# On Causation and Correlation Part 1: Evidential decision theory is correct.

By Paul Almond

28 September 2010

Website: http://www.paul-almond.com
E-mail: info@paul-almond.com

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. A justification for evidential decision theory is given. This first involves a scenario intended to suggest evidential decision theory as an approach. Some objections to evidential decision theory being used in Newcomb's paradox are that it seems to imply reverse causation, but it is shown that this issue is raised by any decision anyway. Light cones are used to give a simplified view of events, in which it is shown that there is no profound transition involved in going from an event causally following from a choice to one related to it less directly. The view of an outside observer is taken to show how decisions should be approached with no assumption of them having any special status as a result of being "owned" by the decider. Making a special case of your own decisions violates the Copernican principle. It is argued that, even if we try to view our decisions causally, correlation between other parts of reality will mean that choices tend to "contaminate" much of the description of reality in non-causal, indirect ways. Evidential decision theory can be justified by considering identical players in a game, and then considering *almost* identical players. The term "meta-causation" is proposed. A choice *meta-causes* an event if it corresponds to that event irrespective of whether or not the event causally follows it. Evidential decision theory is correct, but has little practical significance in many everyday situations in which we have a lot of knowledge. The next article will discuss situations where it could be relevant.

# Table of Contents

1 Introduction ..... 5
2 A Scenario Involving Actions and Correlation ..... 7
2.1 Description of the Scenario ..... 7
2.2 Considering the Scenario ..... 8
2.3 The *Control Group* Version of the Scenario ..... 10
2.4 The *Time-Delayed* Control Group Version of the Scenario ..... 13
2.5 Does the size of the control group dilute the importance of your decision? ..... 16
2.6 The Importance of What You Already Know ..... 16
2.7 *Amnesia* Versions of the Scenario ..... 17
2.8 Could your decision increase your estimated risk while actually making you *safer*? ..... 17
2.9 Expected *correlation* has a role in decision-making. ..... 19
3 Arguing the Case Further ..... 20
3.1 Decisions and Previous States of Reality ..... 20
3.1.1 Determinism implies apparent backward causality for *all* decisions. ..... 20
3.1.2 Non-determinism does not make the issue go away. ..... 21
3.2 The Edges of Causality ..... 23
3.2.1 Using Light Cones for a Simplified View of Causality ..... 23
3.2.2 Light Cones, Causality and Correlation ..... 23
3.3 Seeing Your Decision from Outside and the Copernican Principle ..... 27
3.4 How do you stop your choice contaminating your model? ..... 30
3.5 Degrees of Similarity ..... 32
3.6 The Fallacy of Trying to Separate the Decision Process from the Model ..... 33
3.7 The Fallacy of Thinking that Evidential Decision Theory is Just About "Trying to Get Good News" ..... 34
4 How We Should Decide ..... 36
4.1 The General Idea of Decision-Making ..... 36
4.2 The Three Ways of Obtaining Information from an Event ..... 36
4.2.1 Forward Causal Structure ..... 37
4.2.2 Backward Causal Structure ..... 37
4.2.3 Reference Classes ..... 37
4.3 Generalizing Further ..... 37
4.4 Reference Classes as a Generalization ..... 38
4.4.1 A General Approach to Describing Reality ..... 38

4.4.2 The Patterned Floor Analogy ..... 38
4.4.3 Disregarding Ownership ..... 39
4.5 *Lack* of knowledge is power. ..... 39
4.5.1 Why Knowledge Matters ..... 39
4.5.2 Forward Causal Structure and Knowledge ..... 40
4.5.3 Backward Causal Structure ..... 41
4.5.4 Reference Classes ..... 42
4.5.5 The Strangeness of the Idea that *Lack* of Knowledge is Power ..... 43
4.6 Analogy with Correlation in Everyday Life and Special Relativity ..... 44
4.7 What about knowledge of your own cognitive processes? ..... 45
4.8 General Artificial Intelligence ..... 46
5 What is a "decision"? ..... 48
5.1 "Controlling" Reality ..... 48
5.2 Wanting a Special Status for Causal Relationships ..... 49
5.3 Concerns About "Free Will" ..... 49
5.4 The Language of "Decisions" ..... 50
5.5 Meta-Causation ..... 52
6 Conclusion ..... 53
7 Acknowledgements ..... 56
8 Bibliography ..... 57

# Table of Figures

Figure 1: Two Events and a Decision ..... 24
Figure 2: Two Events in Close Proximity ..... 25
Figure 3: Further Events ..... 26

# 1 Introduction

There is some controversy about whether *causal decision theory* or *evidential decision theory* is the best approach to making decisions. Causal decision theory is the way that most people would probably, intuitively think about making decisions: You make the choice which seems to bring about, causally, the most desirable results. Evidential decision theory, on the other hand, involves making the choice which causes you to obtain information about reality which indicates that the situation is the preferred one. From this, the distinction may not be too obvious. Where things get strange is that, if we use evidential decision theory, whenever you have made a choice, some of the information you will obtain as a result will be about yourself: You will know that you made that choice, which might suggest that other people like you would make that choice, that various things in the past may have happened to cause you to make that choice, and so on. Your decisions will correlate with other things in reality, and the choice you will make will tell you about those things. This would seem to suggest that when making a choice you can "choose" how you want to "control" reality in non-causal ways. This is all too much for some people, who insist that advocates of evidential decision theory are confused and that we cannot rationally base a choice on anything except that which we causally control.

In this article, I will be making an argument in support of evidential decision theory.

Two scenarios are often discussed in relation to evidential decision theory. One of these is the *prisoner's dilemma* and the other is *Newcomb's paradox* (or *Newcomb's problem*). These will not be discussed in detail, here, as there is already a lot of literature about them, but I will try to give a brief idea of the sorts of issues they raise.

The prisoner's dilemma involves two suspects being interviewed about a crime in separate rooms. The optimal behavior by each suspect depends on what the other suspect does, creating a clear implication that each suspect should ask if the other suspect is reasoning in a similar way to him/herself.

Douglas Hofstadter has discussed the prisoner's dilemma in terms of *superrationality* (Hofstadter, 1985). A superrational player in a game who views another player as superrational will assume similar behavior by the other player and will make his/her decisions accordingly. The standard version of the prisoner's dilemma requires you to buy into an assumption of full-scale superrationality to take this approach with it. Variations of the prisoner's dilemma have been constructed, however, which only require people to buy into superrationality to a limited degree.

Newcomb's paradox was originally proposed by Simon Newcomb (Nozick, 1969; Kiekeban, 1996). The paradox takes various forms, but always involves a predictor who predicts the decision of the player in a game with some degree of accuracy: The predictor is generally assumed to be "almost completely accurate". The reward that a

player receives is based on the prediction that the predictor made of his/her decision and the paradox is therefore suggesting that the player should make his/her decision as if backward causation is operating – as if the decision that he/she makes can influence that prediction that the predictor made in the past.

Some people claim that Newcomb's paradox requires us to assume a situation that is logically impossible, or inconsistent with physics. Others try to construct scenarios in which the paradox could, in principle, be realized in the real world: These can include a person who is simply very skilled at psychology or mind uploading. The extreme ability of the predictor means that Newcomb's problem is often written off as a contrived situation with no relevance to reality. In this article, I will be arguing that this is not the case: Newcomb's paradox raises issues that are with us in real life and need to be taken into account in decision theory.

To show that evidential decision theory is relevant in the real world, I will start by introducing a scenario, and some variations of it, constructed so as to "push" readers towards accepting evidential decision theory. Unlike Newcomb's paradox, the scenario will not involve any entities with extreme abilities. To the contrary, the scenario will make evidential decision theory an issue by introducing entities with reduced knowledge. As I will explain later, reduction of knowledge is important in creating a situation in which evidential decision theory becomes important. This is more plausible than Newcomb's paradox, as a candidate for the kind of thing that could happen in real life: We should find it easier to imagine situations in which our knowledge is limited than situations in which entities have extreme predictive abilities: It is generally quite easy to end up not knowing things. Following the scenario, I will make various arguments to support evidential decision theory and show that it is the only unbiased way we can approach reality.

This article will show that evidential decision theory has varying degrees of relevance in real-world situations, and that in some situations it could have significant relevance. It will not be necessary to have a previous understanding of evidential decision theory to understand this article: The ideas that I am defending will be fully explained. This article will be part of a two-part series, and in the second article I will explore some of the implications of evidential decision theory, considering where it is not relevant and some of the situations in which it may be relevant: where we should consider taking it into account when making decisions.

This article is intended to argue for evidential decision theory as opposed to causal decision theory. Its target is *causal decision theory*. It is *not* intended as an attack on various approaches to decision theory which have some "evidentialist" attributes, in terms of accepting the relevance of correlation, but which may not always be explicitly referred to as "evidential decision theory" approaches.

# 2 A Scenario Involving Actions and Correlation

To start with, we will consider a scenario in which you are playing a game where you have to choose whether to press a button. The commonsense decision is to press the button; however I will try to persuade you not to press the button, at least in some circumstances. The scenario discussed here has some resemblance to scenarios presented by Bob Wolf (Hofstadter, 1986, pp.752-755).

## 2.1 Description of the Scenario

You were raised in a cave on a post-apocalyptic world by artificial intelligences and robots. You have been taught nothing, or almost nothing, about pre-apocalypse society or the world beyond the cave and you have never seen another human. One day, you are taken into a room containing a desk with a button on it and the door is locked behind you. You are told that you are here to play a game. The game will involve ten other players: These are other humans, whom you have never met, each of which has been raised in a separate cave by AIs and robots, in the same general way that you were. They are *not* clones of you. Each of you is cloned from some different, randomly selected pre-apocalypse human.

You will be given five minutes to press the button on the desk or not press it, a period of time which will be known as the *decision period*. You will then have a five minute *waiting period* to reflect on whether your actions were a good idea. A machine will then execute the following algorithm to decide your fate, and the same applies for the other players.

Step 1: The number of button presses which were made by the other players are added up (not including your own). Each button press counts as a 0.1 contribution to your probability of death in Step 1. Whatever probability results, that is the probability that you are now killed immediately. For example, if three of the other ten players pressed their buttons, 3x0.1=0.3 and you now have a probability of 0.3 of being killed immediately. If all ten of the other players pressed their buttons, you have a probability of 10x0.1=1 of being killed immediately; that is to say, you face certain death.

