
Causation and Correlation

Part 2: Implications of

Evidential Decision Theory

By Paul Almond

9 October 2010

Website:

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

E-mail:

info@paul-almond.com

This is the second in a two-part series of articles about *evidential decision theory*. The first article, *On Causation and Correlation – Part 1: Evidential decision theory is correct.*, is available at <http://www.paul-almond.com/Correlation1.pdf> or <http://www.paul-almond.com/Correlation1.doc>. There are two different approaches in decision theory: evidential decision theory and causal decision theory. Evidential decision theory seeks to maximize the utility of a choice, taking into account what the choice tells you about yourself, and therefore, about other parts of the world that may correlate with your own behavior. The first article provided a justification of evidential decision theory. This article discusses implications of it. Evidential decision theory has ethical implications, because your own actions can be viewed as correlated with the actions of others, meaning that the ethical consequences of your actions go beyond the causal ones. This is useful in dealing with the problems caused for ethics if reality is infinite and the utility of a particular state of reality cannot be measured, because it allows us to think of ethical actions in terms of altering the *density* of ethical outcomes, due to correlation, throughout all of reality. Evidential decision theory might also be relevant when applied at the level of our entire civilization. Evidential decision theory is most relevant when we know the least about the reference class, and we only have experience of our own civilization, meaning that our decisions might provide us with significant information about the behavior of other civilizations, and that we might *meta-cause* them to behave in certain ways. This has implications, in terms of being able to control our situation with regard to interaction with alien civilizations, the simulation argument, the Fermi paradox and the Doomsday argument. Evidential decision theory may also be relevant to *self-control*.

Table of Contents

1 Introduction	3
2 Personal and Civilization-Level Decisions	4
3 Ethical Behavior	5
3.1 Ethics in General	5
3.2 Being Good in an Endless Reality	5
3.2.1 The Exploding Puppies Scenario	5
3.2.2 The Problem of Being Good in an Endless Reality	7
3.2.3 Temporal Discounting?	7
3.2.4 Evidential Decision Theory as a Solution	8
4 How not to find out that you are in a bad situation	10
4.1 Civilization-Level “Karma”	10
4.2 The Simulation Argument	10
4.3 The Fermi Paradox	14
4.4 The Doomsday Argument	15
5 Self-Control and Reference Class	16
6 Conclusion	17
7 Bibliography	19

1 Introduction

This is the second in a two-part series of articles about evidential decision theory. The first article, *On Causation and Correlation – Part 1: Evidential decision theory is correct.*, is available at <http://www.paul-almond.com/Correlation1.pdf> or <http://www.paul-almond.com/Correlation1.doc>.

Most people probably think of causal decision theory as the correct way to make decisions. With causal decision theory, you evaluate a possible choice in terms of the causal consequences following on from making it. With evidential decision theory, on the other hand, you take into account everything that you find out about reality as a result of making the decision, and this includes what the decision tells you about yourself and your past, and the behavior of anything else in reality that would be expected to correlate with your behavior due to similarity: Evidential decision theory takes into account the behavior of entities like you, and views you as deciding on their behalf, to some extent, when you make your own decision.

The first article provided a justification for evidential decision theory. In that article, it was stated that evidential decision theory would be hard to distinguish from causal decision theory in everyday life, because evidential decision theory is most relevant when we have the least information: In everyday life, we have an enormous amount of information about how other people around us should behave and about the past. This was summarized with the following statement.

Lack of knowledge is power.

This means that the more we know, the less scope we have for obtaining new information from a decision that we make, and therefore the less control over reality is given to us by that decision.

This article will discuss some of the implications of evidential decision theory. There are two obvious ways in which evidential decision theory may become relevant. It can be relevant to discussions of ethics. It can also be relevant when we apply it at the level of our entire civilization: We know very little about the reference class of possible civilizations as we only have our own to observe, so our own decisions might give us significant information about the expected behavior of other civilizations, meaning that we could regard ourselves as having the ability to meta-cause other civilizations to behave in various ways.

We will start by discussing the difference between applying evidential decision theory at the level of decisions taken by individuals and applying it at the level of decisions taken by an entire civilization.

