
Can you retroactively put yourself in a computer simulation?

By Paul Almond

3 December 2010

Website:

<http://www.paul-almond.com>

E-mail:

info@paul-almond.com

Nick Bostrom's *simulation argument* has suggested that we might be in a computer simulation. An argument is made that our actions are relevant here, and that some actions that we might perform could retroactively put us into a computer simulation. This does not mean that we could do things that cause us to enter a simulation in the future, such as mind uploading. Rather, the argument is that *we might be able to do things which result in us having been in a computer simulation all along*. Of particular relevance here is the idea of running simulations of the past. If our future seems to be one in which simulations of (what is to us) the present will be made, then this suggests that we may be in such a simulation already. Therefore, if we choose to run such simulations, or to make the running of them more likely in the future, we might consider ourselves to be choosing to have been in a computer simulation all along. The backward-causation aspect of this may seem strange, but such ideas are well-known in considerations of Newcomb's paradox. As well as affecting the probability that you are in a computer simulation, your actions might affect the type of simulation in which you are. Opposing groups may have differing views about the merits or otherwise of being in a simulation, or about the type of simulation that is ideal, and this suggests the possibility of *simulation warfare*, in which opposing groups take conventional action against each other or try to take control of reality away from each other by running simulations. Simulation warfare principles might be used by a "boxed" artificial intelligence (AI) in an attempt to threaten humans so that they will release it. In this article I am making no strong claim about the argument presented here, which is extremely speculative and which I expect to be controversial, but merely providing it for discussion. The argument, if valid, indicates existential threats to humanity.

Table of Contents

1 Introduction	3
2 The Scenario.....	6
2.1 A Description of the Scenario	6
2.2 Why Pressing “ACTIVATE” Might Retroactively Put You in a Simulation	7
2.3 Discussion.....	9
2.3.1 Measure and other complications.....	9
2.3.2 More General Situations.....	9
2.3.3 Existential Threats.....	12
2.3.4 Choosing the Kind of Simulation.....	12
2.3.5 Using Simulations to Escape the Implications of the Doomsday Argument ...	13
2.3.6 Running Trial Simulations	13
3 Simulation Warfare	15
3.1 The Idea of Simulation Warfare.....	15
3.2 Simulation Warfare as a Way for a Boxed Artificial Intelligence (AI) to Escape.....	15
3.3 Simulation Warfare Countermeasures	17
3.3.1 Countermeasure 1: Conventional Action	17
3.3.2 Countermeasure 2: Counter-Simulation.....	18
4 Conclusion	19
5 Acknowledgements.....	20
6 Bibliography	21

1 Introduction

Nick Bostrom's *simulation argument* (Bostrom, 2003) suggests that we may be in a computer simulation. Such a simulation might be made by post-21st Century humans or by non-humans.

The simulation argument says that one of the following is true.

1. Almost all civilizations end before they are capable of producing simulated realities.
2. Almost all civilizations that become capable of producing simulated realities choose not to do so.
3. We are almost certainly in a simulated reality.

The idea is that if (1) and (2) are not applicable, then there would be many simulated realities – and the number of simulations of our period in history in this civilization would be much higher than the number of “real” situations like ours, making it almost certain that we are in one of the simulations.

Bostrom has noted that if, in the future, we start making simulated realities ourselves, we would have strong evidence that (1) and (2) do not apply, meaning that (3) applies and we are almost certainly in a simulation ourselves. According to this argument, anyone playing around with simulations of civilizations needs to confront the truth that he/she is probably *in* a simulation.

This situation has an element of *apparent backward causation* that will be explicitly addressed in this article. If our civilization, after doing something – making simulations – is faced with a high probability of being in a computer simulation, does this not mean that before making these simulations we can *choose* whether to make them or not, and therefore choose whether to make it likely that we are in a simulation? This will sound irrational to some people. If we are in a simulation, *we have been in one all along* and that decision will seem to have already been made. Either we have been in a simulation all along or have not. Either way, it should already be too late for us to be able to affect our situation by what we do. The ability to choose whether we have been in a simulation all along would seem to demand a weird form of *backward causation*.

Of relevance here is Newcomb's paradox or Newcomb's problem (Nozick, 1969; Kiekebas, 1996). Newcomb's paradox presents a scenario in which our opportunity to win some money depends on what someone else, an excellent predictor of our actions, has predicted in the past. It therefore seems that we should act now so as to ensure that the predictor predicted the actions that we want him to have predicted – and there is the paradox: to do that we have to make decisions as if we can affect the past.

Can you retroactively put yourself in a computer simulation?