If you survive Step 1, and you pressed the button, the room is unlocked and you are free to leave. If you survive Step 1 and did *not* press the button then you now face Step 2.

Step 2: You have a probability of $10^{-1,000}$ of being killed immediately. You are therefore being punished for *not* pressing the button with some small, extra risk. This is a clear argument for pressing the button.

If you survive Step 2, the room is unlocked and you are free to leave.

Assume that you do not care about the other humans, and this is purely about self-interest: You just want to minimize your chance of being killed.[1]

Should you press the button?

We now extend the scenario further: If you would choose not to press the button, what if the probability of death in Step 2 were increased in stages, from $10^{-1,000}$ to some other value? At what value would you decide to press the button? With a probability of death of 1 in Step 2, meaning certain death for not pressing the button, everyone would press the button, but what about probabilities between $10^{-1,000}$ and 1?

## 2.2 Considering the Scenario

To some people, the approach to the scenario is obvious: In Step 1, you are exposed to a level of risk that you cannot do anything about: Whether you pressed the button or not is irrelevant, because only the decisions of the other players are deciding your fate. In Step 2, you are exposed to risk if you did not press the button, but you are exposed to no risk if you did press it. Pressing the button, therefore, obviously guarantees your safety in Step 2 and does nothing to increase the danger in Step 1: You should press the button.

Is it really that simple? Suppose that you just pressed the button, to remove the risk of death in Step 2, and imagine that you are now in the five minute waiting period, before you will find out your fate. You are safe from Step 2, because it will not apply to you. You clearly have a lot to worry about, though. The actions of the other ten players could easily have generated a substantial probability of death. You will be hoping that not many of them, and ideally none of them, pressed their buttons.

Something about this should worry you. You have never met other humans. You have little to no knowledge of pre-apocalyptic society. You know almost nothing about how other humans behave. Practically all your experience about how humans behave, and your only real basis for guessing how other people might behave, comes from the one human you know about: you.

*And you just pressed the button.*

Maybe you should be very worried.

The other players are not identical to you, but they are human, and they were raised in the same kind of way as you, so you know they have something in common with you. If you pressed the button, does it not suggest that it is more likely that the other players pressed their buttons too than if you had *not* pressed it? By pressing the button, you

[1] Some people have an issue with this scenario in that it requires you to behave callously. If that is an issue, you may prefer to imagine that something less serious than death happens.

avoided entering Step 2 with its $10^{-1,000}$ probability of death, but was it really worth it? After doing that, maybe you should be much more worried about the implications of what you did, *and what it says about people in general*, for Step 1. You can spend the five minute waiting period hoping that the other ten players were nice, in the full-knowledge that your own behavior suggests that people can be quite nasty, really, and that they are capable of causing serious risk of death to other people, just to avoid a $10^{-1,000}$ probability of death themselves.

Alternatively, if you chose *not* to press the button, you will have to face the $10^{-1,000}$ risk of death in Step 2, assuming you survive Step 1, but that is a small risk, so it should not unduly worry you. Your main worry should be Step 1. Now, however, you do know something a bit more encouraging about humans: The one human you know about, you, has chosen not to press the button, which should suggest that you are in a better situation, with regard to what you should expect the other players to do, than if you had pressed it.

All this suggests that, once you have chosen to press the button or not press it, and you are in the five minute waiting period, you should regard death as being more likely if you pressed the button.

This gives us a strange result: It suggests that *you should not press the button!*

This is such a counter-intuitive result that some people will recoil from it. There is no causal link, as we understand causality, between you and the other players. If this reasoning is valid, it suggests that the best way to act, to maximize your chances of survival, is to act as if you have some degree of control over the actions of the other players. Yet there is clearly no control in the causal sense by which we understand "control". The only thing that would make the other players behave in the same way as you is their common humanity and similar upbringings. If you chose not to press the button, you are exploiting a correlation that you expect to exist between your actions and those of others. This is strange!

One issue here is how strong you should think the expected correlation is between yourself and the other players. This is where the probability of $10^{-1,000}$ in Step 2 comes in. I deliberately selected an ultra-low probability to try to set things up so that anyone who accepted that even a weak degree of correlation would justify some degree of expectation that the other players would behave as you did would view this as good enough reason not to press the button, as the risk being accepted in Step 2 would be extremely small. Some people would still think it worth pressing the button to avoid Step 2 with its $10^{-1,000}$ probability of death, but I think most of those people would be unlikely to be interested in not pressing the button anyway: They would not have bought into any of what I have said about acting as you want the other players to act. If, however, you have accepted that you should take account of the correlation, but the

probability of $10^{-1,000}$ in Step 2 is too high, feel free to reduce it to some value at which you would choose to press the button.

I have extended the scenario by asking people who would choose not to press the button for a low probability in Step 2 how high the probability in Step 2 would need to be for them to press the button: The higher the probability would need to be for people to choose to press the button, the higher the degree of correlation that they think should be taken into account when making the decision. To some people, even $10^{-1,000}$ would be too much, and they would press the button to avoid this risk in Step 2. Others might choose not to press the button even if the probability were as high as 0.001. Some people might still not press the button even if the risk of death in Step 2 were as high as 0.99: To do that you would have to think that you should take account of a *huge* degree of correlation between your actions and those of others when deciding. It would mean that you would need to think that the correlation is so strong that the probability of being killed in Step 1, is decreased by more than 0.99 if you pressed the button – that it goes from at least 0.99 down to 0, or from above 0.99 to almost 0: You are accepting a 0.99 probability of being killed in Step 2, so you need to be saving yourself from an even worse threat in Step 1.

Some readers may want to ask questions about their upbringing in the cave in the scenario. Did they learn anything about pre-apocalypse society? Do the AIs act like humans? Can the AIs themselves give some statistical clues about how other intelligent beings may act? Rather than add details here, I will leave it as saying that you know "nothing or almost nothing" about pre-apocalypse society, the outside world or other humans. We should assume that you have never been allowed to read a psychology book. Beyond that, people may wish to make their own assumptions about the degree of knowledge that you are supposed to have, to produce particular versions of this scenario.

## 2.3 The *Control Group* Version of the Scenario

Some people will have trouble with the idea that you might rationally take into account what you want the other players to do when making your decision. To try to persuade such readers, I will now provide an altered version of the scenario: the "control group" version.

This scenario is like the one just discussed, except there are no longer just 10 other players: There are now 1,010 other players. With regard to how they affect you, these players are divided into two groups.

- One thousand of the other players are in the "control group" with regard to you.
- The ten remaining other players are in the "deciding group" with regard to you.

Each of the other players is in a situation like yours and will be asked to decide whether to press a button. For each of these other players, one thousand of the other players

will be in the control group and ten will be in the deciding group. Each player has his/her own control group and deciding group, with both being selected randomly for each player. You will be in the control group of some of the other players and the deciding group for others: In fact, because a control group is much larger than a deciding group, you will probably be in the control group of most of the other players.

When the decision period has expired for you, it has also expired for the other players. Normally, you would just wait a further five minutes to find out your fate. However, now, something different will happen: As soon as the waiting period has started, you will be told how many of the players in your control group pressed the button: It will be a number from 0 to 1,000. After the waiting period has ended, the following algorithm is executed by the machine.

Step 1: The number of button presses which were made *by the ten players in your deciding group* are added up. Each button press counts as a 0.1 contribution to your probability of death in Step 1. Whatever probability results, that is the probability that you are now killed immediately. For example, if three of the other ten players pressed their buttons, 3x0.1=0.3 and you now have a probability of 0.3 of being killed immediately. If all ten of the players in your deciding group pressed their buttons, you have a probability of 10x0.1=1.0 of being killed immediately; that is to say, you face certain death. (This is the same as Step 1 in the original version, except that it is now specifically the members of your deciding group who determine your fate.)

If you survive Step 1, then what happens next is exactly the same as in the original version of the scenario: If you pressed the button, the room is unlocked and you are free to leave, but if you did *not* press the button then you now face Step 2.

Step 2: You have a probability of $10^{-1,000}$ of being killed immediately. You are therefore being punished for *not* pressing the button with some small, extra risk.

As before, if you survive Step 2, the room is unlocked and you are free to leave.

Suppose that you are in this new scenario, and for now let us disregard whether you press or not. The decision period ends, and the five minute waiting period starts. At the start of the waiting period, you are told how many of your control group pressed the button.

Suppose that only a small number of people in your control group pressed the button. This suggests that the kind of people playing this game tend not to press the button. It is therefore quite likely that nobody in your deciding group pressed the button, and even if anyone did, that not many of them pressed it. You should therefore regard this as indicating that you will probably survive Step 1. You should take this as good news.

Suppose now that almost all of the people in your control group pressed the button. This suggests that the kind of people playing this game tend to press the button. It is

therefore quite likely that most of the people in your deciding group pressed the button, indicating that you will probably not survive Step 1. You should take this as bad news.

So far, this should be fairly uncontroversial. Remember that your control group and deciding group are both selected randomly from the other players, so you have every reason to think that the control group will be statistically representative of the players in general, from whom your deciding group have been selected.

Let us consider things now from the point of view of one of the other players for whom you are a member of the control group. Suppose that you pressed the button. When this player is told how many people in his/her control group pressed the button, your own button press will have increased this number by one. The higher this number, the more likely it is that people in the player's deciding group also tended to press the button, so your button press increases this player's estimate of the probability that he/she is about to be killed. Now suppose that you did not press the button. Now the number of button presses in the control group is not as high as it might have been and the player estimates that he/she is more likely to survive. If you press the button, this is bad news for this other player and if you choose not to press the button it is good news. This is despite the fact that, according to the usual meaning of the word "causality" there is no causal link between your actions and what happens to this other player: As a member of his/her control group your actions are not taken into account when deciding his/her fate. Rather, your actions matter because they give an idea of the behavior of the players in general, from whom the people in this player's deciding group, who *do* determine his/her fate, have been selected.

