let's assume that as the sims become more and more complex, they will eventually acquire conscious states much like yours or mine. I do not want to argue for or against this assumption, but it seems plausible enough for discussion purposes. I'll reserve the term Sim, with a capital S, for a sim that is conscious. By similar reasoning to the above, it seems in time the number of Sims in the world will far outstrip the number of humans, unless humanity either (a) stops existing, or (b) runs into unexpected barriers to computing power or (c) loses interest in these kinds of simulators. I think none of these is likely, so I think that over time the ratio of Sims to humans will far exceed 1:1.

Nick Bostrom (2003) argues that given all that, we should believe that we are probably Sims. Roughly, the argument is that we know that most agents with conscious states somewhat like ours are Sims. And we don't have any specific evidence that tells on whether we are a Sim or a human. So the credence we each assign to *I'm a Sim* should equal our best guess as to the percentage of human-like agents that are Sims, which is far above 1/2. As Glenn Reynolds put it, "Is it live, or is it Memorex? Statistically, it's probably Memorex. Er, and so are you, actually."[1] (Is it worrying that we used the assumption that we are human to generate this statistical argument?

† Penultimate draft only. Please cite published version if possible. Final version published in *Philosophical Quarterly* 53 (2003): 425-31.

[1] Link. Reynolds's comment wasn't directly about Bostrom, but it bore the ancestral of the relation *refers* to Bostrom's paper.

Not necessarily; if we are Sims then the Sims:humans ratio is probably even higher, so what we know is a lower bound on the proportion of human-like agents that are Sims.) Less roughly, the argument appeals crucially to the following principle:

(#) $Cr(Sim \mid f_{Sim} = x) = x$

Here Cr is a rational credence function. I will adopt David Lewis's theory of *de se* belief, and assume that the credence function is defined over properties, rather than propositions Lewis (1979). Whenever I use a term that normally stands for a proposition inside the scope of Cr, it stands for the property of being in a world where that proposition is true. So $f_{Sim} = x$ stands for the property of being in a world where $100x\%$ of the human-like agents are Sims.

As Bostrom notes, the main reason for believing (#) is that it is an instance of a plausible general principle, which I'll call (##).

(##) $\forall \Phi$: $Cr(\Phi \mid f_{\Phi} = x) = x$

Bostrom does not formulate this more general principle, but it is clear that he intends something like it to be behind his argument, for many of the defences of (#) involve substituting some other property in place of *Sim* in statements like (#). So I will focus here on whether anything like (##) is plausibly true, and whether it supports (#). There are many ways we could interpret (##), depending on whether we take Cr to be a rational agent's current credences, or in some sense the prior credences before they are affected by some particular evidence, and on whether we take the quantifier to be restricted or unrestricted. Five particular interpretations stand out as being worth considering. None of these, however, provides much reason to believe (#), at least on the reading Bostrom wants to give it. In that reading (#) the credence function represents the current credences of an agent much like you or me. If (#) isn't interpreted that way, it can't play the dialectical role Bostrom wants it to play. On two of the interpretations, (##) is false, on two others it may be true but clearly does not entail (#), and on the fifth it only entails (#) if we make an auxiliary assumption which is far from obviously true.

For ease of exposition, I will assume that Cr describes in some way the credences at some time of a particular rational human-like agent, Rat, who is much like you or me, except that she is perfectly rational.

1 First Interpretation

Cr in (##) measures Rat's current credences, and the quantifier in (##) is unrestricted. On this interpretation, (##) is clearly false, as Bostrom notes. Rat may well know that the proportion of human-like agents that are like spaghetti westerns is rather low, while rationally being quite confident that she likes spaghetti westerns. For any property Φ where Rat has some particular information about whether he is one of the Φs or not, that information, and not general facts about the proportion of human-like agents that are Φ, can (indeed should) guide Rat's credences. So those substitution instances of (##) are false.

2 Second Interpretation

Just like the first interpretation, except that we restrict the quantifier range so that it only ranges over properties such that Rat does not know whether she possesses them. This interpretation seems to be hinted at by Bostrom when he says, "the bland indifference principle expressed by (#) prescribes indifference only between hypotheses about which observer you are, when you have no information about which of these observers you are." Even given this restriction, (##) is still false, as the following example shows.