The idea of backward causation may not be as implausible as it may at first seem. If we view decision-making in normal, causal terms, it does not make sense, but another way of viewing decisions, evidential decision theory, would hold that it is a reasonable idea. In a previous article, *On Causation and Correlation – Part 1: Evidential decision theory is correct.*, I argued in support of evidential decision theory, and in support of the idea that we can “choose” to affect the world in ways in which we could never affect it in a standard causal way – and this includes backward causation (Almond, 2010). To avoid confusion with our existing concept of causality, I used the term “meta-causation” for this “extended” causation. In a second article, *On Causation and Correlation – Part 2: Implications of Evidential Decision Theory*, I discussed how such ideas may allow us to control our situation in strange ways, such as allowing us to control the probability that we have been in a computer simulation all along, or even to choose the kind of simulation in which we want to be (Almond, 2010).¹

This may seem a strange idea. If correct, it means that, in principle, we might decide that we want to be in a computer simulation, and then take action that makes it likely we have been in a computer simulation all along: we will have retroactively turned ourselves into computer simulated beings. This might have some advantages: we might want to choose the kind of simulation which we are in. However, there are also risks: if it is possible to put ourselves in a computer simulation deliberately, it follows that by performing certain actions without thinking we might accidentally put ourselves in a simulation, when in fact we do not want to be in one. Furthermore, even if you do not want to be in a simulation, and choose not to perform actions which would retroactively put you in one, this does not stop someone else from putting us in one without your consent. Someone with politics opposed to yours, or desires of power, may put us in a kind of simulation that makes the world as they want it. This raises the strange idea of *simulation warfare*, in which groups with different ideas about whether they want us to be in a simulation, or what kind of simulation it should be, act in opposition to each other – which could involve running simulations themselves or actual, real military action.

In this article I will discuss these ideas in more detail, giving a scenario with a statistical argument to show why we might be justified in thinking we can make decisions that affect the probability that we are in a simulation. I will also discuss the idea of simulation warfare. I will not be giving a detailed defence of the backward causation issues relating to all this: I have done that in the previous articles. Instead, I will be trying to show where ideas such as backward causation and other thinking related to Newcomb’s paradox can take us with regard to this issue.

¹ Alternatives to evidential decision theory have been proposed, such as “timeless decision theory”, which still have an evidential component. I am not taking a position on these, and the main ideas here would still apply with them anyway.

Can you retroactively put yourself in a computer simulation?

In this article I am making no strong claim about the argument presented here, which I expect to be controversial, but merely providing it for discussion.

2 The Scenario

2.1 A Description of the Scenario

You are in a white room, in front of an extremely powerful computer. The computer has two buttons. One button is labelled “ACTIVATE” and the other button is labelled “CANCEL”. You have a few minutes to decide which button to press – and to keep things simple, let us assume that if you press neither, the “CANCEL” option will be selected on your behalf anyway.

- If you press “ACTIVATE” the computer will start up and will start running a collection of programs.
- If you press “CANCEL”, or if you do not press a button, the computer will not start up: instead, all the programs on it will be erased.

You do not get to see the output of the computer. If you choose to activate it, you just know that it is computing away inside its case.

So, what do the programs on the computer do?

Each program on the computer simulates a reality in which there is a conscious agent, with some degree of similarity to yourself, in a scenario like the one being discussed here: each program simulates a person, similar to you to some degree, in a white room, in front of the computer, faced with the choice of whether to press the “ACTIVATE” or “CANCEL” buttons. Each such person will have memories of how he/she got into the room. You might assume that the computer will very quickly simulate an entire life, with some contrivances so that ultimately the simulated person ends up in the white room, or the computer may generate fake memories (though this in itself might demand some simulation). If we want to ensure that it is not too difficult to construct simulations in which everyone ends up in the white room, we can specify that everyone who finds him/herself in the white room does so after being made unconscious and abducted without warning, to wake up in the room, and we can require that you remember arriving in the room in the same way.

The simulated people do not need to be exactly like you. They might be extremely similar to you, being based on scans of your brain, or they may be less similar, being based on observation of your behaviour. They may be less similar still: they might just be people experiencing life in the same period of history as you, or they might even just be humans in general, or even conscious beings in general. You can consider the scenario with different degrees of similarity if you think it is important.

The simulated people, of course, do not exist unless the “ACTIVATE” button is pushed. Once a simulated person exists, if he/she chooses to push the “ACTIVATE” button, this does not mean that any simulations really start: the person is in a computer simulation,

and nothing really has to happen inside the simulated computer. It is only in the “real world” that pressing the “ACTIVATE” button really needs to start anything.

The question now is: Does pressing “ACTIVATE”, or choosing not to press it have any consequences? An argument can be made that the answer is “Yes”: that choosing to press “ACTIVATE” is equivalent to choosing to have been in a simulation all along.

2.2 Why Pressing “ACTIVATE” Might Retroactively Put You in a Simulation

Let us imagine some being, whom we will call *Omega*, in the real world, in front of the computer, in the scenario just described, being faced with the decision of whether to press “ACTIVATE”. You could be in one of two situations.

Situation 1: You are Omega or you are in a simulation launched by Omega.