2 Personal and Civilization-Level Decisions

In the previous article, it was said that lack of knowledge is power, and that evidential decision theory becomes distinguishable from causal decision theory when we have very limited knowledge about a reference class of events of which our decision is a member, or about our past. It was pointed out that in everyday life we tend to have a lot of information about other people, so it is unrealistic to expect to be able to acausally control other people's decisions to any significant degree.

This might suggest that evidential decision theory is of no consequence, beyond that of contrived situations. There is a way in which it might become relevant, however: if we apply it at the level of a *civilization*. The word "civilization" is being used here in the way in which it tends to be used when we are talking about extraterrestrial aliens. With this usage of the word, all of humanity is a single civilization and, at the time of writing, no extraterrestrial civilizations have been discovered, so we only know of one civilization: ours. This means that it could be argued that our lack of knowledge of other civilizations means that how our civilization behaves is one of our main sources of information about them – that if we behave in a certain way we should think it more likely that other civilizations behave in that way. *Evidential decision theory, therefore, suggests that we might be able to control alien civilizations, acausally, to a significant degree.* If true, this could have good or bad consequences. We might be able, inadvertently, to meta-cause an alien civilization to do something that threatens us. On the other hand, it might mean that we can deliberately behave so as to meta-cause alien civilizations *not* to threaten us.

An objection to this might be that other civilizations may have had contact with each other, and so would know much more about the reference class than we do. This might be answered by saying that the fact that we are in a civilization which has not made contact with others suggests they are typical, which is an application of the Copernican principle, by saying that even if some civilizations have had contact with each other, we can at least meta-cause the ones that have not had such contact to behave in a particular way or by saying that if any number of civilizations are in contact with each other, we can regard them as a single civilization. Furthermore, the simulation argument will be discussed later, in 4.2, and some of the civilizations of interest here will be ones that are very similar to a possible near-future of our own civilization.

3 Ethical Behavior

3.1 Ethics in General

If we can acausally affect the behavior of other beings, then this has implications for ethics. It implies that the ethical consequences of an action are not limited to what follows causally from it, but by what the action *meta-causes*: It makes the ethical affects of actions potentially greater. When you decide to be good, you can consider yourself meta-causing other beings to be good, and when you decide to be bad, you can consider yourself meta-causing other beings to be bad: It makes our actions *more important*, in ethical terms.

This could be considered at the level of the individual and at the level of an entire civilization. In the previous article, I pointed out that evidential decision theory would have little practical relevance in everyday life. This situation is a bit different: As an individual, your actions would still not acausally affect anyone else to a significant degree, so from a practical view, in terms of anything that is likely to happen to you, you probably do not need to worry about acausal effects of your actions, but from an *ethical* point of view, things may be different. The fact that this slight effect may be occurring all over reality – that a single action by you might be meta-causing other beings to perform actions like it all over reality – might be ethically very significant: A good action or bad action gets duplicated many times.

If you think you should be ethical, then this is extra support for being ethical. One way of imagining this is as resembling the common situation in which someone offers to increase the donations that others make to a charitable cause; for example “For every \$1 you give, I will give a dollar!” For each of your good acts, evidential decision theory makes an “extra contribution”, increasing your effectiveness.

One thing that may determine how much ethical value evidential decision theory adds to an act is the size of reality: If there is a lot of reality for evidential decision theory to operate in, maybe that suggests that it can add a lot of value, even if the correlation is very weak? We can take this further, however. Evidential decision theory might give us an answer to what should be regarded as a serious problem: how to justify ethical behavior in an *infinite* reality. This will now be discussed.

3.2 Being Good in an Endless Reality

3.2.1 The Exploding Puppies Scenario

We will start this part of the discussion with a scenario.

You are standing at the end of an infinite line of boxes: The line of boxes extends into the distance for infinity. In each box is a puppy and each box has an explosive charge placed on it. Near the first box are two buttons. One button is labeled “Detonate” and

On Causation and Correlation - Part 2: Implications of Evidential Decision Theory

the other button is labeled “Deactivate”. If you press the “Detonate” button, the explosive charge on the first box is detonated, and the box is blown up, killing the puppy inside. If you press the “Deactivate” button, the explosive charge on the first box is deactivated and the puppy is spared.

Ethically (and assuming you care about puppies) the choice here is quite simple, right? You can kill a puppy or spare it and the *good* thing to do is to spare it. However, I will now reveal something else about the scenario.