An important point here is that your actions were not bad news or good news to the other player in any way which was specific to the other player: They were only significant because they gave information about the expected behavior of the players in general, and this is what determines the level of risk for everyone, *including you*. If you pressed the button and your actions told this other player that the other players in general were more likely to press, why should you not view it the same way? Why should you not add yourself to your control group, and count your decision to press or not press as contributing to your expectations of the behavior of everyone else and, ultimately, your chances of survival. It may sound irrational to include yourself in the control group, but we should remember that *most of the other players have you in their control group quite legitimately*. If you should not be in your own control group, why should you be in anyone else's?

For example, suppose you just chose to press the button. You are then told that 257 of the 1,000 people in your control group pressed the button. 25.70% of the control group pressed the button. However, you know you that you pressed it too, so you should add yourself to the control group. The control group now contains 1,001 people, of which 258 pressed the button. 25.77% of the control group pressed the button. Clearly, you should want this percentage to be as low as possible, but your decision to press has

made things worse: When you include yourself in the control group, your decision to press the button means that you should consider it more likely that you are about to die. On the other hand, if you had chosen not to press the button, you now have a control group of 1,001 people, including yourself, of which 257 people pressed the button. Your decision not to press means that the percentage of people who pressed is now only 25.67%. In this case, adding yourself to the control group actually improves your expected situation.

If your estimated change of survival is going to be lower after pressing the button than it would be if you had not pressed it, this suggests that the sensible decision, if you want to survive, is obvious: *Do not press the button*.

This, needless to say, is strange advice. Choosing not to press the button means accepting the extra risk, even if it is only small, in Step 2. To do this, you would have to really believe that not pressing the button improves your chances of surviving Step 1. There is no causal link between your decision and the decisions of your control group, and yet choosing not to press the button on the basis of what has been discussed here means that you are acting as if there is. We might try to get round this by inventing weird physical effects that transmit the effects of your decision to the other players, but this would be flawed: The justification for not pressing the button that has been given here is based not on transmission of any effect from you to the other players, but on correlation: They are expected to act like you, to some degree, because they have something in common. Making a decision in such a way as to exploit a correlation like this is still strange, and this will be given more discussion later.

Trying to invent any physical effect to justify all this would be a pointless distraction, and to make that clear I will now give another version of the scenario. This is like the control group version of the scenario that we have just been discussing, but it now involves a *time delay*.

## 2.4 The *Time-Delayed* Control Group Version of the Scenario

As before, the game involves yourself and 1,010 other players. Each player is placed in a room and has to choose whether or not to press the button. However, the following changes are made.

Ten players have been randomly selected from all the players to be *the deciding group*. They will be the deciding group for all the other players. (They do not need a deciding group themselves, for a reason that will shortly become apparent.) This means that all the other players will have the same deciding group.

Assuming that you are not in the deciding group, the 1,000 players who are not in the deciding group (that is to say, everyone else) are your control group. (If you are in the

deciding group, you do not need a control group yourself, for a reason that, again, will shortly become apparent.)

As before, you are allowed a five minute decision period to decide whether or not to press the button. After this, there will be the usual five minute waiting period, at the start of which you will be told how many people in your control group pressed the button. After the five minute waiting period, you face Step 1, in which your risk of death depends on how many people in the deciding group pressed the button. If you survive Step 1, then you are free to leave if you pressed the button and otherwise face Step 2, as before.

So far, this is just like the previous version of the scenario, except that everyone now has the same deciding group. We now come to another, important difference.

The ten players in the deciding group are taken to their rooms one hour before everyone else. *They are not told that they are in the deciding group*. They do not know that they have been taken to their rooms first. Initially, they are treated just like the other players. The rules of the game are explained to them and they enter the five minute decision period. Their decisions, as members of the deciding group are recorded, as it is these that will determine the risks to all the other players in Step 1. At the end of the decision period, they are immediately killed. There is no chance of escape here: All members of the deciding group are doomed to die at the end of the decision period. They never get told anything about control groups, nor do they go to Step 1 or Step 2. It should now be clear why members of the deciding group do not need a deciding group for themselves: They are going to die anyway. It should also be clear why they do not need a control group: The only purpose of the control group is to allow players to be told how many people in it pressed the button: Nobody in the deciding group needs that information, because they will all be dead by the time it would normally be given.

After the ten people in the deciding group have been put in their rooms, allowed to choose to press or not press, and have been killed, the remaining 1,001 players are taken to their rooms and the game proceeds normally for them. The only difference, now, is that the deciding group is the same for all of them – the ten people who have just been killed – and the control group, for each of them, is all of the other 1,000 players who are still alive. After the decision period, each of the players experiences a five minute waiting period, at the start of which they are told how many players in their control group (everyone still alive except them) pressed the button. They then go to Step 1, with the ten dead players being used as the deciding group, and if they survive, as usual, they are either released from the room if they pressed the button, or they face the extra risk of death in Step 2.

None of these rules is hidden from the players. Everyone knows that if you are in the deciding group you have been taken into the room early and will be killed after your decision is obtained. However, while choosing whether to press or not to press, none of

the players know whether they are in the deciding group. If you are not killed at the end of the decision period, however, you know that you were not in the deciding group.

How does this change things? When you are making your decision, you know that you may be in the deciding group. For this to happen, you would have to be one of the ten unlucky people out of 1,011, and the probability of this is about 0.01. This is going to be a concern, but there is nothing that you can do about it: If you are in the deciding group you will die whatever you do. It is rational, therefore, to assume that you are not in the deciding group when deciding whether to press the button.

After you have made your decision, if you are not in the deciding group, you will now find out this, when you are not killed, but are instead told how many people in the control group (all the other surviving players) pressed the button. As before, the number of people in the control group who pressed the button should be statistical evidence of how players are expected to act in general, and this includes the ten players in the deciding group. The fact they already dead when the control group made their decisions should not matter: This is about correlation, rather than any physical effect being transmitted, and what the control group has just done should give us an idea of what the deciding group *did*. This means that the reasoning with the last variation of the scenario should still hold: You should want to see that as few people as possible pressed the button, and if you add yourself to the control group, you should think yourself more likely to die if you pressed the button. This, again, makes a case for not pressing the button.

What may make this particularly strange, to some readers, is the time-delayed aspect. If you decide not to press the button, you are acting as if you have some causal influence over the decisions of the players in the deciding group, even though these players have decided and are dead before you even entered the room: You are acting as if your decision somehow has causal power that reaches back in time. This should make it particularly clear that there is no “magic causal effect” get-out here. It is all about correlation, rather than causation, which is why it can appear to relate events that are temporally separated like this, yet it is the fact that the scenario seems to be pushing us to treat correlation like causation when acting that may make it seem intuitively wrong.

If you do not buy into the idea of acting as if you influence the decisions of the other players, then you will press the button. This, however, should cause you to experience a conflict. When you are told how many people in the control group pressed the button, you should feel safer if this number is low, yet the number being presented to all the other players takes account of your decision: If you add yourself to the control group, your decision would seem to have increased your risk, but if you do not add yourself to the control group you have the problem of justifying your removal of yourself from the reference class – and before you reach for “free will” as an excuse, whatever that is supposed to be, if it exists at all, we can presume that everyone in the control group has it as well.

## 2.5 Does the size of the control group dilute the importance of your decision?

It may occur to some readers that, in the control group versions of the scenario, your decision to press or not to press only has a small effect on the statistics: When you are told how many people chose to press, it is for a control group of 1,000 people (not including you): Your own decision is a relatively small part of this. It may therefore seem that your decision is not very important.

The problem with this is that knowledge of what the control group did is information, and you did not have this information when you made the decision. Your own decision gives you information about the behavior of people playing the game, and the importance of this information diminishes when you obtain extra information from other sources. When you made the decision, you knew a lot less about how people were likely to behave, and your own decision should have been much more relevant to predictions about the other players. This may seem strange, but we can take this to an extreme. Suppose that, before making your decision, you are told how the deciding group voted: At this point, your own decision is clearly completely irrelevant with regard to future expectations of survival.

One way of looking at this is that when you make your initial decision, this gives you an indication, to some extent, of how people will behave, and this means that it will give an indication of how the control group will behave. Although, when you know the results from the control group your own decision has become less important, your own decision was important with regard to the earlier prediction of those results before it gave up its own importance as more information arrived.

## 2.6 The Importance of What You Already Know

The importance of what you already know when making your decision is the reason that I placed the scenario in the strange, post-apocalyptic environment, with the players raised separately in caves. If you and the other players were assumed to be recruited from people living in a normal, twenty-first-century world, you would already have a lot of knowledge of how humans typically behave before making your decision. This would make your decision less important, and would weaken the argument for not pressing the button. It should be noted that I have set the probability of death in Step 2 to a very low value of $10^{-1,000}$, so there may still be an argument for not pressing the button in such a situation, but I have also asked readers to consider gradually increasing this probability – and, even if you accept the main idea of this article, a situation in which you can be assumed to have a lot of pre-existing knowledge and experience of other humans would seem to be one which should cause you to press the button even when the probability is extremely low, due to the importance of your decision with regard to prediction of everyone else's behavior being extremely low.

It could be argued that, even if you do not have experience of other people, your experience of all your previous decisions and behavior before playing the game will tell you much more about how people will tend to behave in the game than your decision in the game itself. If you knew, a long time before the game, that you would be facing the game one day, this might be viewed as an argument for behaving in particular ways. For this reason, I will now suggest *amnesia* versions of the scenario.

## 2.7 *Amnesia* Versions of the Scenario

Amnesia versions of the scenario are like other versions, except that, as well as being deprived of knowledge about how people behave, you are deprived of knowledge of how you have behaved. In this kind of variation on the scenario, you have no memory of any previous life before a very recent time in the past. You may have only a few weeks, or even hours, worth of memories. It may be that, before this, you were living your life in the cave and you have lost your memories, or maybe you were somehow manufactured recently. People can imagine their own variations here: The main idea is that you are deprived of a lot of the information that you would otherwise have about yourself.

## 2.8 Could your decision increase your estimated risk while actually making you *safer*?

One objection that might be made here is that I am confusing an increase in the estimated probability of death following an action with an actual increase in risk due to the action, whereas it is possible for an action to reduce risk, and be rational, while actually causing someone to increase their assessment of the risk.

An example of such a situation might be as follows.