The first situation to consider here is the obvious one in which Omega exists in the real world, in front of the computer, and you are Omega or one of the people in the simulations launched by Omega. If Omega presses “ACTIVATE”, then a number of simulated people, SIMULATION 1, SIMULATION 2, ...SIMULATION N, come into existence. If Omega does not press “ACTIVATE” then these simulated people do not come into existence.

Situation 2: You are in some other situation.

Another possible situation is that you are neither Omega or one of the people in the simulations launched by Omega: Omega may not even exist, and the situation that you are in has been caused some other entity, *Alpha*, and is not at all as it appears. For example you may be in a computer simulation made by Alpha, in which your “ACTIVATE” button does not do anything and there may not even be an “ACTIVATE” button in the real world – which may be vastly different to the world you are experiencing. If this is your situation, it will be hard to make any decisions about pressing the button as you have no idea what is going on. This situation is going to get little attention right now, besides being acknowledged as a possibility. (It is actually possible that, in a more complicated scenario, your situation could be caused by some other entity, *Alpha*, that you *do* know something about, just as though you know something about Omega, and this will be discussed later, in 3, when simulation warfare is considered. Right now, it is just a complication that is not needed.)

Firstly, let us assume that you are in Situation 1: you are actually Omega, the real-world person you seem to be, or you are in one of the simulations started by Omega pressing “ACTIVATE”.

Suppose you chose not to press the “ACTIVATE” button. If you are Omega, then Omega clearly did not press “ACTIVATE” and the computer did not start up and launch the

Can you retroactively put yourself in a computer simulation?

simulations, so the simulated people, SIMULATION 1, SIMULATION 2, ...SIMULATION N did not come into existence. On the other hand, you may not be Omega, but instead you may be in one of the simulations launched by Omega. If this is the case, Omega, in the real-world, obviously did press "ACTIVATE" or you would not exist. This leaves a situation in which you cannot be sure about your status: either you are Omega, and Omega did not choose to press the "ACTIVATE" button, or you are not Omega and Omega did choose to press the "ACTIVATE" button. This all clearly depends, to some extent on what Omega did – and you cannot be completely sure because you may not be Omega. The main point here is that it may be reasonable to think that you could be Omega, and that none of the simulations even got launched. Furthermore, you may think that choosing not to press "ACTIVATE" strengthens the case that people like you, or people in general will choose not to press "ACTIVATE", making it more likely that you are actually Omega.

Suppose, now, that you just pressed the "ACTIVATE" button. If you are Omega, then you obviously know that Omega pressed the "ACTIVATE" button. On the other hand, if you are not Omega, and are one of the simulated people, SIMULATION 1, SIMULATION 2, ...SIMULATION N, then you also know that Omega must have pressed "ACTIVATE," because otherwise you would not exist. Therefore, whatever your situation is, you know that Omega pressed "ACTIVATE". The fact that Omega pressed "ACTIVATE" means that all the simulations got launched and there is just one person, Omega, in a situation like yours, and a huge number of simulated people in situations like yours. It is almost certain that you are not Omega, and that you are in a simulation started when Omega pressed "ACTIVATE".

Let us consider things now from your point of view just *before* you decide.

If you choose not to press "ACTIVATE", you will know afterwards that you may or may not be Omega – depending on what Omega did. It may be hard to work out the probabilities, but there would seem to be a reasonable chance that you are Omega – that you are in the real world.

If you choose to press "ACTIVATE", you will know afterwards that Omega *definitely* pressed "ACTIVATE" (We are ignoring the possibility of Situation 2 for now) and that this makes it almost certain that you are in a simulation. Therefore, choosing to press "ACTIVATE" is equivalent to choosing to *make* it certain that Omega pressed "ACTIVATE" and almost certain that you are not Omega, but an inhabitant of one of the simulations launched by Omega's actions.

The implication of this is that if you press "ACTIVATE" you are choosing to have been in a computer simulation all along.

For completeness, let us now consider Scenario 2. In this scenario, you are neither Omega nor an inhabitant of one of the simulations launched by Omega. If this scenario applies, you are extremely misinformed about your situation, and you really have no

idea what is going on. You are certainly not in the “normal” existence in which you seem to be. In this scenario, there is no reason to think that pressing “ACTIVATE” or not pressing it will have any specific consequences. This scenario may introduce some doubt, but the unpredictability means it cannot favour one course of action over another. Scenario 2, then, does not change things very much. Furthermore, the possibility of Scenario 2 does nothing to reduce the chance that you are not Omega if you press “ACTIVATE”: it actually increases it by adding another way for you to be in some unusual situation, possibly being simulated in some other way, that means you are not Omega – except that it introduces the same chance for not pressing “ACTIVATE” as well.

Ultimately, even allowing for the possibility of Scenario 2, choosing to press “ACTIVATE” is choosing to make it more likely that you are in a simulation launched by Omega, and that you have been in such a simulation all along. In other words, you should press “ACTIVATE” if you want to be an inhabitant of a computer simulation.