The button that controls the explosives in the first box is also connected to a mechanism that controls the explosives in the second box, which is in turn connected to a mechanism that controls the explosives in the third box, and so on for infinity. When the explosive charge in a box is detonated or deactivated, a signal indicating whether detonation or deactivation occurred for that box is sent to the mechanism controlling the explosives in the next box. This mechanism is preset either to detonate if and only if the explosive charge in the last box was detonated, and otherwise deactivate, or to detonate if and only if the explosive charge in the last box was *deactivated*, and otherwise deactivate. The mechanism controlling the explosive charge for every box (except the first one, which you control directly) is like this, waiting for its prompt from the previous box and then acting according to this and its internal logic. You are unaware of any pattern with which the two kinds of mechanisms occur. As far as you are concerned, boxes which detonate if the last box detonated and boxes which detonate if the last box were deactivated are distributed randomly. Each, however, occurs with the same frequency. If you press the “Detonate” button, you will see the first box explode, and if you press the “Deactivate” button the first box will be deactivated; however, whichever button you press, you will then see a chain reaction of detonations/deactivations progressing along the line of boxes, with each box either detonating or deactivating in turn. You will not see any pattern to this: If you choose to detonate for the first box, you will see one chain reaction of detonations/deactivations start and if you choose to deactivate for the first box, you will see another chain reaction of detonations/deactivations.

The question, now, is: What is the *good* thing to do in this scenario? In either case, the first puppy is killed or saved, and then *an endless sequence of puppies being killed or saved starts*. Whatever you do, there is no end to the effects of your actions and *there is no final number of puppies that is blown up as a result of your actions*. The easy answer seems to be to press the button to deactivate the explosive charge on the first box and spare the first puppy, but what does this achieve? Is it supposed to blow up fewer puppies in total? An easy way out may be to say that if you press the “Detonate” button you get the first puppy blown up + an endless sequence and if you press the “Deactivate” button you get the first puppy saved + an endless sequence, so the two situations are ethically the same, differing only in respect of what happens to the first puppy, but this is a cheap way out really: If we are to take a utilitarian view on this, whatever happens to the first puppy should be considered within the context of what

happens to all the puppies – and when we add the first puppy to the sequence and just look at the overall results of everything we do, we are left with an uncomfortable result: The overall situation is no different whichever button we press. Of course, the local situation is different: We may feel some responsibility to that first puppy, but that is just the local situation and there will be endless puppies to follow which our decision will be blowing up or saving. Any one of these puppies could be imagined to be “the local situation” by the simple expedient of transferring your perspective there, if you want to feel some responsibility over it as well.

3.2.2 The Problem of Being Good in an Endless Reality

Some cosmological views suggest that what we think of as “the universe” may only be a small part of reality, and that reality may be unlimited – containing an infinity of things and events. As just one example, Max Tegmark has proposed that reality consists of every mathematically describable object – and that the observable universe is just one of these. Tegmark’s reality is unlimited (Tegmark, 1998).

Our actions would affect an endless reality in one of two ways.

Each action we performed might have an unlimited chain of consequences, like the sequence of puppy explosions in the exploding puppies scenario, above.

Each action we perform might have a limited affect on reality, even though reality is infinite. It would simply be the case that we live in some finite part of reality which cannot interact with the rest of reality in any way, so that what happens elsewhere is nothing to do with what happens here.

Right now, which of these would be more likely is unimportant. It should be noted that the “unlimited chain of consequences” in (1) might not be exclusively temporal: This does not necessarily mean assuming infinite time. It may be, for example, that temporal relationships are only a special case of relationships and that more general relationships are possible, so that what happens in “our” space-time is related to what happens “somewhere else” in an atemporal way.

3.2.3 Temporal Discounting?

In a limited reality, the problem is not quite as serious, because there can be, in principle, an overall result of your actions. A less-serious problem does still arise, however, just from our actions having consequences that become increasingly difficult to predict further into the future. One solution that has been proposed to deal with limitations in prediction is a variation on *temporal discounting*. This would involve assigning less importance to consequences which we are less able to predict as following from our actions. We would regard ourselves as responsible for foreseeable consequences of our actions in the near future, but would effectively ignore consequences in the distant future, or events which had no clear relationship to our actions. We could easily extend this to ignoring spatially distant consequences of our

actions, or to ignoring events that are remotely connected in some atemporal way to our actions in some possible cosmologies, or events that are not connected at all with our actions.