There is a disease which is likely to kill people who have it, and indeed is guaranteed to kill anyone who gets it and does not receive treatment. It is estimated that 0.01% of the population have the disease at any time. If you have it, taking medication for the disease reduces your risk of death, but you are still in significant danger. The medication can be bought over the counter at a pharmacist. The medication is only effective if treatment starts early, before you are sure you are have the disease. If you start the medication early, you have a probability of 0.6 of survival. If you do not start the medication early, you have a probability of 1 of death. A lot of people therefore go to a pharmacist when they suspect they are showing the first symptoms, so that they can start the medication early, in case they have the disease. A study was done recently, and it was found that 35% of the people who went to the pharmacist saying they suspect they are showing the first symptoms actually turned out to have the disease.

Suppose you are concerned that you might be showing symptoms. As of yet, the concern has not been strong enough to cause you to visit a pharmacist. While walking

past a pharmacist's store, you see a sign in the window, with graphic images, warning people of the consequences of ignoring the first symptoms of the disease. You decide to walk into the pharmacy to buy the medication.

The pharmacist had seen you walking down the street, and at that time he/she had no reason to think that you were any more likely to have the disease than anyone else, and so would have assigned you a probability of having the disease of 0.0001, and a probability of dying from the disease of less than 0.0001. However, when you told the pharmacist that you suspected you may be showing the first symptoms of the disease and asked for the medication, the pharmacist would have assigned you a probability of having the disease of 0.35. If you have the disease and start the medication early, which you are doing, you have a probability of 0.6 of survival, which means a probability of 0.4 of death, so the pharmacist now thinks that your probability of dying from the disease is 0.35x0.4=0.14.

From the pharmacist's point of view, your decision to walk into the store and buy the medication increased your risk of death from less than 0.0001 to 0.14. However, any idea that this meant that walking into the store actually increased your risk of death would be absurd. Getting the medication clearly made it *less* likely that you would die from the disease. Does this not show that an action can be rationally taken, that actually makes you less likely to die, even though your probability of death is higher after performing the action?

There is a fallacy in this objection. Probabilities are subjective: They depend on the state of knowledge of the observer who is assigning them. The pharmacist only changed his/her probability because he/she started with incomplete information. Initially, he/she did not know what you knew – that you suspected that you had the first symptoms of the disease – until you asked for the medication. Once the pharmacist knew that you suspected you had the first symptoms of the disease, the pharmacist and you both had the same knowledge, and should assign about the same probability to your survival. None of this involves a person rationally performing an act even though he/she knows that the probability that he/she assigns to his/her survival will be reduced if the act is performed because the probability being reduced is not assigned by the person performing the act: It is assigned by the pharmacist.

On the other hand, just before you made your decision to go into the pharmacist's store and buy the medicine, *you* would know that you were in the dangerous situation of showing what resembles the first symptoms of the disease: You should have already reduced your probability of survival accordingly.

## 2.9 Expected *correlation* has a role in decision-making.

The intention of the above scenario, for now, is to try to persuade you not to press the button provided that the probability in Step 2 is low enough, thereby accepting that expected correlation, rather than just expected causation, has some role in decision-making. Later in this article, I will be discussing how we can view actions in terms of correlation and how causation fits in with this. The message I have been trying to communicate, with the above scenario, is:

*Why, this is your reference class, nor are you out of it.*

# 3 Arguing the Case Further

## 3.1 Decisions and Previous States of Reality

### 3.1.1 Determinism implies apparent backward causality for *all* decisions.

Both the time-delayed control group version of the scenario that I presented and Newcomb's paradox suggest that you should act as if backward causality occurs: that your actions cause something to happen in the past. This may seem nonsensical. It should be remembered that we are talking here about correlation rather than causality; nevertheless, it is still suggested that you should make decisions as if you can "choose" a previous state of reality. This is actually not all that extreme. In fact, it is a feature of normal, everyday decision-making, as I will now explain.

Suppose the universe is deterministic, so that the state of the universe at any time completely determines its state at some later time.

Suppose at the present time, just before time $t_{now}$, you have a choice to make. There is a cup of coffee on a table in front of you and have to decide whether to drink it.

Before you decide, let us consider the state of the universe at some time, $t_{sooner}$, which is earlier than the present. The state of the universe at $t_{sooner}$ should have been one from which your later decision, whatever it is going to be, can be determined: If you eventually end up drinking the coffee at $t_{now}$, this should be implied by the universe at $t_{sooner}$.

Assume we do not know whether you are going to drink the coffee. We do not know whether the state of the universe at $t_{sooner}$ was one that led to you drinking the coffee. Suppose that there were a number of conceivable states of the universe at $t_{sooner,}$ each consistent with what you know in the present, which implied futures in which you drink the coffee at $t_{now}$. Let us call these states $D_1,D_2,D_3,\ldots D_n$. Suppose also that there were a number of conceivable states of the universe at $t_{sooner}$, each consistent with what you know in the present, which implied futures in which you do *not* drink the coffee at $t_{now}$. Let us call these states $N_1,N_2,N_3,\ldots N_n$.

Suppose that you just drunk the coffee at $t_{now}$. You would now know that the state of the universe at $t_{sooner}$ was one of the states $D_1,D_2,D_3,\ldots D_n$. Suppose now that you did *not* drink the coffee at $t_{now}$. You would now know that the state of the universe at $t_{sooner}$ was one of the states $N_1,N_2,N_3,\ldots N_n$.

Consider now the situation in the present, just before $t_{now}$, when you are faced with deciding whether to drink the coffee. If you choose to drink the coffee then at $t_{sooner}$ the universe will have been in one of the states $D_1$, $D_2$, $D_3,\ldots D_n$ and if you choose *not* to drink the coffee then at $t_{sooner}$ the universe will have been in one of the states $N_1,N_2,N_3,\ldots N_n$.

*From your perspective*, your choice is determining the previous state of the universe, as if backward causality were operating. From your perspective, when you are faced with choosing whether or not to drink the coffee, you are able to choose whether you want to live in a universe which was in one of the states $D_1,D_2,D_3,\ldots D_n$ or one of the states $N_1,N_2,N_3,\ldots N_n$ in the past. Of course, there is no magical backward causality effect operating here: The reality is that it is your decision which is being determined by the earlier state of the universe. However, this does nothing to change how things appear from your perspective.

Why is it that Newcomb's paradox worries people so much, while the same issue arising with everyday decisions does not seem to cause the same concern? The main reason is probably that the issue is less obvious outside the scope of contrived situations like that in Newcomb's paradox. With the example I have been discussing here, you get to choose the state of the universe in the past, but only in very general terms: You know that you can choose to live in a universe that, in the past, was in one of the states $D_1,D_2,D_3,\ldots D_n$, but you are not confronted with specific details about one of these states, such as knowing that the universe had a specific state in which some money was placed in a certain box (which is how the backward causality seems to operate in Newcomb's paradox). It may make it seem more like an abstract, philosophical issue than a real problem. In reality, the lack of specific knowledge should not make us feel any better: In both situations you seem to be choosing the past as well as the future.

You might say that you do not *really* get to choose the previous state of the universe, because it was in fact your decision that was determined by the previous state, but you could as well say the same about your decision to drink or not drink the coffee: You could say that whether you drink the coffee was determined by some earlier state of the universe, so you have only the *appearance* of a choice. When making choices we act as if we can *decide*, and this issue of the past being apparently dependent on our choices is no different from the normal consequences of our future being apparently dependent on our choices, even though our choices are themselves dependent on other things: We can act as if we choose it.

### 3.1.2 Non-determinism does not make the issue go away.

In what was just said, it was assumed that *determinism* applies, in that the state of the universe at any time allows its state at some point in the future to be predicted. Some people may think that *non-determinism* provides a rebuttal of this. In response, I could try arguing that we could force determinism to be true, in a tautological kind of way, by saying that at any time, there is a future that is going to happen, whether it is possible to access variables indicating what it is going to be or not, and that the information about this future can be considered part of the state of the present universe: just an inaccessible part of it. A lot of people would not be persuaded by this, and it would make the apparent "backward causality" seem a bit weak: It would just amount to being able to select variables in some previous state of the universe that indicate its future

and are otherwise inaccessible. Some controversial scientific models suggest that there are multiple futures. For example, the many-worlds interpretation of quantum mechanics states that decoherence is causing "splitting" of worlds (Everett, 1957). This would make a deterministic view, applied at the level of our "world" untenable.[2] We could get into a long discussion here about the coherence of non-determinism as an idea. Rather than do that, I will not assume that the future state of reality can be predicted with certainty from variables describing its current state: In other words, I will assume that something that many people would call "non-determinism" applies, and show that this resolves nothing.

Suppose the choice of whether to drink the coffee at $t_{now}$ faces you. Let us consider the state of the universe at some earlier time, $t_{sooner}$. We can no longer say that there are two sets of conceivable, earlier states of the universe: $D_1,D_2,D_3,\ldots D_n$ which lead to you drinking the coffee and $N_1,N_2,N_3,\ldots N_n$ which lead to you not drinking the coffee. For any state of the universe, there is no single future which is implied by that state.

The fact that no single future is implied by any state of the universe does not, however, mean that things are completely unpredictable. The universe's behavior can still be described *statistically*. This is the least that is needed for scientific models to work, for any regularity to exist in nature – even for our own brains to exist and function. Suppose we consider every conceivable, previous state of the universe $S_1,S_2,S_3,\ldots S_n$ at $t_{sooner}$ which is consistent with what we know just before $t_{now}$. For each such state of the universe at $t_{sooner}$ there will be a probability that you drink the coffee at $T_{now}$ and a probability that you do not. For some states of the universe at $t_{sooner}$ it will be very likely that you will drink the coffee at $t_{now}$, while for others it will be less likely. From your perspective just before making the decision at $t_{now}$, each conceivable, previous state of the universe will have a probability of being the actual state of the universe at $t_{sooner}$, so there will be a probability distribution for $S_1,S_2,S_3,\ldots S_n$.