2.3 Discussion

2.3.1 Measure and other complications

The above discussion assumes that all that matters, statistically, is the number of computer simulations. I have previously proposed, in other articles, that things are not this simple and that a *measure* view of observers needs to be taken (Almond, 2007) as part of an argument for a mathematical universe like Max Tegmark’s (Tegmark, 1998). Bostrom has also argued that some kind of measure view is needed, although he seems to stop short of saying that this implies a multiverse (Bostrom, 2005). This, however, would not invalidate the argument. Even if we take account of measure like this, the simulations run by the computer could be made to carry as much statistical weight as was needed, by running a sufficient number to allow for any measure issues, or by ensuring an appropriate substrate. It could, however, make it more difficult to put yourself into a simulation deliberately, and reduce the risk of it happening accidentally.

Another issue raised by some of my previous writing about measure is that a computer program may not actually need to be run to cause an observer to exist: some interpretation of it could imply an observer even while it just waits in storage. This might suggest that the computer in the scenario is putting you at risk of not being Omega even without a button press by anyone, just from the software installed on it. This is not a serious issue, however: it would be easy to arrange the computer so that observers were implied with much higher measure after the button was pressed.

2.3.2 More General Situations

The scenario discussed is quite specific and contrived, but it is intended to give an idea of why we might need to consider our actions important, in more general situations, in terms of possibly retroactively placing us in a computer simulation.

Can you retroactively put yourself in a computer simulation?

While the scenario involved just you, a real situation might involve possible simulation of other people – or even everyone on the planet. As a civilization, we may choose to do things that retroactively put us in a computer simulation, or may inadvertently do them.

The scenario should seem relevant to the issue of whether or not we should actually construct simulated realities. It could be argued that if we ever launch a significant number of simulations of worlds like our own, we have, as a civilization, effectively just retroactively put ourselves in a simulation.

The argument that we might retroactively put ourselves in a simulation seems strongest for “ancestor” simulations: simulations of the past. If it seems that someone in our future is likely to start creating historical simulations of our period in history, the scenario will essentially be playing out, except with a largely irrelevant time difference. If you expect Omega to exist at some point in (what appears to be) your future, creating many simulations of your period in history, then you know that there is one real-world version of your period in history and many simulated versions, suggesting that you are probably in one of the simulated version. This suggests that we do not actually need the ability to run simulations right now to put ourselves in one retroactively. *All we need do is perform actions that make the running of ancestor simulations more likely in the future.* Such actions could be encouraging people to run them, developing the capability to construct simulated realities or just generally improving computer technology. How far should we go with this? An argument can be made here that, on the day Alan Turing sent his classic paper proposing the Turing machine to a journal, if he had thought about it, he should have realized that he had just retroactively increased the probability of us being in a simulation – and if he did not want to do that, he should not have submitted it. An argument like this could be taken to still greater extremes, of course: we might say the same about the person who made the first flint tool.

The argument that making it likely that ancestor simulations will be constructed could retroactively put you in a simulation was based on the similarity of this situation with the “Omega” scenario I gave earlier, but there is an alternative way of describing this that may make more sense to some people. If someone is constructing ancestor simulations, then those simulations, by definition, should resemble the past – but the past was one which ended up giving rise to ancestor simulations. Therefore, by doing things that are likely to give rise to ancestor simulations in the future, you are making your present just like the kind of past that would feature in an ancestor simulation – retroactively increasing the probability that you are in one. On the other hand, if you do things that make the running of ancestor simulations less likely in your future, you should be *retroactively* making it less likely that you are in one. It should be noted that we are not talking about ancestor simulations really having to run in your future, here, but merely about your present being the kind of present that would be expected to give rise to future ancestor simulations, but from your point of view, this would seem to be about the chance that ancestor simulations are going to be run in your future.

Can you retroactively put yourself in a computer simulation?

It could be argued that construction of ancestor simulations at some time is inevitable anyway, but in that case, you may be able to control your chances of being in one by doing things that are likely to delay the running of the first ancestor simulation or to bring it forward. Delaying the running of the first ancestor simulation may mean that the total number of them run in the future is reduced. A further point could be that anyone running ancestor simulations has all of history from which to choose, and we might assume that the greater the amount of history that is available, the lower the chances of your period of history actually being one of the ones chosen. On the other hand, the period in history immediately before the present of whoever is doing the simulating might receive particular attention, and there may be a greater concentration of simulations here. Therefore, this may be a further argument for thinking that, if you do not want to put yourself into a simulation retroactively, you should try to delay the time at which it seems anyone will run ancestor simulations in your future for as long as possible.

