Nick Bostrom 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 content is unclear. I set out four 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 third it does not entail the conclusion; and on the fourth it only entails the conclusion given an auxiliary hypothesis which 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. At present 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.

I shall 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 shall reserve the term ‘Sim’, with a capital S, for a sim who is conscious. By reasoning similar to the above, it seems that 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 argues that given all that, we should believe that we are probably Sims.1 Roughly, the argument is that we know that most agents with conscious states somewhat like ours are Sims. And we have no specific evidence that bears on whether we are Sims or humans. So the credence we each assign to I am a Sim should equal our best guess as to the percentage of human-like agents who are Sims, which is far above 50 per cent. As Glenn Reynolds put it, ‘Is it live, or is it


© The Editors of The Philosophical Quarterly, 2003. Published by Blackwell Publishing, 9600 Garsington Road, Oxford 0X4 2DQ, UK, and 350 Main Street, Malden, MA 02148, USA.
Memorex? Statistically, it’s probably Memorex. Er, and so are you, actually.2 Is it worrying that I have used the assumption that we are human to generate this statistical argument? Not necessarily; if we are Sims, then the ratio of Sims to humans is probably even higher, so what we know is a lower bound on the proportion of human-like agents who are Sims.

Less roughly, the argument appeals crucially to the following principle:

\[ C_r(\text{Sim} \mid \text{Sim} = x) = x. \]

Here \( C_r \) is a rational credence function. I shall adopt David Lewis’ theory of de se belief, and assume that the credence function is defined over properties, rather than propositions.3 Whenever I use a term that normally stands for a proposition inside the scope of \( C_r \), it stands for the property of being in a world where that proposition is true. So ‘\( \text{Sim} = x \)’ stands for the property of being in a world where 100 per cent of the human-like agents are Sims.

As Bostrom notes, the main reason for believing (1) is that it is an instance of the following plausible general principle:

\[ \forall \Phi: C_r(\Phi \mid \text{Sim} = x) = x. \]

Bostrom does not formulate this more general principle, but it is clear that he intends something like it to lie behind his argument, for many of the defences of (1) involve substituting some other property in place of Sim in statements like (1). So I shall focus here on whether anything like (2) is plausibly true, and whether it supports (1). There are many ways in which we could interpret (2), depending on whether we take \( C_r \) to be a rational agent’s current credences, or in some sense the prior credences before they are affected by some particular evidence, and also on whether we take the quantifier to be restricted or unrestricted. Four interpretations stand out as being worth particular consideration. None of these, however, provides much reason to believe (1), at least on the reading Bostrom wants to give it. On that reading of (1) the credence function represents the current credences of an agent like you or me. If (1) is not interpreted in this way, it cannot play the dialectical role Bostrom wants it to play. On two of the interpretations, (2) is false; on another it may be true but clearly does not entail (1); and on the fourth it only entails (1) if we make an auxiliary assumption which is far from obviously true.

For ease of exposition, I shall assume that \( C_r \) 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 he is perfectly rational.

**First interpretation**

\( C_r \) in (2) measures Rat’s current credences, and the quantifier in (2) is unrestricted. On this interpretation, (2) is clearly false, as Bostrom notes. Rat may well know that the proportion of human-like agents who like spaghetti westerns is rather low, while rationally being quite confident that he likes spaghetti westerns. For any property \( \Phi \)

---

2 http://www.instapundit.com/archives/003465.php#003465. Reynolds’ comment was not directly about Bostrom, but it bore the ancestral of the relation refers to Bostrom’s paper.

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, should guide Rat’s credences. So those substitution instances of (2) are false.

Second interpretation

This is just like the first interpretation, except that we restrict the quantifier range so that it ranges only over properties such that Rat does not know whether he possesses them. This interpretation seems to be hinted at by Bostrom when he says “the bland indifference principle expressed by [(1)] 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, (2) is still false, as the following example shows.