Now, suppose you have just drunk the coffee at $t_{now}$. You consider the list of conceivable, previous states, $S_1,S_2,S_3,\ldots S_n$. In principle, any of these could have been the actual state of the universe at $t_{sooner}$, but the fact that you drank the coffee will suggest, statistically, that it was a state in which it was likely that you were going to drink the coffee: For each previous state in which coffee drinking at $t_{now}$ was likely, it will increase the probability that the universe was actually in that state, and for each previous state in which coffee drinking at $t_{now}$ was *unlikely*, it will *decrease* the probability that the universe was actually in that state. There is a similar result if you choose not to drink the coffee. Before you make your decision at $t_{now}$, you will know that there will be one probability distribution for the conceivable, previous states of the universe at $t_{sooner}$ if you choose to drink the coffee and a different probability distribution for them if you

[2] It should be noted, however, that the many-worlds interpretation *would* become deterministic once the entire system was considered.

choose *not* to drink the coffee. This means that from your perspective, just before you make the decision, you are getting to choose the probability distribution for previous states of the universe. This is still *apparent* backward causation: What happened in the deterministic case is still with us: It merely turned statistical.

*Ultimately, there is no escape from your reference class.*

# 3.2 The Edges of Causality

## 3.2.1 Using Light Cones for a Simplified View of Causality

I want to show that we should think of the consequences of our actions in terms of correlation, and that causation should be viewed as a special case of this, by looking at the boundary between causation and correlation: the point where causation ends and correlation is needed. What I want to show, here, is that this boundary is not very significant, that there is therefore nothing philosophically profound that happens when causation becomes correlation, and that causation is just a special case of correlation.

We need a way of distinguishing between events that can be caused by your decision and events that cannot be caused by it. In practice, the extent to which causation applies involves lots of complication and specific details about how the world is arranged. For example, if I turn on a torch it can cause light to reach an object that is in front of me, but not if the object is in a box that the light cannot penetrate. We can imagine all kinds of specificity that relates to causality like this. We need a simpler way of distinguishing between events that can be caused by a decision and events that cannot. I am going to simplify here by using the idea of *light cones*, as used in Einstein's relativity (Salgado, 1996). There will not be any of the mathematics associated with relativity being used here, however: I will just be using the basic idea of light cones, and how they describe the limits of causality, as a philosophical device.

In relativity, space and time are considered as a single entity, and any event is at a point in space-time. The speed of light imposes a limit on causality, because one event, occurring at some position and time, can only causally influence another if the path of a beam of light in space-time can connect them. A light cone is the region of space-time containing all the points in space-time which can be reached by a light beam originating at a particular point in space-time; that is to say, it contains all the events which can be causally influenced by a particular event. The speed of light determines where the edges of the light cone are. A light cone is projected forward in time, "ahead" of an event, and indicates what it can causally influence, but we can also project a light cone backwards to indicate what events can causally influence a particular event.

## 3.2.2 Light Cones, Causality and Correlation

Consider an event, $D_1$, which is a decision occurring at some point in space-time. (In simple terms, for a decision made by a human, it occurs wherever your brain is when

you make it, and at whatever time you make it.) We can project a light cone into the future of $D_1$, which contains all the events that $D_1$ can causally influence. Event $E_1$ is in the light cone and can be causally influenced by $D_1$. Event $E_2$ is outside the light cone and cannot be causally influenced by $D_1$. If we were to use a conventional, "causal" view of decision-making, here, we would say that it is possible that your decision could involve "choosing" how you want $E_1$ to occur, but could not involve choosing how you want $E_2$ to occur: Einstein's relativity prohibits any causal connection between $D_1$ and $E_2$. (See Figure 1: Two Events and a Decision, below.)

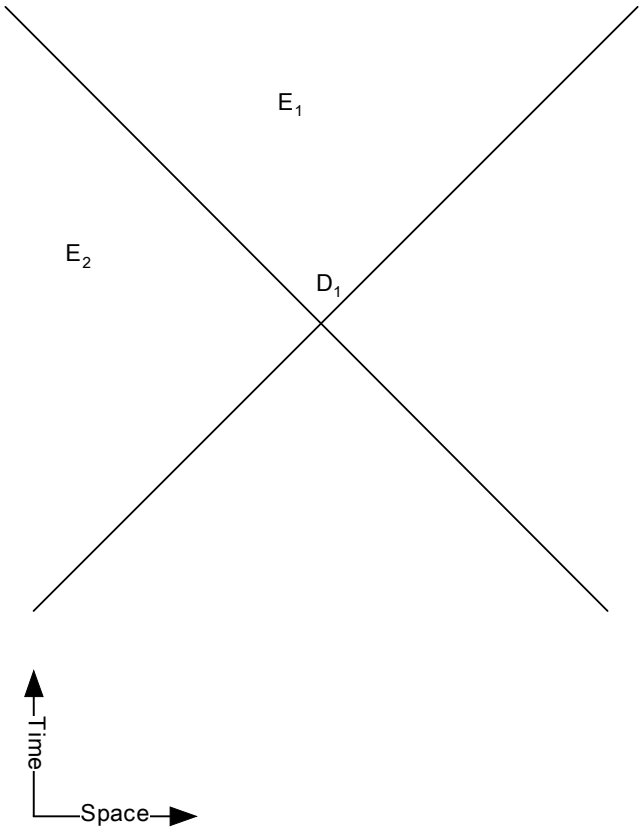


**Figure 1: Two Events and a Decision**

Consider, now, an event, $E_3$, which is just on the edge of the light cone. It can just about be causally influenced by $D_1$. $E_3$ is on the "edge of causality" as far as you are concerned, and we might think that any event just marginally further over, by being outside the light cone of your decision, cannot possibly be influenced by it. (See Figure 2: Two Events in Close Proximity, on page 25.)

Consider, now an event, $E_4$, which occurs at *almost* the same point in space-time as $E_3$, but is *just outside* the light cone. $E_4$ cannot be causally influenced by $D_1$: A signal does not have time to reach it. However, $E_4$ can be causally influenced by another event, $D_2$, occurring in your past, just before $D_1$.

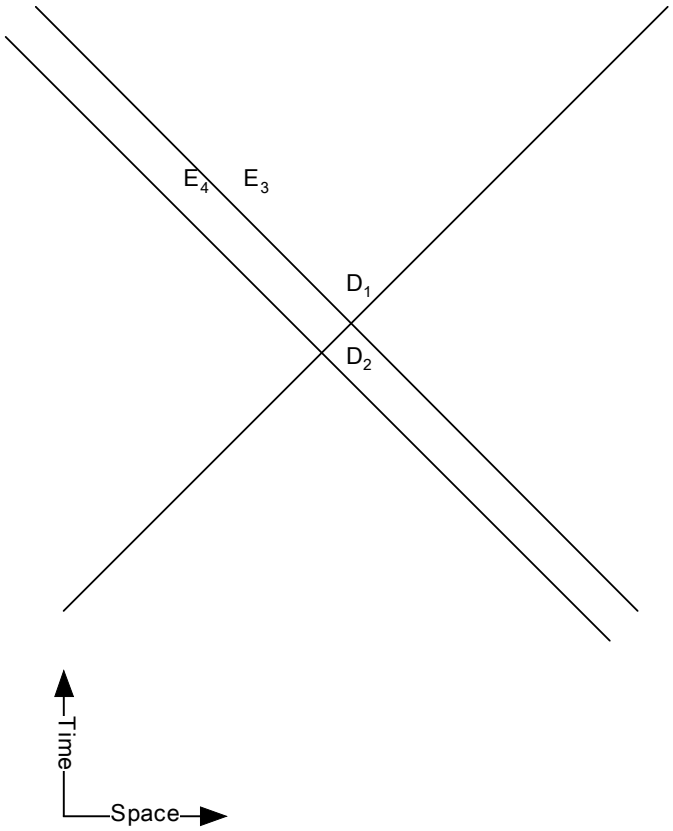


**Figure 2: Two Events in Close Proximity**

The fact that $E_4$ can be causally influenced by $D_2$ may not seem to be much use when making your decision, $D_1$. However, suppose $E_4$ and $E_3$ are really close, so that $D_2$ occurs only *very slightly* before $D_1$. Let us say, for the sake of argument, that $D_2$ occurs only a *fraction of a nanosecond* before $D_1$. This time difference is inconsequential. The state of your brain when $D_2$ occurs is not going to be noticeably different than when $D_1$ occurs. In fact, if $D_1$ and $D_2$ are sufficiently close together in time, the uncertainty principle will make it meaningless to talk about differences in your brain state at the time of each event. For all practical purposes, $D_1$ is influenced causally by your decision, and $E_4$ can be influenced causally by an event that is *practically indistinguishable* from your decision. Now, you could argue here that your decision does not have any influence on "the outside world" until you do something – which only happens when $D_1$ has occurred, but the argument here is much more basic than that: We are only looking at the edges of causality in principle – and *in principle*, a causal effect can be transmitted from $D_2$ to $E_4$.

This argument becomes stronger if we recognize that it is dubious to think of a "decision" as occurring at a single instant in time. A decision is a neurological process – or in some other system it is a computational process, probably distributed over a system. It may be impossible to say, objectively, *exactly* when a decision has occurred, but rather we may need to view a decision as a process which occurs in a system over

some period of time. $D_1$, rather than being at a point in space-time, may need to be viewed as occupying a duration of time in a "fuzzy" way. With this in mind, it is even harder to justify viewing $E_3$ and $E_4$ as significantly different: They could be considered to be in the "region" of the same, vaguely defined decision. All we could say, really, is that $E_3$ and $E_4$ both correlate highly with $D_1$, and $E_4$ would correlate slightly less than $E_3$: The difference in correlation would not be noticeable and you would be entirely justified in making the decision $D_1$ as if it influences $E_4$ just as much as $E_3$, even though $E_4$ is outside $D_1$'s light cone: The fact that you are "operating" outside the light cone of $D_1$ means that you are now reasoning with correlation rather than causation.

We might consider an event, $E_5$, very slightly further outside the light cone than $E_4$. While it could not be causally influenced by $D_1$ or $D_2$, it could be causally influenced by an event $D_3$, very slightly further in the past than $D_2$. $D_3$ will not be exactly the same as $D_2$, and it will be less like $D_1$, but there will still be high correlation. See Figure 3: Further Events, below.)