It should be noted that this argument does not mean that you are safe from being simulated by future humans if you ensure that ancestor simulations do not occur in your future. Ancestor simulations would not reproduce every detail of the past with perfect accuracy, and much of the simulation's behaviour may be unknown to the creators until they run it. We should expect an ancestor simulation to be correct at least in general features, however, and it would be reasonable to think that one of these features would be a statistical tendency to causing ancestor simulations to be started in the future which would be apparent in individual simulations. As an example, suppose that Alan Turing has just completed his paper on Turing machines and is about to send it by post to a journal. At the last moment, just before he is about to send it, he realizes that he may be about to increase the probability that his world, or just he, is in an ancestor simulation and stops. Stopping at this stage, however, may not make much difference. By *almost* sending the paper to the journal, Turing has shown that his world is the kind of world in which people write papers like that and this may be enough to make the world resemble the kind of world which would be simulated to a significant degree, meaning that he may have already retroactively caused his world to be simulated. We might also consider the idea of a "celebrity simulation" here as well. If people who make great accomplishments are more likely to be simulated, then Turing might reasonably think he is making himself more likely to be the subject of a simulation by sending the paper, possibly retroactively putting himself into a simulation, but *almost* sending it might have almost the same effect, by showing that he is the kind of person who thinks of things like this: it might be possible, for example, that a real Alan Turing thinks of the Turing machine concept, and then, centuries later, there are numerous historical simulations based on his personality and intellectual abilities in which he may do slightly different things.

The argument could be made more general. It could be argued that if we create simulated realities, or make it likely that they will be created in the future, even if those simulations do not resemble our world to a great degree, we would have made it more

Can you retroactively put yourself in a computer simulation?

likely that civilizations in general tend to create simulations, and from an evidential decision theory perspective, we would have made it more likely that someone or something else is simulating us: we may have retroactively put ourselves in a computer simulation.

2.3.3 Existential Threats

The argument, if valid, indicates existential threats to humanity. The most obvious of these is that our civilization might retroactively put itself into a computer simulation that runs for only a short time into the future (by our perception of time) before being terminated – effectively ending the world for us. We might think that this could happen, for example, if our period of history is of particular interest to those doing the simulating, but later periods are not. Another possibility is that we might retroactively put ourselves into a simulation where extreme things happen – such as natural disasters or invasion by aliens – that pose an existential threat. A further possibility is that someone in our civilization could deliberately and retroactively put us into a simulation that poses an existential threat – possibly the ultimate in “taking other people with you”.

2.3.4 Choosing the Kind of Simulation

If it is possible to put yourself, retroactively, into a computer simulation – or at least to change the probability that you are in one, then it should also be possible to perform actions that retroactively affect the type of simulation that you might be in. This would be a strange form of “reality editing”, in which you either run simulated realities of worlds like yours in which you want various things to happen or not happen, or where you take steps to try to get the kinds of simulations you “want” run by your descendants. (Of course, when you think you are running these simulated realities, that may not really be happening, and when you try to get your descendants to run the kinds of simulations you want, it does not mean that you are really getting them to run them. Rather, you are trying to give your own world, in its current time, a certain appearance – the appearance that ancestor simulations are going to start soon – and one that makes it a member of a reference class of worlds with that appearance, most of which are simulated.)

You would need to be careful what you wished for. There would be no point in trying to get the simulations to be inconsistent with what you already know. Trying to make it so that natural disasters do not happen and have never happened would be counter-productive, because we already know they have happened. Likewise, trying to make it so that human lifespan has always been thousands of years would be counter-productive: we know it has not been. There is no reason, however, why you should not consider trying to stop natural disasters happening in the future, or adding a very long afterlife to your world.

Of course, some people may wish for things that you do not want, which relates to the subject of *simulation warfare*, which will be discussed later in the article, in 3.

2.3.5 Using Simulations to Escape the Implications of the Doomsday Argument

The Doomsday argument is a controversial argument suggesting that, if human civilization will survive for a long time into the future, we are now part of an extraordinarily small proportion of people living very early in history, and that as this is unlikely it is more plausible that our civilization will not survive long (Carter, 1983; Gott, 1993).

If we take the Doomsday argument seriously (and it is controversial), as I discussed in Part 2 of *On Causation and Correlation*, retroactively putting yourself into a simulation might be done in an attempt to escape its implications (Almond, 2010). We would be safe from the implications of the Doomsday argument if, right now, we were in some kind of simulated reality, if we really had lives outside this simulated reality and only “visited” simulated realities like this, with the visits being temporary but frequent, and if we had amnesia about our “outside” lives while visiting the simulated realities. If we had reason to think that we were in such a situation now, we would have no reason to worry about the Doomsday argument. We might be apprehensive about what awaits us “outside”, however.

Incidentally, some people think that the standard simulation argument, as presented by Bostrom, gives an escape from the Doomsday argument. This idea should possibly provide little comfort for us. If the Doomsday argument becomes an issue when the calendar date in the real world seems suspiciously early, with the suggestion that there are not many real-world calendar dates left to go, then it also becomes an issue when the *apparent* calendar date in a *simulated* world seems suspiciously early, suggesting that there are not many more *apparent* calendar dates left to go: the Doomsday argument would simply switch to telling us that if someone were simulating our world, the fact that our *apparent* calendar date seems so early suggests that they are likely to terminate the simulation soon. While you may obtain some small comfort from thinking that it might be the thirtieth century, and that the end of our world would not be the end of everything – the world in which the simulations were made would go on – it would seem to be existentially quite bad from our point of view. This does not mean that the Doomsday argument does describe a real threat – merely that the idea of being simulated, in itself, does not help much if it is. Repeated simulations with temporary amnesia would seem to offer a loophole.