We might try to use something like temporal discounting to justify ethical behavior in an endless reality: By assigning importance to the events dependent on the degree with which we can predict them from our actions, we would be dealing with a finite amount of reality, and whatever happened in the wider, infinite reality would be irrelevant. Temporal discounting like this is the obvious answer that people are going to give. At a stroke, it declares the problem non-existent. If we apply temporal discounting, we merely engage in local ethics and ignore everything else.

Does temporal discounting really work, however? I suggest that if we use causal decision theory, temporal discounting is usable as a practical method – it can clearly help us define how we should act coherently – but it does not resolve the problem of ethics in an endless reality in any philosophically strong way. It merely declares those problems to be irrelevant, which is another matter entirely. Even if we temporally discount, it does not change the situation that, in an endless reality, there is apparently no overall “improved state of reality” that results from our actions. This problem may be lost on a lot of people, because temporal discounting is practically forced on us by causal decision theory. If we do not discount, what *are* we supposed to do? We can hardly take account of all of infinity and the idea of disregarding ethics entirely is not going to get much support. If we want to act coherently, we really have to temporally discount. The fact that we have to do it may lead people to think that it must therefore have some strong philosophical justification. I suggest that, in an endless reality, with causal decision theory, it has no strong justification and is merely the last resort if we want to act coherently, that faced with the choice of giving in to despair or caring about those close to us, we do the latter because we can do nothing else.

Ideally, we would have something better than just temporal discounting to justify being good: maybe evidential decision theory can help?

3.2.4 Evidential Decision Theory as a Solution

With evidential decision theory, your decisions might correlate with similar decisions throughout all of reality. There is no limit to this: If reality is infinite then your decisions correlate with an infinity of other decisions, meaning that each decision that you make can be regarded as meta-causing an infinity of other decisions.

How should we regard this? It may seem tempting to say that a “good” act now becomes “infinitely good”, but that would make little sense: *Everything* would correlate with an infinity of things like it over all of reality, no matter how weak the correlation, so there would be no way of enumerating the total consequences of anything. An infinite reality deprives us of the ability to have a final accounting of the effects of a decision on reality, yet it may provide an alternative way of justifying an ethical action in terms of its

On Causation and Correlation - Part 2: Implications of Evidential Decision Theory

effects on the *density* of ethical actions in all of reality. If you make a decision ethically, then this can be regarded as meta-causing similar decisions throughout all of reality, and therefore increasing the density with which such decisions occur in reality. That density would not face any problems from reality being infinite: It would be a real value. Your actions then, while not affecting any total *amount* of “goodness” in reality would affect the *distribution* of “goodness” throughout reality.

This could actually be regarded as a justification of temporal discounting: You can make your decisions while temporally discounting everything except local consequences, knowing that, in the parts of reality that are beyond your observation, other entities are making similar decisions to be ethical in their local neighborhoods.

Some people may think that the idea of justifying ethics in terms of increasing the density of ethical actions is too abstract to mean anything to us. A thought experiment can suggest otherwise. Suppose you are standing on an infinite checkerboard, made of black and white squares. On every square of the checkerboard, a puppy is suffering horribly. Now, if you like puppies, this should distress you. You might try to save one puppy, but in itself this will not do anything to the total amount of suffering, because there is no total amount. Suppose, however, you can rescue every puppy that is on a black square, so that all the black squares no longer have suffering puppies on them and only the white squares have suffering puppies. Again, you have not affected any total amount of anything, but you have affected the density of suffering in a real way. You may doubt that this means anything, but would you not feel better as you wandered over the checkerboard seeing that the suffering on the white squares was gone, and knowing that it was gone over all of the infinite board?

4 How not to find out that you are in a bad situation

4.1 Civilization-Level “Karma”

Applying evidential decision theory at the level of our entire civilization means that we might expect the behavior of other civilizations to correlate with ours, so that we can meta-cause other civilizations to behave in various ways. As we have minimal knowledge of the reference class – we only know of our own civilization – the effect might be worth taking note of.