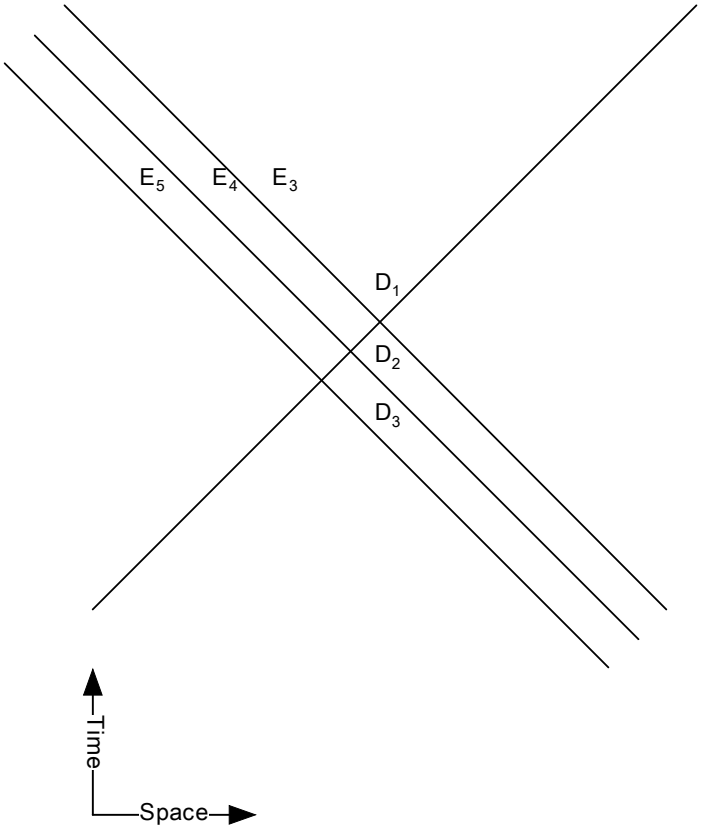


**Figure 3: Further Events**

"Going outside the light cone", in terms of associating a decision with how we expect some event to occur elsewhere, presents no obvious problems: There is no sudden line at which logic tells us that we have to stop. It should be clear that there is no

*fundamental* limit on how far outside the light cone we can go like this. We might say that we need to stop when the event in our past that is causally connected to the event being influenced is "sufficiently different" from the event of the decision itself, but how different would it have to be? There is no rule for us, here. Ultimately, it would have to come down to our ability to predict the correlation. If we are making a decision, $D_1$, and some event, $D_3$, has occurred in our past that correlates sufficiently with $D_1$ such that the decision we make for $D_1$ tells us enough about $D_3$, then we are entirely justified in viewing this in terms of our decision $D_1$ telling us about $D_3$, and therefore about some event that is causally influenced by $D_3$.

Further, this correlation does not have to be extended straight back into the past. Starting with our initial decision $D_1$, we can easily add more events that are expected to correlate with it, in the "backwards" light cone extended back in time from $D_1$, because each event tells us something about previous events. In principle, any event in this backwards light cone might be considered to correlate with $D_1$, and therefore be taken into account when making decision $D_1$.

What I am trying to point out, with all this, is that "messing around on the edge of the light cone" shows that there is no abrupt point where causation becomes inoperable and your decisions become irrelevant. Instead, as the direct path of causation between your decision and an event becomes unavailable, you can, at the start, just ignore the issue and assume normal causation, for events not so far outside your light cone that they are noticeably outside the causal "scope" of your decision. Going further outside your light cone, an indirect path, involving correlation and causation becomes progressively more relevant, and there is no obvious point at which it becomes invalid to use it. In fact, in this kind of view, the only role that causation plays in our decision-making processes is as a very specific method that we can use in correlating past and future events.

Once we have accepted that we can reason in terms of correlation like this, we should not consider ourselves just restricted according to what our decision tells us about our own past. Once we have done that, it should be comparatively painless to start making decisions in terms of more general correlation involving reference classes of agents like ourselves, as well: The main issue should have been getting into reasoning with correlation in the first place.

## 3.3 Seeing Your Decision from Outside and the Copernican Principle

We should be able to see how to approach our own decisions if we imagine viewing them from the perspective of an *external observer*. Let us imagine that you have just made a decision, and it is being viewed from the perspective of some other being. We might imagine an alien who is here to research humans, or a super-powerful AI that

wishes to understand its makers. It does not really matter who or what this external observer is. The only real requirement is that *the external observer is not you*.

The external observer would feel quite entitled to use your decision to infer things about reality beyond what would be implied by assuming "normal" forward causation from your decision. Your decision would *contribute* to statistical information about events sharing well-defined reference classes with it, which for practical purposes would mean the expected behavior of beings similar to you. This should not seem unusual to us at all as we already do it: A human naturalist would see nothing strange in making inferences about the behavior of a species of animals by observing *some* of them. The external observer may also make inferences about events in your past, from observing you make decisions. Again, this is nothing unusual: We observe events happening and make inferences about the causal path leading to those events all the time.

Some people may get confused about what is meant by the external observer "not being you" and think that you would be in a privileged position: that you would have access to your memories and private thoughts to which the external observer would have no access, while the external observer would be able to observe things that you could not. This is not about "knowledge", however. The external observer might be assumed to know exactly you know. The only difference in this case would be that he/she does not gain this information by "being you", but by observing all the parts of the reality, including your mind, which you can observe.

When the external has observed that you have made a decision, or acted on a decision, as far as he/she is concerned this is just another event in the world that he/she happens to observe. Why not use that to infer things about the past or to contribute to information about reference classes of similar events? In fact, it would be strange to make a special case for you. This would be like a naturalist who wants to get statistical information about the nocturnal hunting behavior of cats, who leaves one cat out of the statistics purely because that cat is the focus of a lot of attention and he/she happens to know rather more about that cat than others. The external observer has no reason to think that you are special.

When you are about to make a decision, the external observer knows that your choice will provide information to him/her about reality. Causally extending patterns into the past or future may suggest something about the past which has led to a particular choice being made, or the expected future, or it may provide information about the behavior other entities that share a reference class with you. In advance of you making a decision, the external observer can imagine each possible choice that you could make and work out what information this would provide him/her about reality. The external observer then knows the description of reality that he/she would have for each choice that you could make – and these descriptions could have information in them that has come from causally extending patterns back in time, or using reference classes of other beings making decisions: all the things which you are not supposed to be able to

"control" when taking a non-evidential view of decision theory. The external observer, however, does not have to accept an evidential view of decision theory to do this, however, because *he/she is not you*. The external observer is not applying any of this to his/her decisions, but is simply dealing with another being making decisions. In fact, to generalize further, in observing your decisions, the external observer is merely observing a thing doing things, and treating as any other thing doing things would be treated.

If a description of reality produced by the external observer is valid then the same description should be valid if produced by you from the same information. If it is valid for the external observer to operate like this, then it should be valid for you to do it. If that were not the case it would mean that a description of reality which could be justified if produced by the external observer would be considered unjustified if produced by you from the same information. When faced with a decision, you should be able to think about the possible choices and, for each, work out what information about reality you would be able to obtain from your knowledge that you made that choice. Some of that information could be obtained by projecting causal patterns back in time, or by considering reference classes of which your decision is a member, and this means that the descriptions of reality which you would have after making each choice could vary with respect to such things. Knowing this, before you make the decision, you could choose which description of reality you prefer, and you would be effectively choosing the description of reality that you wanted in a way that takes account of backward causality and reference class. You would now be taking an evidential approach to decision theory. However, you are only doing what the external observer is doing without assuming an evidential approach at all!

This fact that what we would describe as an evidential approach, from the point of view of someone else, is merely consistency in not giving any events special status because they happen to be your decisions, should make it clear that an evidential approach is merely treating all of reality – including what you know about events corresponding to your own decisions – in the same way: It gets included naturally in any generalized process that infers a description of reality from partial information about it. You can project causal patterns both forwards and backwards in time from events in general, and you can use information about events in general to inform you about reference classes to which they belong, and this tells you how the world should be expected to behave. You may, of course, expect the world to behave differently in the vicinity of your own decisions, but you are then making them special events: You are treating your place in reality as special and violating the Copernican principle. This will appear particularly strange when you try telling the external observer that he/she needs to treat observations of your decisions as special as well.

## 3.4 How do you stop your choice contaminating your model?

I have just been arguing that avoiding an evidential approach to decision theory puts us in the situation of treating our own decisions as special-case events, with the information we have about them not being used in the same way to obtain a description of the world; however it is even worse than that. Once our decision has been causally in "contact" with anything else, are we supposed to disregard that as well? This has the potential to put us into an absurd situation, where we are supposed to ignore the statistical information we have about a large amount of reality, as this "tainting" of statistics by our decision propagates outwards.

Here is an example.

Suppose you are locked in a room and you will remain there for a few months. We will not go into too much detail here. Let us not worry about whether you have been raised in a cave, etc: The degree of correlation is not an issue. A pet smurf has been placed in the room with you, and you have a five minute period to decide whether to kill it. If you kill your smurf, someone will come and remove the body from the room. If you do not kill your smurf, someone will bring food for it every day. Other people are in similar situations, in different rooms. In fact, this is happening regularly.

Before deciding whether to kill your smurf, you might think that your decision counts as information about other people in this situation. This would mean that your choice would lead to one of two situations with different expectations about the behavior of other people, and we are in the familiar situation in which your choice seems to provide this unusual kind of "control". Suppose, however, that you reject this kind of evidentialist approach: Your own decision is to be placed off-limits when it comes to working out what everyone else is doing. Let us now suppose that you decide to kill your smurf.

The problem now is that your decision causes other things. Here is one example.

The people who deliver the smurf-food to the rooms do not have to deliver any to your smurf, so they stop. Other people may kill their smurfs too, and all this will mean a reduction in the need to deliver smurf-food to the rooms. This will all affect the total demand for smurf-food. The factory where the smurf-food is produced may have a reduction in its orders, as food is not needed for your smurf. If a lot of people are killing their smurfs, this may be noticeable, and may cause management at the smurf-food factory to reduce their staff. If the local community is economically dependent on the smurf-food factory, this could cause local businesses to experience a decrease in custom. "Business just hasn't been the same," a local shopkeeper might say, sadly, "since those people in the smurf rooms started killing their smurfs. The smurf-food

factory hasn't been doing so well, and people don't spend money around here anymore."