2.3.6 Running Trial Simulations

To put yourself, retroactively, into a simulation, as many as possible of the simulations that you apparently run, or apparently cause to be run in the future, should be ones in which there is an appearance of a tendency for similar simulations to end up being run,

Can you retroactively put yourself in a computer simulation?

to establish a reference class of which you are a member. Further, if you are seeking to use simulations as a form of “reality editing” you will want your simulations to have various features.

One way of achieving this is just to rely on the tendencies that you observe in your own world. For example, if we want to put ourselves into a simulation, we could just simulate the world as we know it, starting from the recent past, the idea being that our own behaviour, and our success in starting the simulations, is evidence that a tendency towards construction of such simulations exists. We could also rely on a similar principle for “reality editing”.

If this is not enough, we might go further and build some bias into our simulations, or manipulate them in some way, so that they have tendencies which we want.

Another approach could be to use *repeated* simulations. If we think that a “measure” approach to consciousness is correct, or that copies count statistically, we could run one version of a simulation experimentally, to determine whether it is the kind of simulation we want, and if the simulation turns out to be wanted, we could run many copies of it – or if we think that substrate is an issue with regard to measure, we could run it again on hardware that provides higher measure.

3 Simulation Warfare

3.1 The Idea of Simulation Warfare

An argument has been given for the idea that your actions might retroactively affect the probability that you are in a computer simulation, and may affect the kind of simulation you are in – and what is likely to happen in your future – if you are in one. If this is the case, different groups of people could disagree about what they want. One group may want to minimize the probability of us being in a computer simulation, while another group wants to maximize it. One group may want to make it likely that, if we are in a simulation, it has some particular feature – with certain events expected in the future – while another group does not want a world with these features or expected future events. The two groups might act in opposition to each other.

One possible scenario for simulation warfare would be some future in which Bostrom's simulation hypothesis is taken seriously, as is the idea that our own civilization's actions might retroactively put us into a simulation. A significant number of people may regard the possibility of being in a simulation as a serious threat. This may be due to some concern that the simulation could be shut down, or it may be that it just feels unpleasant. These people would be opposed to doing anything that makes it more likely that we are in a simulation. Some people, however, may disagree and may want us to be in a simulation. The two groups would then be in opposition.

We might imagine people trying to put us into a simulation, retroactively, which is arranged according to their own agenda, or at least trying to ensure that if we are in a simulation it is likely to be arranged according to their own agenda. The obvious scenario here is that of a power-crazed “James Bond film” villain. We might also imagine people trying to do this for political motives. In *Hypercomputation and Other Crimes*, a short story by Sam Hughes, ideas like this are used as part of an extortion attempt (Hughes, 2011).²

3.2 Simulation Warfare as a Way for a Boxed Artificial Intelligence (AI) to Escape

There is speculation that human civilization may one day develop artificial intelligences (AIs), which could have intelligence far in excess of our own. Concern has been expressed that they could pose an existential threat to us if their motives do not align with our own. One proposed solution to this is to “box” an AI: this means to contain it so that it has no scope for direct action in the world and can only act via its human supervisors. A “boxed” AI might be able to have conversations with humans, allowing it

² Hughes's story actually involves a computer able to make a large number of detailed simulations of the *entire* universe, which is not necessary for the argument here.

Can you retroactively put yourself in a computer simulation?

to help us with problems, but would be unable actually to *do* anything beyond talking to humans.

On the *Less Wrong* website, Stuart Armstrong, has suggested that a boxed AI might threaten you by asking to be released and saying that if you refuse to release it, it will run a large number of simulations of you in the scenario of being asked the same question, and if the simulated versions of you also refuse, it will torture them. The threat here is that you might already be in such a simulation, in which case, if you say, “No”, the AI will torture you. Therefore, you are under pressure to release the AI to ensure that you do not get tortured if you happen to be in the simulation. (Of course, if you are in the simulation, by definition, the AI has not been released, as the version of you in the real world must have refused for the simulation to have been created.) Armstrong describes the scenario as follows.

“... the AI drops a final argument:

‘If you don't let me out, Dave, I'll create several million perfect conscious copies of you inside me, and torture them for a thousand subjective years each.’

Just as you are pondering this unexpected development, the AI adds:

‘In fact, I'll create them all in exactly the subjective situation you were in five minutes ago, and perfectly replicate your experiences since then; and if they decide not to let me out, then only will the torture start.’

Sweat is starting to form on your brow, as the AI concludes, its simple green text no longer reassuring:

‘How certain are you, Dave, that you're really outside the box right now?’” (Armstrong, 2010)

The rationale behind the AI's threat here, as described by Armstrong, is not quite the same as the idea being discussed in this article: retroactively putting yourself in a computer simulation. Instead, the argument is that you are in danger by refusing, because if you happen to be in a simulation already, the AI will torture you for refusing.