This could have practical implications for us if any of these civilizations might interact with us, or might even be involved with our situation in ways that go beyond the simple interaction that might arise from a mere meeting. It suggests that our own behavior, as a civilization, could correlate with the behavior of other civilizations of which we are on the receiving end: Our own behavior could come back at us. This has some similarity with the Buddhist idea of karma, at least at the level of a civilization, though it should be noted that karma, in Buddhism, involves issues such as the moral quality of the intention: The resemblance may be superficial.

A simple way of viewing this is in terms of how ethical we are: If we behave unethically as a civilization, evidential decision theory would suggest that we are meta-causing other civilizations that we may encounter one day to be unethical towards us, and thereby putting ourselves in danger.

There are some more interesting variations on this kind of reasoning, which will now be discussed.

4.2 The Simulation Argument

The simulation argument, by Nick Bostrom, raises the idea that we may be in a computer simulation – made either by humans at a later time than it now appears to be (an “ancestor” simulation) or by a non-human civilization (Bostrom, 2003). 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) do not apply, then there would be a huge number of simulated realities – and the number of simulations of situations like ours – 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.

The simulation argument will not be discussed in detail in this article, and nor will the merits or otherwise of it. What is relevant is that, if you take the simulation argument seriously, then evidential decision theory would seem to allow you to assert some control over the other civilizations that might be building these simulated realities.

One way in which evidential decision theory would be relevant is in the way it allows you to control the probability that you are in a simulation in the first place. If your civilization decides to develop the capability to run simulated realities, then you are meta-causing civilizations in general to do likewise (including civilizations on which our own might be modeled), and making it less likely that almost all civilizations end before they are capable of producing simulated realities, in turn making it more likely that you are in a simulated reality. If, however, your civilization decides *not* to acquire this capability then you are meta-causing civilizations in general to do likewise, making it less likely that you are in a simulated reality. Once your civilization has the capability to produce simulated realities, if your civilization decides to do it, this would make it more likely that other civilizations also do it, again making it more likely that you are in a simulated reality. On the other hand, if your civilization decides not to produce simulated realities, this makes it less likely that other civilizations would choose to do so, and therefore less likely that you are in a simulated reality yourself.

This is not “all or nothing”. If your civilization *does* decide to acquire the capability to produce simulated realities and/or decides to go ahead and produce them, the earlier this happens, the greater is the probability of being in a simulation, because this can be regarded as meta-causing other civilizations to acquire the capability or use it earlier as well – making it more likely that they will have time to do this while still in existence. If your civilization is producing simulated realities then the number it produces will be relevant: If your civilization produces a huge number of reality simulations, then this could be regarded as meaning that civilizations in general are expected to do likewise, increasing the likelihood that you are in one, whereas if your civilization produces few simulated realities then it will suggest that civilizations in general produce few of them.

Some readers may say, about all this, that it will be practically impossible to prevent the capability being developed – that it was already too late when the first flint tool was made – and they might also say that such a capability would almost inevitably be used, and that even if human civilization starts with only a small number of simulated realities, once the line has been crossed there are going to be a lot one day.

Evidential decision theory is important then, with regard to the simulation argument, because, if we take the simulation argument seriously, it seems to give a civilization some control over whether it is in a simulated reality or not. The very small amount of information we have about the reference class – just one civilization – means that, here,

On Causation and Correlation - Part 2: Implications of Evidential Decision Theory

conditions are ideal for evidential decision theory to become relevant and for this to be a serious matter.

Evidential decision theory is not restricted to the issue of whether we are in a simulated reality. If we are in a simulated reality, it might be relevant in allowing us to control the probabilities that we are in various kinds of simulation. If we construct many simulated realities in which various things happen, then if another civilization is simulating us, we might be meta-causing it to make those things happen to us. This creates an argument for being kind to the inhabitants of any simulated realities that you do make. A particularly obvious issue here is how long the realities that you simulated last for: If you run huge numbers of simulated realities that only go through a relatively short period of history and then terminate, then evidential decision theory would suggest that you might be meta-causing your own world to end soon as someone else terminates their simulation.