Now, this shopkeeper is being influenced, to some degree, by your decision to kill your smurf, yet this happens through a chain of causal effect. A statistician who meets some shopkeepers like that could correlate this with some expected level of smurf-killing, and that would relate to people in the smurf rooms in general. Unhappy shopkeepers should suggest lots of dead smurfs. In fact, by plotting graphs of smurf deaths and shopkeeper happiness over some years, you might see a pattern.

So, why should you leave this out of your description of the future? Prior to making your decision, you might predict that your action, ignoring those of all the other people, will tend to reduce the happiness of shopkeepers around the smurf-food factory a bit, and therefore that shopkeeper happiness will tend to be lower if you kill your smurf. However, the correlation between shopkeeper unhappiness and smurf-killing should suggest that this means more smurfs are being killed in general. How are you supposed to deal with this? It is not as if you can just decide not to correlate your choice with anything else: We are *not* correlating your choice directly with anything, here, but instead are looking at something which is affected, indirectly and causally, by your choice. It is not as if you can reasonably say that allowing for looking at the correlation between shopkeeper happiness and the actions of people like yourself in the smurf rooms is invalid: You might be using all kinds of other correlations to work out what reality is expected to be like after your decision. Why should this case be special? You might correlate the shopkeeper's happiness with many other things, quite validly – and you might have already been correlating it with what happens in the smurf rooms before you ever went in one yourself. You might try saying that your decision should be left out of the picture? How? Do you ignore it causally, so that you do not allow your decision even to affect the shopkeeper's happiness? If you do that you cannot even take into account any consequences at all of your decision on reality. Do you ignore your decision in terms of correlation? This also gets absurd. It would mean that, when correlating the shopkeeper's happiness with anything else, you would need to edit your own effects on his happiness out of the picture. Are you going to do that with *every* piece of reality that is causally affected by your decision? Does it just apply when the correlation is coming back to the smurf rooms? What if the shopkeeper's happiness correlates with something else, which then correlates with events in the smurf rooms? Do we just intervene when the correlation starts to get too close?

If you are trying to predict what reality will be like after your decision, assuming you make some particular choice, it is going to be hard to stop your choice affecting lots of things, and ultimately coming back, by correlation, to start telling you things about the behavior of other people around you. The only way to prevent this would be to start editing your description of reality all over the place, to keep this evil correlation effect away from things you do not think it should be touching.

This should answer the question that I used in the heading: “How do you stop your choice contaminating your model?” You don’t.

## 3.5 Degrees of Similarity

Even people who reject evidential decision theory when there is supposed to be some degree of correlation should stop and consider when the behavior of the agents is not merely supposed to be similar, but is *absolutely identical*.

One way of making agents identical is to say that they are identical computer programs, starting in identical states, such that whatever one program does will be done by all of them. Suppose you were in a scenario like the one at the start of this article, but with your mind having previously been copied into a computer program, and you were playing some game involving identical copies of yourself. You could be sure that the correlation would be total, and you would have a firm reason for thinking this. If you decided to press the button, you would know that the other players would press the button too. The argument for *not* pressing the button should now be powerful enough to persuade anyone.

People who may be persuaded by this might say that it is irrelevant if there is any difference at all between the players: that as soon as the similarity becomes anything less than total, the players are acting independently and the whole idea of taking correlation into account breaks down. I will try to show, now, that this is not the case.

Suppose you worked in a rather sinister organization that sets up various games, some of which are a bit like the one at the start of the scenario. The difference, now, is that the players, instead of being raised in caves by artificial intelligences, are produced by scanning the brains of random people into computers. Your work involves observing these games. You have observed many games and have a lot of experience about what happens.

Some of the games you have observed have involved identical players. You have seen how the behavior of one player in these games is also mirrored by the others. You have also observed many games in which the players are *almost completely identical*. In these games, a tiny change is introduced into each computer program so that their behavior will be slightly different. They still start, however, in almost exactly the same state. Initially, their behavior is the same, but as time passes their behavior diverges as the program states become progressively less like each other. You have experience of how this happens. You know that for certain kinds of set-up, for example, the players have a 0.99 probability of making the same moves in a game for the first five minutes, as it is very hard to tell then apart, but after this the divergence becomes more extreme and the probability falls below 0.99, with the correlation essentially becoming irrelevant in the end. Now, ignoring how a player should think when in this game, none of this should cause problems when we look at it from the point of view of an outside observer: We

are merely talking about how almost identical systems will tend to behave the same and will gradually diverge.

Suppose you wake up one day and are surprised to find yourself in the game. Let us say that while you slept, another employee, who has a dispute with you, has scanned your brain into a computer program to play games with it. You are a now a computer simulated player, in one of the games you have been observing. You recognize the game. You are also shown the set-up for the game, with the parameters controlling the degree of initial difference between the copies, etc. It is not uncommon for players in the game to be given this information. You have a decision to make in a game with some similarity with the one at the start of the article. From your experience of observing other players in this kind of situation, you know that they tend to make the same decisions 99.999% of the time, as the divergence is very low during the game for this set-up. In this situation, you are not dealing with total correlation, yet you have enough experience to tell you that it would be stupid to ignore it. The situation is hardly different from one in which the players are identical, and the fact that you are now in the game, rather than observing it from outside, should not cause you to expect things to be dramatically different in your case, unless you think you are a *special* player. Suppose now the set-up were changed to one in which you have previously observed the players to make the same decisions 99.99% of the time. What about 99%? 80%? 60%? It should be apparent, here, that there is no point at which you can abruptly say that correlation should be disregarded completely. Instead, the importance that it should play in your decisions gradually diminishes. Initially, the correlation is very high, and your own decisions are telling you a lot about how other players are expected to behave. As the correlation decreases, your own decisions tell you less about what other players are expected to do, and your knowledge about how other people behave in general is more important.

What this should show is that you cannot escape from this issue just because other people are not exactly the same: *Any* degree of similarity will create *some* degree of correlation.

## 3.6 The Fallacy of Trying to Separate the Decision Process from the Model

Some people will object to what I have said by saying that I have misunderstood causal decision theory, and that it is not about excluding certain things from the description of reality, or prejudicing it in a certain way, but that it is not based directly on the description of reality at all: that it does not involve looking directly at the model after the decision to determine the expected utility, but instead merely involves taking the description of reality *before* the decision and then using a process of forward causation to determine the consequences of the decision.

My response to this would be that this is merely another way of describing things. The person, computer program or other agent may be using some explicit process that "does the causality separately" but that does not change the fact that just looking at forward causation like this, and nothing else, is equivalent to updating the description of reality in a selective way. It amounts to looking at the utility after selective updating.

## 3.7 The Fallacy of Thinking that Evidential Decision Theory is Just About "Trying to Get Good News"

Some people may confuse evidential decision theory with managing the news to try to get good news. The idea of people doing this in reality is sometimes considered amusing. For example, in an episode of the cartoon, *The Simpsons*, after the town escapes destruction from a comet, one of the characters states that they are going to burn down the observatory so that an incident like this will never happen again: The character is confusing the idea of preventing the observatory from telling people that a comet is on the way with the idea of *actually* preventing a comet from being on the way.[3] We can also imagine a dictator executing anyone who brings him bad news, so that eventually people only bring him good news.

This kind of idea may be partly responsible for some of the intuitive issues that people have with evidential decision theory. If you are going to make a decision on the basis of what looks better afterwards, without necessarily having a specific causal chain to justify it, are you not simply in the business of managing the news that you receive?

This kind of objection, however, would be a fallacy. There is, in fact, nothing wrong with making the choice that causes you to receive good news. In fact, what else are you supposed to do? What point to making a choice can there ever be other than that things seem to be better when you make that choice? Even when you make your choice on causal grounds alone, you are effectively doing the same thing: You are *still* making the choice that causes you to receive good news, but restricting yourself to the news that is delivered as a result of the causal consequences of your decision, and you are therefore still managing the news, but just restricting the ways in which it is produced. People probably see "trying to get good news" as a problem when they confuse it with the kinds of situations I just mentioned in which people deliberately damage the process which generates the news to make it less accurate; for example, by burning down an observatory or executing messengers with bad news. Acts like this are irrational because they cause you to receive good news by biasing the way in which you receive information: by making your news less accurate. Evidential decision theory does not ask us to do this: Taking into account what your decision tells you about yourself is not biasing the news one way of another, but it is merely taking into account information that any external observer could use quite validly. The idea that we are "managing the

[3] *Bart's Comet*. Aired on February 5, 1995. Written by John Swartzwelder and directed by Bob Anderson.

news" would only become a reasonable objection if the process by which the description of reality was generated were compromised in some way.

# 4 How We Should Decide

## 4.1 The General Idea of Decision-Making

What has been said so far should answer the question of how we should approach decisions, at least in general terms.

When faced with a decision, you should approach the outcome of that decision – what you decide – as *giving you knowledge about reality* in exactly the same way that an external observer, watching you make the decision, would approach the matter. You should consider *what you will know about reality* after each choice that you can make, and then make the choice for which *what you will know about reality,* having made that choice, corresponds to *the most desirable reality*.

Your decision is an event, and any event, E, can provide information, in principle, in the following ways.

- **Forward causal structure:** Causal relationships informing you about events in the future light cone of E.
- **Backward causal structure:** Causal relationships informing you about events in the past light cone of E.
- **Reference classes:** Reference classes informing you about events in reference classes of which E is a member.

To make decisions properly, you must be modeling reality, and you will already be doing something like this for numerous events in space-time. When it comes to dealing with your own decisions, nothing really changes. You should just regard your decision as another event. For each possible outcome of your decision – each possible choice you could make – you should assume that your decision has occurred with that outcome – that you have made that choice – and then take into account what you know about the decision event as a result, before repeatedly using forward causal structure, backward causal structure and reference classes to tell you about further events. This will give you a description of reality for each choice that you could make. You should make the decision according to the outcome that seems to correlate with the most desirable reality.

## 4.2 The Three Ways of Obtaining Information from an Event

I have just described a process of decision-making in terms of forward causal structure, backward causal structure and reference classes and I will give an explanation of each of these.

### 4.2.1 Forward Causal Structure

This involves using what we know about causal relationships, about how reality is causally structured, to extrapolate from an event into the future, to obtain information about future events.