Assume that Rat knows that $f_{\text{Sim}} > 0.9$, 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 his conscious life. If Rat is anything like you or me, he will have experiences that he can be fairly sure are unique to him. Last night, for instance, while Rat was listening to Go-Betweens bootlegs, watching baseball, drinking vodka, rocking in his rocking chair and thinking about Bostrom’s simulation argument, he stubbed his 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. He is very confident, though not certain, that he is the only human-like C. For much of the paper I shall be concerned with the following two properties:

\[
\begin{align*}
\text{x is a suman} & \equiv x \text{ is a human C or a Sim who is not a C} \\
\text{x is a him} & \equiv x \text{ is a Sim C or a human who is not a C.}
\end{align*}
\]

Following Bostrom, I shall assume that Rat does not know whether he is a Sim, so he does not know whether he is a suman. Given that almost no one is C, it follows that $f_{\text{suman}} = f_{\text{Sim}}$. Hence $f_{\text{suman}} > 0.85$, for if $f_{\text{suman}}$ is less than $f_{\text{Sim}}$ it is not much less. But if $C'(\text{SUMAN}) > 0.85$ and $C'(\text{SIM}) > 0.9$, and Rat is coherent, it follows that $C'(C) < 0.25$. But we assumed that Rat knows that he is 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 0.25. Hence it must be false that $C'(C) < 0.25$, but we inferred this from given facts about the story and (2) as interpreted here. Hence (2) as interpreted here is false.

Third interpretation

One natural response to the previous objection is that there should be some way of restricting (2) so that it does not apply to properties like being a suman. The thought is that even though Rat does not know whether he is a suman, he knows something which is relevant to whether he is a suman, namely that he is a C. The problem with this response is that any formal restriction on (2) that implements this thought ends up by giving us a version of (2) so weak that it does not entail (1).

© The Editors of The Philosophical Quarterly, 2003
The idea is that what is wrong in the previous case is that even though Rat does not know whether he is a suman, he knows something relevant to this. In particular, he knows that if he is a suman, he is one of the sumans who are human, rather than one of those who are Sims. My third interpretation avoids the difficulties this raises by restricting the quantifier in (2) even further. Suppose a property \( \Phi \) is in the domain of the quantifier iff (a) Rat does not know whether he is \( \Phi \), and (b) there is no more specific property \( \Phi' \) such that Rat knows that if he is \( \Phi \), then he is \( \Phi' \). This will rule out the applicability of (2) to properties like being a suman. Unfortunately it will also rule out the applicability of (2) to properties like being a Sim. For Rat knows that if he is a Sim, then he is a Sim who is a him, rather than a sim which is a suman. So (2), thus restricted, does not entail (i).

This kind of problem will arise for any attempt to put a purely formal restriction on (2). The problem is that as Goodman noted in a quite different context, 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 all must receive credence:

3a. \( x \) is a suman iff \( x \) is a human or a Sim who is not a C
3b. \( x \) is a him iff \( x \) is a Sim or a human who is not a C
4a. \( x \) is a human iff \( x \) is a suman or a him who is not a C
4b. \( x \) is a Sim iff \( x \) is a Sim or a human who is not a C.

If the obvious truth of (3a) implies that Rat cannot apply (2) to the property of being a suman once he knows that he is a C, for (3a) makes that evidence look clearly relevant to the issue of whether he is suman, then similar reasoning suggests that the obvious truth of (4a) implies that Rat cannot apply (2) to the property of being a human once he knows that he is a C, for (4a) makes that evidence look clearly relevant to the issue of whether he is human.

The point is that a restriction on (2) that is to deliver (i) must find some epistemologically salient distinction between the property of being a human and the property of being a suman, if it is to rule out one application of (2) without ruling out the other; and if we only consider formal constraints, we shall not find such a restriction.

My final attempt to justify (i) from something like (2) attempts to avoid this problem by appealing directly to the nature of Rat’s evidence.

Fourth interpretation

The problems with the three interpretations of (2) so far have been that they applied after Rat found out something distinctive about himself, that he was a C. Perhaps (2) is really a constraint on prior credence functions. A priori, Rat’s credences should be governed by an unrestricted version of (2). We then have the following argument for (i). (As noted above, (1) is a constraint on current credences, so it is not

\[ I \text{ I think it is this interpretation of (2) to which Adam Elga implicitly appeals in his solution to the Sleeping Beauty problem: A. Elga, ‘Self-Locating Belief and the Sleeping Beauty Problem’, Analysis, 60 (2000), pp. 145–7. } \]

\[ ^{1} \text{ N. Goodman, Fact, Fiction and Forecast (Harvard UP, 1955).} \]
immediately entailed by a constraint on prior credences such as (2) under its current interpretation.)