All this may seem removed from the world of individuals: We got evidential decision theory to matter, here, by using it far above the level of individuals, at the level of an entire civilization. This does not mean, however, that all this is irrelevant to an individual. An individual who thinks that his/her civilization should behave in a particular way, on evidential decision theory grounds, can attempt to influence his/her civilization's decisions, for example by voting or persuading others. The individual can do this by making decisions in the conventional, causal way, so as to influence the civilization in such a way as to make the appropriate decisions according to evidential decision theory. The reality, of course, is that this would work as a result of your decisions correlating with those of entities that have a similar role in the decision making process of another civilization. In this respect, this would involve evidential decision theory at the level of the individual; however the correlation can only affect the individual significantly by being expressed through a civilization and the individual can act as if evidential decision theory only applies at a higher level.

If evidential decision theory might matter at the level of a civilization because of our lack of knowledge of the reference class, this suggests the idea of a civilization attempting to find out more about the reference class – and about its own probable status – by running a large number of simulations of possible civilizations and finding out how many of them construct their own simulations. It might be thought that doing this could provide you with the information about the reference class needed to make your own decisions irrelevant and free you from the implications of evidential decision theory. This, however, would be a fallacy. Your own decision to do this in the first place would suggest that other civilizations would tend to do likewise, so the decision to start all these simulations would actually be *meta-causing* the simulated civilizations to engage in constructing simulated realities themselves. The decision to launch the simulations would be a kind *fast-forwarding*, as discussed in the previous article. This does not necessarily mean that any correlation is going to be particularly strong, but doing this would not insulate you from any correlation that does exist.

On Causation and Correlation - Part 2: Implications of Evidential Decision Theory

This reasoning might be taken further. Even if our civilization does not decide to launch large numbers of simulated realities, if we do *anything* to try to find out about the reference class of possible civilizations, this in itself could be viewed as meta-causing civilizations in general to do things to find out about the reference class – and some of them might do this by constructing simulated realities. Taking this to an extreme, it could be argued that even *thinking* about the reference class risks meta-causing your civilization to be in a simulation.

If evidential decision theory suggests that a way of meta-causing yourself not to be in a simulated reality is to avoid constructing simulated realities yourself, or to limit their numbers, this might be considered an objection to the simulation argument: If we choose not to construct simulations, or to limit their numbers, and if we think the correlation with civilizations in general will be strong enough, then we may presume that civilizations in general tend not to construct simulated realities so as to reduce the “risk” of being in simulated realities themselves. This argument may have limited force, however: We may not think that the correlation is so high as to feel so safe, and it may be that a typical civilization tends to be unable to restrain themselves before vast numbers of simulations are made and it is too late – at which point the civilization has to face the uncomfortable realization that this suggests that it may be too late for all the other civilizations too.

While evidential decision theory might be applied to try to reduce your “risk” of being in a simulated reality, some people, and some civilizations, might not see it that way: They might think that being in a simulated reality could have benefits if the entity that constructed the simulation is kind; for example, the inhabitants of the simulation might be protected from existential risks to their civilization, or they might be provided with an afterlife. Evidential decision theory suggests the possible tactic of making large numbers of simulated realities in which the inhabitants are treated kindly as a way of trying to meta-cause civilizations in general to do the same thing. This would be going further than what I said previously about treating the inhabitants of your own simulations kindly: This would be done so as to make it more likely that you are in a simulation, and that it is one in which you will be treated kindly. We might imagine a civilization doing this as a way of trying to use evidential decision theory to pluck an afterlife out of nowhere for itself, if it has recently acquired the computing power to simulate many civilizations, and provide them with an afterlife, but does not yet have technology such as mind uploading which it might use to obtain an afterlife more directly. A civilization might attempt this even if it does not yet have the computing power to construct simulated realities: It might set up some kind of legal or corporate framework to ensure that large numbers of ancestor simulations, complete with an afterlife, are constructed in the future, the idea being to strengthen the case that it is itself in such a simulation, made by a civilization with a past that is strongly correlated with its own present. Someone might even set up some organization for this purpose as a result of reading this article!

There is the issue of how strong the correlation will be in all this. There is one situation in which we might expect the correlation to be strong. If we are in a simulation constructed by future humans to simulate their own past – an ancestor simulation – we might expect the correlation between ourselves and the real past of that civilization to be particularly high, because it should have been contrived that way. Evidential decision theory would suggest to us that we could take steps to use the high correlation to manage the probability that we are in such a simulation. If we perform actions that make it seem less likely, from our point of view, that our descendants will construct large numbers of ancestor simulations, then it is less likely that there is any strong correlation between our decisions and the civilization that we are supposed to resemble, and this would mean that it would be less likely that we are in such an ancestor simulation. Alternatively, we might perform actions to make it *more* likely that we are in an ancestor simulation: maybe one with an afterlife.