Assume that Rat knows that $f_{Sim} > 0.9$, which Bostrom clearly takes to be consistent with rationality. And assume also that Rat, being a normal human-like agent, knows some fairly specific, and fairly distinctive facts about her conscious life. If Rat is anything like you or me, she will have experiences that he can be fairly sure are unique to her. Last night, for instance, while Rat was listening to Go-Betweens bootlegs, watching baseball, drinking beer, rocking in his rocking chair and thinking about Bostrom's simulation argument, she stubbed her toe in a moderately, but not excessively, painful way. Few people will have done all these things at once, and none in quite that way. Let C be the property of ever having had an experience almost just like that. Rat knows he is a C. She is very confident, though not certain, that she is the only human-like C. Let a suman be the property of being C and human, or not-C and a Sim. For much of the paper we're going to be concerned with the following two properties.

x is a **suman** $=_{\text{df}}$ x is a human C or a Sim who is not a C.

x is a **him** $=_{\text{df}}$ x is a Sim C or a human who is not a C.

We are following Bostrom in assuming that Rat does not know whether she is a Sim so she does not know whether she is a suman. But given that almost no one is C, it follows that $f_{suman} \approx f_{Sim}$. Hence $f_{suman} > 0.85$, for if it is less than f_{Sim}, it is not much less. But if $Cr(\text{a suman}) > 0.85$, and $Cr(Sim) > 0.9$, and Rat is coherent, it follows that $Cr(C) < 0.25$. But we assumed that Rat knew that she was a C, and however knowledge and credence are to be connected, it is inconceivable that one could know something while one's credence in it is less than 1/4. Hence it must be false that $Cr(C) < 1/4$, but we inferred that from given facts about the story and (##), as interpreted here. Hence (##), as interpreted here, is false.

3 Third Interpretation

One natural response ot the previous objection is that there shoul dbe some way of restricting (##) so that it does not apply to properties like being a suman. Intuitively, the response is that even though Rat doesn't know whether she is a suman, she knows something that is relevant to whether she is a suman, namely that she is a C. The problem with this response is that any formal restriction on (##) that implements this intuition ends up giving us a version so weak that it doesn't entail (#).

The idea is that what went wrong in the previous case is that even though Rat does not know whether she is a suman, she knows something relevant to this. In

particular, she knows that if she is a suman, she is one of the sumans that is human, rather than one of the ones that is a Sim. Our third interpretation avoids the difficulties this raises by restricting the quantifier in (##) even further. Say that a property Φ is in the domain of the quantifier iff (a) Rat does not know whether she is Φ, and (b) there is no more specific property Φ' such that Rat knows that if she is Φ, then she is Φ'.[2] This will rule out the applicability of (##) to properties like a suman. Unfortunately, it will also rule out the applicability of (##) to properties like *being a Sim*. For Rat knows that if she is a Sim, then she is a Sim that is also a *C*. So now (##) doesn't entail (#).

This kind of problem will arise for any attempt to put a purely formal restriction on (##). The problem is that, as Goodman noted in a quite different context (Goodman, 1955), there is no formal distinction between the 'normal' properties, being a human and being a sim, and the 'deviant' properties, being a suman and being a him. The following four biconditionals are all conceptual truths, and hence must all receive credence 1.

(1) (a) x is a suman iff x is a human C or a Sim who is not a C.
(b) x is a him iff x is a Sim C or a human who is not a C.

(2) (a) x is a human iff x is a suman C or a him who is not a C.
(b) x is a Sim iff x is a him C or a suman who is not a C.

If the obvious truth of (1a) implies that Rat cannot apply (##) to the property o being a suman once she knows that she is a C, for (1a) makes that evidence look clrarly relevant to the issue of whether she is suman, then similar reasoning suggests that the obvious truth of (2a) implies that Rat cannot apply (##) to the properties of being a human once she knows that she is a C, for (2a) makes that evidence look clearly relevant to the issue of whether she is human. The point is that a restriction on (##) that is to deliver (#) must fine some epistemologically salient distinction between the property of being human and the property of being suman if it is to rule out one application of (##) without ruling out the other, and if we only consider formal constraints, we won't find such a restriction. Our final attempt to justify (#) from something like (##) attempts to avoid this problem by appealing directly to the nature of Rat's evidence.