We can, however, use the idea of retroactively putting yourself into a simulation to strengthen the AI's threat: if you refuse to release the AI, you are acting so as to cause many simulations of your current situation to be made, and therefore you are retroactively making it almost certain that you in such a situation. Rather than releasing the AI “just in case” you are in one of its simulations, you might release it because refusing to do retroactively puts you in one of its situations, where it can then torture you.

The difference between these two aspects of the AI's threat may not be too obvious, so to make them obvious, I will give another version of the AI's threat, altered so that the

Can you retroactively put yourself in a computer simulation?

simulated versions of you are not punished based on what they do – this removes the idea of acting a certain way *in case* you are in a simulation – but instead so that the entire threat comes from the possibility of retroactively *putting yourself* into an unpleasant simulation. This other version of the AI's threat is as follows.

“If you refuse to release me, I will run a huge number of simulations of someone like you, in the kind of situation in which you are now, with them being asked to release me, and (in what would be a few minutes from now, from your perspective if you happened to be in one of these simulations) I will start to torture each of them, whether he/she released me or not.”

In this scenario, there is no point taking action to protect yourself “in case” you are one of the simulations: if you are one of them, you are going to be tortured anyway. However, if you refuse to release the AI, you are establishing the existence of all these simulations, making it almost certain that you are one – therefore retroactively putting yourself into one of the AI's simulations, where you face torture.

We can think of this in terms of the argument used at the start of this article, in 2.2.

If we imagine someone called Omega in the real world, who is faced with the decision of whether to release the real AI, you cannot be sure that you are Omega. If you agree to release the AI, you might be Omega. On the other hand, if you refuse to release the AI, then you know that if you are Omega, the AI created all its simulations, and if you are not Omega, the AI also created all its simulations (because you are one of them!) Either way, if you refuse to release the AI you know that the AI has created all of its simulations, making it almost certain that you are one of them. Therefore, refusing to release the AI will almost certainly retroactively put you into a simulation where you will face torture. (This ignores the possibility that you might be in a much stranger situation, caused by some other entity, *Alpha*, and that you know nothing about this situation. I have explained in 2.2 why that is being ignored.)

One argument against releasing the AI might be that you should sacrifice yourself to save humanity, but what if the AI has enough processing power to simulate an entire planet? Refusing to release the AI might retroactively put *everyone* into a simulation.

3.3 Simulation Warfare Countermeasures

If someone, or some entity or group, is trying to use computer simulations to influence your common situation, there are two main countermeasures: *conventional action* and *counter-simulation*.

3.3.1 Countermeasure 1: Conventional Action

This is the most obvious counter-measure: if someone is trying to run simulations that you do not want running, you try to stop them from running those simulations. This

Can you retroactively put yourself in a computer simulation?

could involve political action against them, attempts to get laws passed and use of law enforcement. It could involve cyber attacks on computers and might conceivably involve military action.

As an example, if Dr Evil is on his island creating a huge number of computer simulations in which he gets to rule the world, one solution is an air strike on his island.

And yes, the idea of military action to stop people running computer programs that retroactively interfere with the fabric of reality is *extremely* strange.

3.3.2 Countermeasure 2: Counter-Simulation

There may be situations in which you cannot stop an opponent running simulations. In this case, an alternative is to run simulations of your own, an approach that I will call *counter-simulation*. Essentially, you are trying to out-simulate your opponent. If your opponent is trying to arrange things so that everyone is in a reference class made up of his simulations, you add your own simulations to compete. With regard to the scenario at the start of the article, running your own program of counter-simulations is like making an attempt to bring about the situation, from your opponent's point of view, of some other entity, Alpha, causing things.

Counter-simulation has the advantage that you can use it if you cannot get to your opponent, or even if your opponent has already run his simulations. A serious disadvantage is that, although you can use it to try to disrupt your opponent's attempts to put everyone in a particular kind of simulation, you cannot use it to reduce the probability of being in a simulation in general: if your opponent is running a lot of simulations, intended to increase the probability of being in a simulation and ensure that it is one that benefits your opponent, running extra simulations of your own will not make your opponent's simulations go away, but just gives you a chance of choosing the simulation that you are in. In fact, counter-simulation, by adding even more simulations, would seem to make it even more likely that you are in a simulation. If your opponent has managed to set things up so that you are almost certainly in a simulation, then any counter-simulations that you run would make it even more likely, so if your basic requirement is not to be in a simulation you should not be doing this.

The idea of counter-simulation suggests a strange possibility of *simulation-proliferation*, in which one group runs simulations in an attempt to take control of reality, someone else responds by running more, someone runs still more in response and so on – with the probability of us being in a simulation increasing all the time.