An idea like this does not rely just on reducing the probability of ancestor simulations ever being made. Any action in your world that delays the construction of the first ancestor simulation might make it less likely that you are in one, though the argument would become more complex here. It should be noted that an ancestor simulation is only one type of simulation that can be imagined: You might take steps that affect your chances of being in an ancestor simulation, but they may not have much effect on the chances of being in a simulation made by non-humans. Nevertheless, changing the probability that you are in an ancestor simulation would still affect the probability of being in a simulation in general, as it contributes to it.

No claim is being made in any of this that the correlation would *necessarily* be strong. We should consider the issue and the degree of correlation we should expect, because our minimal knowledge of the reference class suggests that this could be an issue.

4.3 The Fermi Paradox

The Fermi paradox (Jones, 1985) is the issue of why we do not have any observationist evidence of alien civilizations, given that the size and age of the galaxy is, in the belief of many people, more than adequate to have allowed many such civilizations to have come into existence by now.

Numerous explanations have been offered for the Fermi paradox. Some people think that aliens are simply so far away that they have not had time to get here yet, or that alien civilizations tend not to expand into space. Those are the *nice* answers. Darker possibilities can be imagined. One of these is that civilizations almost always destroy themselves, or otherwise cease to exist, before they are sufficiently advanced to expand on an interstellar scale. This suggests the idea that we might apply evidential decision theory at the level of our own civilization to make it more likely that one of the “nice” possibilities is the right one: If we decide not to expand into space, this could be viewed as meta-causing other civilizations not to expand into space, providing a “nice”

resolution of the Fermi paradox, whereas, on the other hand, if we decide to expand into space, this suggests that other civilizations do likewise, increasing the probability that an unpleasant explanation – such as technological civilizations having a short lifespan - is the answer.

As well as suggesting a possible way to act for a civilization, evidential decision theory is providing, with all this, a possible mechanism to explain the Fermi paradox: Everyone is applying evidential decision theory and staying at home to allow the Fermi paradox to have a *nice* answer.

As with what was just said about the simulation argument, in 4.2, there is no claim being made here that the correlation would be strong, but our minimal knowledge of the reference class suggests it as a possible issue.

4.4 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 a civilization thinks that the Doomsday argument suggests the existence of a serious existential threat, it might attempt to apply evidential decision theory in strange ways to escape this threat. If human civilization came to this conclusion, for example, we might take steps to ensure that the actual period in history is unknown to us for a lot of the time. One way of doing this would involve humans regularly entering simulated realities in which their knowledge that they were in a simulation was temporarily taken away. For example, people in the twenty-third century might choose to enter simulations, on a temporary but regular basis, which cause them to think that they are in the twenty-first century. The idea of this would be that it could be viewed as meta-causing civilizations in general to take “vacations” in simulated earlier eras like this, so someone who is having the experience of living in the twenty-third century might reasonably think he could be living in a much later era, and that he just happens to be spending one of his periods in the simulation before returning. If you have reason to think that you are in a much later period of history, and will shortly be returned there, the Doomsday argument may be less of a worry.

5 Self-Control and Reference Class

Evidential decision theory may be relevant in exercising self-control, and this could be at the level of an individual person. Any decision you make might be regarded as part of a statistical pattern on which your future behavior is based, so that when you make a decision you are also deciding, to some degree, what your future self is going to do.

This will probably not be news to many people. It is routine for people trying to diet or to give up smoking to act as if their behavior at any time has consequences for future behavior.

One aspect of this that may be interesting is that the statistical link between our past behavior and our future behavior may be explicit. Some views of the mind are based on the idea that your “self” is just an object in a model, based on previous events, and that intentions and decisions are just attributes of that object (Metzinger, 2003, 2009). If such a view were correct, past behavior would affect future behavior by providing a pattern which is projected into the future by a modeling system.