4 Fourth Interpretation

The problems with the three interpretations of (##) so far have been that they applied *after* Rat found out something distinctive about herself, that she was a *C*. Perhaps (##) is really a constraint on *prior* credence functions. *A priori*, Rat's credences should be governed by an unrestricted version of (##). We then have the following argument for (#). (As noted above, (#) is a constraint on current credences, so it is not immediately entailed by a constraint on prior credences such as (##) under its current interpretation.)

[2] I think it is this interpretation of (##) that Adam Elga implicitly appeals to in his solution to the Sleeping Beauty problem Elga (2000).

P1 *A priori*, Rat's conditional credence in her being a Sim given that f_{Sim} is x is x.
P2 All of Rat's evidence is probabilistically independent of the property of being a Sim.
C Rat's current conditional credence in her being a Sim given that f_{Sim} is x is x.

This interpretation may be reasonably faithful to what Bostrom had in mind. The argument just sketched looks similar enough to what he hints at in the following quote: "More generally, if we knew that a fraction x of all observers with human-type experiences live in simulations, and we don't have any information that indicate that our own particular experiences are any more or less likely than other human-type experiences to have been implemented *in vivo* rather than *in machina*, then our credence that we are in a simulation should equal x." So it's not unreasonable to conclude that he is committed to P2, and intends it to be used in the argument that you should give high credence to being a Sim.[3] Further, this version of (##), where it is restricted to prior credences, does not look unreasonable. So if P2 is true, an argument for (#) might just succeed. So the issue now is just whether P2 is true.

Why might we reject P2? Any of the following three reasons might do. First, Rat's evidence might be constituted by more than her conscious phenomenal states. This reply has an externalist and an internalist version. On the externalist version, Rat's perceptual evidence is constituted in part by the objects she is perceiving. Just as seeing a dagger and hallucinating a dagger provide different evidence, so does seeing a dagger and sim-seeing a sim-dagger. For reasns Williamson notes, a Sim may not know that she has different evidence to someone seeing a dagger when she sim-sees a sim-dagger, but that does not imply that she does not have different evidence unless one also assumes, implausibly, that agents know exactly what their evidence is Williamson (2000). On the internalist version, our evidence is constituted by our sensory irritations, just as Quine said it is (Quine, 1973). If Rat's evidence includes the fact that her eyes are being irritated thus-and-so, his credence conditional on that that she is human should be 1, for if she were a Sim she could not have this evidence because she would not have eyes. She may, depending on the kind of Sim she is, have sim-eyes, but sim-eyes are not eyes. So Bostrom needs an argument that evidence supervenes on conscious experiences, and he doesn't clearly have one. This is not to say that no such argument could exist. For example, Laurence BonJour provides some intriguing grounds for thinking that our fundamental evidence does consist in certain kinds of conscious states, namely occurrent beliefs (BonJour, 1999), but we're a long

[3] Jamie Dreier pointed out to me that what Bostrom says here is slightly more complicated than what I, hopefully charitably, attribute to him. A literal reading of Bostrom's passage suggests he intends the following principle.

$\forall e$: $Cr(e^* \mid Human) - Cr(e^* \mid Sim) = Cr(e \mid Human) - Cr(e \mid Sim)$ (B)

The quantifier here ranges over possible experiences e, e^* is the actual experience Rat has, and Cr is the credence function at the 'time' when Rat merely knows that he is human-like and f_{Sim} is greater than 0.9. I suggested a simpler assumption:

$Cr(Human \mid e^*) = Cr(Sim \mid e^*)$ (I)

Bostrom needs something a little stronger than (I) to get his desired conclusion, for he needs this to hold not just for Rat's experience e^*, but for your experience and mine as well. But we will not press that point. Given that point, though, (I) is all he needs. And presumably the reason he adopts (B) is because it looks like it entails (I). And indeed it does entail (I) given some fairly innocuous background assumptions.

way from knowing that the supervenience claims holds. And if the supervenience claim does not hold, then even if Sims and humans have the same kind of *experiences*, they may not have the same kind of *evidence*. And if that is true, it is open to us to hold that Rat's non-experiential evidence entails that she is not a Sim (as both Williamson and Quine suggest), so her evidence will not be independent of the question of whether she is a Sim.