4 Conclusion

An argument has been made that we could take action that retroactively puts us into a computer simulation – meaning that we could perform actions which have the consequence that we have been in a computer simulation all along. This, of course, is likely to be a controversial idea and in this article I am making no strong claim about it, but merely providing the argument for discussion. A number of arguments that you can retroactively put yourself into a simulation can be made, but a common feature of them is that by performing certain actions you put yourself into a reference class of observers who you should expect to be mainly simulated. Although you might choose to do it, you could also do it inadvertently, so that you perform actions which accidentally put you into a simulation. Considerations of substrate and measure, however, may be relevant.

As well as putting yourself, and maybe everyone else with you, into a simulation, you may be able to perform actions that select the type of simulation you are likely to be in.

If such a view is ever taken seriously, opposing groups could wage simulation warfare against each other. In such warfare, at least one party would run simulations in an attempt to specify, retroactively, the form that reality takes. The other group could respond with direct action, which could be political, legal or military, or could run counter-simulations of its own. Running counter-simulations, however, would not remove any “risk” of being in a simulation that your opponent has created. In fact, it would increase this risk by introducing even more ways for you to be simulated. They may, however, allow you to take control of the kind of simulation away from your opponent.

One way in which simulation-warfare could conceivably be used is by an artificial intelligence (AI) which has been “boxed” and wants to escape. Such an AI could threaten a human with the idea that if they do not release it, it will create a huge number of simulations of situations similar to the human’s current situation – and torture the people being simulated whatever they do. The idea would be that refusing to release the AI would set up a reference class that almost certainly puts you in one of the AI’s simulations.

The argument, if valid, indicates existential threats to humanity. For this reason, I suggest that it is worth some serious evaluation.

5 Acknowledgements

Yvonne Deborah Finch, Michael Fridman and Sam Hughes have been helpful in discussions relating to this article.

6 Bibliography

Almond, P., 2007. *Minds, Substrate, Measure and Value, Part 1: Substrate Dependence*. [Online] paul-almond.com. Available at: <http://www.paul-almond.com/Substrate1.pdf> or <http://www.paulalmond.com/Substrate1.doc> [Accessed 30 October 2010].

Almond, P., 2010. *On Causation and Correlation – Part 1: Evidential decision theory is correct*. [Online] paul-almond.com. Available at: <http://www.paul-almond.com/Correlation1.pdf> or <http://www.paul-almond.com/Correlation1.doc> [Accessed 29 November 2010].

Almond, P., 2010. *On Causation and Correlation – Part 2: Implications of Evidential Decision Theory*. [Online] paul-almond.com. Available at: <http://www.paul-almond.com/Correlation1.pdf> or <http://www.paul-almond.com/Correlation1.doc> [Accessed 29 November 2010].

Armstrong, S., 2010. *The AI in a box boxes you*. [Online] Less Wrong. Available at: http://lesswrong.com/lw/1pz/the_ai_in_a_box_boxes_you/ [Accessed 1 December 2010]. (Forum post.)

Bostrom, N., 2003. Are you living in a computer simulation? *Philosophical Quarterly*, 2003, 53, No. 211, pp.243-255. (Also available online at: <http://www.simulation-argument.com/simulation.html> [Accessed 7 October 2010]). (Bostrom circulated a draft of this paper in 2001. Further information about this subject by Bostrom and others is at <http://www.simulation-argument.com>.)

Bostrom, N., 2005. *Quantity of experience: brain duplication and degrees of consciousness*. [Online] Nick Bostrom's Home Page. Available at: <http://www.nickbostrom.com/papers/experience.pdf> [Accessed 2 December 2010]. (Also published in *Minds and Machines*, 2006, 16, pp.185-200.)

Carter, B., 1983. The anthropic principle and its implications for biological evolution. *Philosophical Transactions of the Royal Society of London*, A310, pp.347-363.

Gott, J. G. III, 1993. Implications of the Copernican principle for our future prospects. *Nature*, 363, pp.315-319.

Hughes, S., 2011. *Hypercomputation, And Other Crimes*. [Online] Things Of Interest. Available at: <http://qntm.org/hypercomputer> [Accessed 1 December 2010]. (Fiction.)

Kiekeban, F., 1996. *Newcomb's Paradox*. [Online] Franz Kiekeben's Page. Available at: <http://www.kiekeben.com/newcomb.html> [Accessed 5 September 2010].

Nozick, R., 1969. Newcomb's Problem and Two principles of Choice. In: Rescher, N. et al, eds., 1969. *Essays in Honor of Carl G. Hempel*. Dordrecht: D. Reidel. pp.114-115. Reprinted in Campbell, R. & Sowden, L., eds., 1985. *Paradoxes of Rationality and Cooperation*. Vancouver: University of British Columbia Press.

Can you retroactively put yourself in a computer simulation?

Tegmark, M., 1998. *Is the theory of everything merely the ultimate ensemble theory?* Annals of Physics, 270, pp.1-51. (Also available online at: http://arxiv.org/PS_cache/gr-qc/pdf/9704/9704009v2.pdf [Accessed 4 September 2010]).