6 Conclusion

The implications of evidential decision theory have been discussed. For practical purposes, evidential decision theory does not give results that are very different to causal decision theory in most everyday situations involving decisions made by individuals. The degree of control – the ability to meta-cause things – offered by evidential decision theory depends on the information gained as a result of the decision, and the amount of this information will be reduced if we already know a lot: Evidential decision theory becomes most relevant when we know the least about our situation. Decisions made by a civilization, however, may be a different matter. Our civilization has little information about the likely behavior of any other civilizations that may exist: We have little knowledge, beyond our own behavior, of the reference class of which we are a part. This means that evidential decision theory may be more important at the level of a civilization and that, from the point of view of a civilization it may be possible to *meta-cause* other civilizations to behave in a certain way by making particular choices.

Viewing decisions at the level of our civilization suggests that our own decisions could meta-cause an alien civilization to act towards us in a particular way.

Civilization-level application of evidential decision theory could have some relevance to the simulation argument. Our actions may be regarded as meta-causing other civilizations, including ones with a past very similar to our present day, that might be simulating us to behave in a similar way. This means that we might view ourselves as having some degree of control over whether we are in a simulation. Some readers will find this utterly objectionable, saying that the issue of whether we are in a simulation or not has already been decided before we do anything; however, viewing things this way is entirely consistent with evidential decision theory, as justified in the previous article, and the idea of taking meta-causation into account when making decisions. We might also be able to control the way that we are treated if we do happen to be in a simulation. We might imagine taking steps to reduce the chance of being in a simulation, but some people might want to *increase* the chance, particularly if they can also take steps to make it likely that the simulation has attractive features, such as an afterlife. Correlation between an ancestor simulation and the civilization constructing it should be particularly high, so ideas like this are more relevant here.

Civilization-level application of evidential decision theory could have some relevance to the Fermi paradox. Some answers to the Fermi paradox suggest a bleak outlook; for example, one answer to it says that technological civilizations tend not to survive long – and that would apply to us. A civilization might choose not to spread deep into space, so as to meta-cause other civilizations to behave likewise and provide an answer to the Fermi paradox that is not an existential threat.

Civilization-level application of evidential decision theory could be relevant to the Doomsday argument: A civilization may take steps to ensure that beings in the

On Causation and Correlation - Part 2: Implications of Evidential Decision Theory

civilization do not always know what period of history they are in, for example, by using simulated realities, with the idea that this would meta-cause civilizations in general to behave in this way, meaning that the current date could be wrong.

Decisions made by our civilization are the end-product of decisions made by individuals. An individual in a civilization may make decisions according to causal decision theory to try to bring about behavior by the civilization as a whole which is justified by evidential decision theory. This might apply in general where there is some entity making decisions about which evidential decision theory is relevant, and where the entity is made up of individual agents that act according to causal decision theory.

Evidential decision theory could have some relevance to control of your future behavior: Your behavior in the present could be regarded as a statistical indicator of your future behavior.

Evidential decision theory may be important with regard to ethics. If your behavior correlates with that of other agents, deciding to be ethical or otherwise is not only choosing for yourself, but also, to some extent, for those agents that correlate with you: You can meta-cause other beings, throughout all of reality, to be good or bad. This concept can be useful with regard to justifying ethics in an infinite reality, such as that proposed by Tegmark. If reality is infinite, there is no way of measuring the total utility of any state of reality, but if your actions meta-cause similar actions over all of reality then you can view the consequences of your actions in terms of affecting the density of ethical actions throughout all of reality – a concept which can have meaning even in an infinite reality. Perhaps we should regard this as a hopeful, pleasing conclusion: It suggests that even in an infinite reality, our actions can make a noticeable difference, and that the effort to be good may be worth it.

7 Bibliography

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 9 October 2010].

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>.)

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.

Jones, E.M., 1985. "Where Is Everybody?": An Account of Fermi's Question. *Los Alamos National Laboratory Report*, LA-10311-MS UC-34B, March 1985. (Also available online at: <http://www.fas.org/sgp/othergov/doe/lanl/la-10311-ms.pdf> [Accessed 7 October 2010]).

Metzinger, T., 2003. *Being No One: The Self-Model Theory of Subjectivity*. Cambridge (MA): MIT Press.

Metzinger, T., 2009. *The EGO Tunnel: The Science of the Mind and the Myth of the Self*. New York: Basic Books.

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 8 September 2010]).