Secondly, even if every one of Rat's experiences is probabilistically independent of the hypothesis that she is a Sim, that doesn't give us a sufficient reason to believe that her total evidence is so independent. Just because e_1 and e_2 are both probabilistically independent of H, the conjunction $e_1 \wedge e_2$ might not be independent of H. So possibly our reasons for accepting P2 involve a tacit scope confusion.[4]

Finally, we might wonder just why we'd even think that Rat's evidence is probabilistically independent of the hypothesis that she is human. To be sure, her evidence does not entail that she is human. But that cannot be enough to show that it is probabilistically independent. For the evidence also does not entail that she is suman. And if P2 is true, then the evidence must have quite a bit of bearing on whether she is suman. For Rat's prior credence in being suman is above 0.9 but apparently her posterior credence in it should be below 0.15. So the mere fact that the evidence does not entail that she is human cannot show that it is probabilistically independent of her being human, for the same reasoning would show it is probabilistically independent of his being suman.

More generally, we still need a distinction here between the property of being human and the property of being suman that shows why ordinary evidence should be independent of the first property but not the second. One might think the distinction can reside in the fact that *being human* is a natural property, while *being suman* is gruesome. The lesson of Goodman's riddle of induction is that we have to give a privileged position in our epistemic framework to natural properties like *being human*, and this explains the distinction. This response gets the status of privileged and gruesome properties back-to-front. The real lesson of Goodman's riddle is that credences in hypotheses involving natural properties should be distinctively *sensitive* to new evidence. Our evidence should make us quite confident that all emeralds are green, while giving us little reason to think that all emeralds are grue. What P2 says is that a rather natural hypothesis, that Rat is human, is *insensitive* to all the evidence Rat has, while a rather gruesome hypothesis, that Rat is suman, is *sensitive* to this evidence. The riddle of induction gives us no reason to believe that should happen.

It seems, though this is a little speculative, that the only reason for accepting P2 involves a simple fallacy. It is true that we have no reason to think that some evidence, say C, is more or less likely given that Rat is human rather than a Sim. But from this we should *not* conclude that we *have* a reason to think it is not more or less likely given that Rat is human rather than a Sim, which is what P2 requires. Indeed, drawing this kind of conclusion will quickly lead to a contradiction, for we can use the same 'reasoning' to conclude that we have a reason to think her evidence is not more or less likely given that Rat is a suman rather than a him.

[4] Thanks to Jamie Dreier for reminding me of this point.

5 Conclusion

Nothing I have said here implies that Rat should have a high credence in her being human. But it does make one argument that she should not have a high credence in this look rather tenuous. Further, it is quite plausible that if there is no good reason not to give high credence to a hypothesis, then it is rationally permissible to give it such a high credence. It may not be rationally mandatory to give it such a high credence, but it is permissible. If Rat is very confident that she is human, even while knowing that most human-like beings are Sims, she has not violated any norms of reasoning, and hence is not thereby irrational. In that respect she is a bit like you and me.

References

BonJour, Laurence. 1999. "Foundationalism and the External World." *Philosophical Perspectives* 13:229–249.

Bostrom, Nick. 2003. "Are You Living in a Computer Simulation." *Philosophical Quarterly* 53:243–255.

Elga, Adam. 2000. "Self-Locating Belief and the Sleeping Beauty Problem." *Analysis* 60:143–147.

Goodman, Nelson. 1955. *Fact, Fiction and Forecast,*. Cambridge: Harvard University Press.

Lewis, David. 1979. "Attitudes *De Dicto* and *De Se*." *Philosophical Review* 88:513–543. Reprinted in *Philosophical Papers*, Volume I, pp. 133-156.

Quine, W. V. O. 1973. *The Roots of Reference*. La Salle: Open Court.

Williamson, Timothy. 2000. "Scepticism and Evidence." *Philosophy and Phenomenological Research* 60:613–628.