P1. *A priori*, Rat’s conditional credence in being a Sim, given that $f_{\text{Sim}}$ 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 being a Sim, given that $f_{\text{Sim}}$ 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 passage: ‘More generally, if we knew that a fraction $x$ of all observers with human-type experiences live in simulations, and we have no information to 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 is 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. Furthermore, this version of (2), where it is restricted to prior credences, does not look unreasonable. So if (P2) is true, an argument for (1) might just succeed. So the issue now is whether (P2) is true.

(Jamie Dreier pointed out to me that what Bostrom says here is slightly more complicated than what I attribute to him, I hope charitably. A literal reading of the passage suggests that he intends the following principle:

\[ \forall e: Cr(e^* | \text{human}) - Cr(e^* | \text{Sim}) = Cr(e | \text{human}) - Cr(e | \text{Sim}). \]

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 $f_{\text{Sim}}$ is greater than 0.9. I suggested a simpler assumption:

\[ Cr(\text{human} | e^*) = Cr(\text{Sim} | e^*). \]

Bostrom needs something a little stronger than (B2) 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 I shall not press this. Given the point, though, (B2) is all he needs. Presumably the reason why he adopts (B1) is because it looks as if it entails (B2). And indeed it does entail (B2), given some fairly innocuous background assumptions.

To return to whether (P2) is true, why might we reject it? Any of the following three reasons might do. First, Rat’s evidence might be constituted by more than his 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 he is perceiving. Just as seeing a dagger and hallucinating a dagger provide different evidence, so do seeing a dagger and sim-seeing a sim-dagger. For reasons which Timothy Williamson notes, a Sim may not know that he has different evidence from someone seeing a dagger when he sim-sees a sim-dagger; but that does not imply that he does not have different evidence, unless one also assumes, implausibly, that agents know exactly what their evidence is.⁶ On the internalist

---


© The Editors of *The Philosophical Quarterly*, 2003
version, our evidence is constituted by our sensory irritations, just as Quine said it is. If Rat’s evidence includes the fact that his eyes are being irritated thus and so, then his credence conditional on this that he is human should be 1, for if he were a Sim he could not have this evidence because he would not have eyes. He may, depending on the kind of Sim he is, have sim-eyes, but sim-eyes are not eyes. So Bostrom needs an argument that evidence supervenes on conscious experiences, and he does not 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, but we are a long way from knowing that the supervenience claim 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 he is not a Sim (as both Williamson and Quine suggest); so his evidence will not be independent of the question of whether he is a Sim.

Secondly, even if every one of Rat’s experiences is probabilistically independent of the hypothesis that he is a Sim, that does not give us a sufficient reason to believe that his total evidence is similarly independent. Even if $e_1$ and $e_2$ are both probabilistically independent of $h$, the conjunction $e_1 \land e_2$ might not be independent of $h$. So possibly our reasons for accepting (P2) involve a tacit scope confusion.

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

---

7 W.V.O. Quine, *The Roots of Reference* (La Salle: Open Court, 1953).
9 Thanks to Jamie Dreier for reminding me of this point.

grue. What \((P_2)\) 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 this should happen.

It seems, although this is speculative, that the only reason for accepting \((P_2)\) 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 \((P_2)\) 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 his evidence is not more or less likely given that Rat is a suman rather than a him.

**Conclusion**

Nothing I have said here implies that Rat should have a high credence in being human. But it does make one argument that he 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 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 he is human, even while knowing that most human-like beings are Sims, he has not violated any norms of reasoning, and hence is not thereby irrational. In that respect he is a bit like you and me.\(^{10}\)

*Brown University, Rhode Island*

---

\(^{10}\) Thanks to Jamie Dreier and referees for very helpful comments.