Bostrom's Simulation Argument

David Braddon-Mitchell

Jason Grossman

draft from 2006 — new version coming in 2012, we hope

Introduction

Nick Bostrom has recently presented us with an argument designed to raise our credence in the thought that we are all simulations running on an advanced civilization's computer. His premises are that there is or will be enough computing power to simulate many individuals and perhaps whole universes, that simulated mental systems (at least in complex simulated environments) are conscious beings, and on one reading that beings like us are likely to have the desire to form such simulations. He then argues that there are many more simulated individuals than 'root' individuals—individuals in the world that produces the simulations. Via a weak indifference principle that claims we are equally likely to be in any of these situations, he then argues that we are more likely to be simulations than members of the root community.

We have some doubts about the soundness of this argument. For the current purposes, we grant the premises.

A Bad Argument

Here is an argument that is poor. In the sense of merely logically possible, there are more possible worlds where things are not as they seem than worlds where they are as they seem: for there are many more ways that things can seem as they are, than both be as they are as well as seem as they are. There is a set of words that is consistent with what reason and evidence says about my experience, but there are many more worlds where solipsism is true, where me and one other conscious agent exists, where we are simulations made by people with property A, simulations made by people with property B, produced by malevolent demons and so on. Since

the skeptical worlds outnumber vastly the non-skeptical worlds, via an indifference principle we can argue that we are likely in a skeptical world.

This is a bad argument. Exactly why is controversial, but many might take it that it is a constraint on the right theory of evidence or probability that it shows this argument to be bad.

The strength of Bostrom's position is that it does not seem to be equivalent to this bad argument.

Bostrom's argument differs from the bad argument.

Bostrom's argument however does not seem to be trading in merely logically possible worlds. For Bostrom's premises about the actual world tell us that there really will be (or that it is very likely there will be) many actual simulated universes located in the future. These are not mere possibilities; they are (most likely) actual. So if it is equally likely that we are located in any of these or in the root universe—the realm of being that created these simulations—then we would agree, we are very likely living in one of those simulations.

But is it equally likely? First we must ask what it means to say that 'we' are in one universe or another (we use 'universe' here indifferently between a real universe and a merely simulated one). The universe that 'we' are in we take to be picked out indexically. The universe we are in—our world—is the one with whom we have the most direct causal and epistemic contact. It is the direct source of the evidence that various things (like computation) are nomologically possible, and of our evidence about the nature of agents' desires.

Thus the evidence that we have about what is nomologically possible and so forth is evidence about the nature of our world. This is the world that we know to (probably) contain many simulated worlds. Is it possible that our world is simulated? Yes. But is it possible that our world is one of the simulations in our world? No, for this is near enough a logical error.

Think of it this way: either our world is a simulation or it is not. If it is a simulation, then what we have evidence for is many sub-simulations. But we are not a sub-simulations of our world. That would be equivalent to saying that our world is not our world, or our world is identical to a proper part of our word.

If it is not a simulation, then what we have evidence for is many simulations. But if our

world is not a simulation, then it is not a simulation, and in any case cannot (again) be identical to a proper part of it.

So if when Bostrom is counting posthuman civilizations he is counting the ones we have direct reason to believe in: the ones that we will simulate, or will be simulated by those simulations and so on, then the argument gives us no reason to believe we are in any of those.

Bostrom's Argument and merely possible worlds

Is there a way of understanding Bostrom's argument so that it falls somewhere between the bad argument, concerned with merely possible worlds, and the argument above, concerned with indifference over which part of the actual world we are located in?

Perhaps what is needed is something like this: all merely possible worlds which have a segment which presents like our world does, have many such segments. For most possible worlds which have such segments, have a root segment, and many simulated segments. The evidence that we gather about how things are actually gives us reason to think that all the logical possibilities are structured like this.

Now, if most of the possible worlds are like this, then what are we to think of the location of 'our world' in logical space? We do know - from the previous argument - that it is not in a proper part of itself (ie in a simulation relative to *us*). But all these logical possibilities are equally likely, and in each of them there is a separate question - where we are located inside that possible world (ie which *cantered world* we are in). There are many answers to that question—corresponding to the number of root universes and the number of simulated universes in each possible world. If the number of simulated universes exceeds the number of root universes in enough of the worlds, then a principle of indifference over which logically possible world (consistent with evidence) we are in combined with a principle of indifference over which *centred* world we in are in will give us Bostrom's result.

But the principle of indifference over which logically possible world we are in (consistent with evidence) leads us directly to standard skeptical arguments. For there are many such worlds that don't have the Bostrom structure, but rather are solipsistic worlds, demon worlds

and so on. So the fact that there is a special subset of these worlds with the multi-simulation structure (henceforth the Bostrom Structure) should not change our attitude.

One last thought on Bostrom's behalf: his arguments tell us that there are (probably) actual simulations. They do not tell us that there are logically possible simulations, because we already knew that. But they do tell us that there are nomologically possible worlds with the Bostrom structure.

Is that enough to defend the argument? For it is a principle of scientific reason that we have reason to believe we are in a nomologically possible world. We do not know exactly which one, because we do not know all the facts. Consistent with knowing that our world is not in one of the simulations that are located within in, we might think that we could be in any location (consistent with evidence) in any of the nomologically possible worlds, many of which have the Bostrom structure. By a Principle of indifference over centered worlds, we could be in any location in any world, and in most of these location/world pairs we are simulations. So we are probably simulations.

This seems to us the best articulation. But still something is wrong. If we are a world contained by another world (a simulation, if you like) why should we suppose that the containing world has laws of nature like ours? We knew before Bostorm's argument that we *could* all be simulations (perhaps run on computers that use different laws of nature from the ones around here). Indeed very likely that the simulators would use simplified or different laws of nature. The premises of Bostrom's argument only tell us that *around here* the laws of nature permit (possibly further) simulation. That tells us about the likelihood of sub-simulations, but it is unclear whether it tells us about the chances of our world being itself a simulation, unless it was always a premise that *if our world is a simulation it is a simulation in a nomologically very simular 'upper level' world.* We do not see any reason to believe this premise.

This then boils down to three ways to count the 'civilizations' in Bostrom's arguments. We can count them by the ones we actually simulate. But we know we are none of those. We can count them by mere logical possibility, but then we are back with old-fashioned skepticism. Or we can count them by near enough nomological possibility. But what reason is there to restrict us to that? Which takes us back to old-fashioned skepticism.

Not quite: perhaps there is some strange bare identity between containing worlds and contained worlds. As far as we can tell this is nomologically impossible, because the containing world might need more computing power. But at least it is logically possible. But it's a bare skeptical possibility.

There may be more than one creation is a merely possible world. Indeed some cosmology says that there are multiple universes actually.