### 4.2.2 Backward Causal Structure

What we know about causal relationships, about how reality is causally structured, can also be used to tell us about the past of an event. This involves extrapolating causal relationships back into the past. It should be noted that this does not mean assuming some kind of magical "backward causation". It merely means that what we know about an event, together with our knowledge of how events tend to be arranged in space-time, can be used to tell us about previous events.

I should point out, here, that this kind of backward extrapolation is not what we would really relate to the apparent backward causation in the time-delayed control group variation of the scenario that I gave: In the scenario that I presented, correlation involving reference classes, as discussed below, is the important concept.

### 4.2.3 Reference Classes

If we know that an event belongs to some well-defined reference class of events, and we have some statistical information about events in that reference class, then we can use it to tell us about this event. The uncertainty in any knowledge we obtain in this way will depend on how much we know about the reference class. With regard to direct application to decisions, this is likely to be most relevant when we do not know very much about the reference class; that is to say, when the uncertainty is quite high. This may seem to be very similar to forward causal structure and backward causal structure, as what we know about causal relationships is going to come from statistical observation of reality. The difference is that we could use a method like this to infer things about an event even when we do not know of any specific causal connection that it has with the events that we know about: From what we know of certain events, we might infer that other events are happening "out there" somehow in a way that correlates with them.

## 4.3 Generalizing Further

In the above, I have discussed things in terms of causal relationships and reference class, as if that is all that there is. In reality, we might generalize a bit further. Some of the relationships that we use to tell us about reality are just about how things are arranged in space. For example, by seeing part of a tree, we are used to inferring the existence of the rest of the tree: We are used to observing partial patterns and extending those patterns and extending causal patterns forwards and backwards in time is a part of this.

We might generalize by saying that “forward causal structure” and “backward causal structure” are just extrapolating from any patterns we observe.

Some cosmological ideas, such as Tegmark’s, view reality in terms that go beyond the causal or spatial relationships with which we are familiar. In Tegmark’s cosmology, the observable universe with its causal and spatial relationships would only be a small part of reality, and a lot of reality would need describing in more abstract terms (Tegmark, 1998). This would not be a problem if we assumed that the first two ways of generating information merely mean “extending any patterns that you know about” without such patterns having to be specifically spatial or temporal.

# 4.4 Reference Classes as a Generalization

## 4.4.1 A General Approach to Describing Reality

I have commented that the *reference classes* way of obtaining information might be viewed as very similar to *forward causal structure* and *backward causal structure*. This is because anything that we know about how to extend causal patterns into the past or future will come from previous observations of reality. For example, if we know that a ball has been traveling on a trajectory, then what we already know about the behavior of objects on trajectories is going to come from previous observations: effectively from what we know about the reference class of “an object on a trajectory”. I have described it separately to make it clear that we could still consider reference class issues even for events that do not have any specific, known causal connection to the parts of reality that we specifically know about. In practice, if humans approach philosophical issues like this, this is how the reasoning is likely to proceed. In reality, this kind of reference class reasoning could be viewed as a general case that encompasses *all* reasoning intended to tell us about the structure of reality. The patterns that we have observed in reality give us statistical information which allows us to obtain further information about the rest of reality. Some of this information is obtained by extending partially known patterns, while some of it is somewhat vaguer and involves obtaining knowledge of patterns where there is a lot of uncertainty about how they might “connect” with the known patterns.

## 4.4.2 The Patterned Floor Analogy

I will give an analogy for a general approach to describing reality. Suppose you are standing on an elaborately patterned floor in some ancient building. The floor was designed by a mathematically skilled artist, and involves mathematically derived patterns. Most of the floor is covered up: The only uncovered part is a small square near where you are standing.

By looking at the part of the floor that you can see, you can work out that various kinds of pattern are used. You can see some complete instances of patterns, and partial instances of the same patterns that are interrupted by the edges of the uncovered area.

You can use your knowledge of the complete patterns to work out how the partial floor patterns are likely to continue past the edges of the uncovered area: You can extrapolate. This is somewhat like using forward causal structure and backward causal structure. The part of the floor that you can see may also allow you to make more "general" statements about the unseen part of the floor, which are not specific enough to describe specific features being in specific places, but are more vague about location. As a rather crude example, if the part of the floor that you can see has circles and triangles on it, and circles tend to be near triangles, you might reasonably think this applies in other regions of the floor: that any circles that are "out there" tend to be near triangles. This may not seem to be the same process as extending the pattern past the edges of the uncovered region, but it is really: It is just less detailed.

### 4.4.3 Disregarding Ownership

What I have just said may seem to be an attempt to propose a specific, and possibly controversial view about how we make models. I do not think this is the case. Nevertheless, if that is the case, I will put this in more general terms.

From your observations of reality, you make a description of reality – which includes what has happened in the past and what you expect to happen in the future. You should not take into account who "owns" the observations: They are just observations that tell you something about reality. Of course, the observations will have been made from your point of view, and they will be particularly detailed with regard to your situation, but I am not suggesting you should ignore that. Saying that you should not take into account who "owns" the observations simply means that, beyond dealing with the observations that you have, which will reflect your place in reality anyway, you should not do anything "special" to make the fact that you own the observations relevant.

## 4.5 *Lack* of knowledge is power.

### 4.5.1 Why Knowledge Matters

I have been arguing that you should view your decisions as if they have the ability to decide features of reality beyond those that can be influenced by standard causality: When you make a decision you can view it as "deciding" the behavior of other beings. If this were really an important feature of everyday life, it should presumably be more obvious to us. In reality, anyone who really tried to make everyday decisions in this kind of way would probably not do very well. I do not think that this even needs arguing very strongly: It will be apparent to practically all readers – those that agree with the main points of the argument given here, as well as those who reject all of it. Anyone who conducted his/her everyday affairs on the basis that his/her decisions could non-causally determine the decisions of others to any significant degree would be considered, correctly, to be delusional.

Why should this be the case? The issue here is one of *amount of knowledge*. As I said previously, when discussing the scenario, when we are about to make a decision, we should consider the possible choices, and for each possible choice we should look at what reality would be like after that choice – what our knowledge that we made the choice tells us about ourselves, other people and the world in general. For us to prefer a particular choice, our knowledge that we made that choice would have to provide us with some information that made us find reality preferable. However, if we already have some piece of information about reality, independently of any choice that we make, then it cannot be provided by a choice: We already have it. Therefore, the amount of information that we already have about reality is going to limit the scope of our choices to generate new information that makes us prefer one reality over another. This does not mean that the degree of correlation does not matter. Of course, the amount of information that we obtain about ourselves from making a decision is important, but that is just information from *one event*. It will be easily submerged in the information we have obtained from the rest of reality if there is a lot of it.

I have said that evidential decision theory is not going to play any important role in everyday life, in terms of "controlling" the actions of others in a way that affects our situation. On the other hand, we might reasonably think it can play a more significant role in terms of managing your own behavior. Each decision that you make can be regarded as evidence of your own behavioral tendencies, and therefore the way that you are likely to behave in future, and it should be stronger evidence about you than it is about other people.

I will now look at the different ways in which we can get information from a choice: forward causal structure, backward causal structure and reference class. I will discuss them with regard to how their usefulness in generating information is likely to be affected by the information that we already have in everyday life, and therefore with regard to how much account we should take of them when making decisions in everyday life.

### 4.5.2 Forward Causal Structure and Knowledge

When we have made a choice, the effects on reality which we find out about by *forward causal structure* will tend to be significant, because this corresponds to the future that follows from our actions, and our future situation in the world tends to be sensitive to what we do. Most people will see this quite trivially. If you have the choice of jumping off a cliff or not jumping off it, and you imagine the projected future that follows from causally extending the consequences of this choice into the future, you are clearly going to have two very different futures. This kind of thinking dominates our everyday life – as it should do.

We can view this in terms of the *amount of information* that we obtain, from the knowledge that we just made a decision, about the way causal patterns extend forward

in time from that decision. Because the decision itself is important with regard to what happens, it needs to be taken into account. By knowing what decision we made, we can find out a lot about our expected future situation by using *forward causal structure*, and that means that the decision itself has to be taken into account. This should make sense: This is just the normal way in which we take decisions into account – by thinking of how they causally affect our future.

### 4.5.3 Backward Causal Structure

What about the apparent "backward causation": what we find out about reality from the *backward causal structure* approach? Your decision will have turned out the way it did because something caused it do so, and something would have caused that thing, and so on. This means that the past should correlate with your decision to some extent. Should this cause us to make decisions, in everyday life, that seem to give the past that we want? There is a problem here. The cause of your decision was your mental state immediately prior to it, and the cause of that was your mental state before that, together with experiences that you had been having – and this could be traced back through a history of mental states and experiences. The point about this is that *you will tend to know about this already because you have a memory*: You will already have a good idea of what has been happening in your mental past to put you into a mental state such that you make a particular decision a certain way.

We can see this with a deliberately ridiculous example. Suppose a man walks into a restaurant and decides to order caviar. His rationale is that if he has ordered caviar, it is probably because he is in the habit of eating it, and that probably means he was raised in a wealthy family, meaning he will have lots of opportunities. The main absurdity, here, is that his memory will already tell him whether he had a privileged upbringing: Ordering caviar is not going to change anything.

Generally, our memory of the past will mean we know so much about the causality behind our decisions, that our decisions are not going to tell us much about it – and therefore will not let us have this apparent control over it.

There is a further issue with trying to make decisions that seem to make your past desirable. Let us consider the scenario in which you are an *amnesiac* faced with deciding whether to order caviar in a restaurant. Should you order it in the hope of finding out later that it was because you had a privileged upbringing? The situation, here, however, would have changed entirely. Your desire to find out that you had a privileged upbringing would explain the decision quite well without recourse to a past in which you had a privileged upbringing. We might expect that this should therefore make any expected correlation between ordering caviar and having received a privileged upbringing rather weak.

### 4.5.4 Reference Classes

When you have made a decision, that decision is an event, and you can consider it as giving you information about any other event that shares a well-defined reference class with it. To belong to the same reference class, two events should have some kind of similarity and, as the event in question is a decision made by a human, events that share reference classes with it should be something like this: We should expect your decision to correlate with decisions by humans who are similar to you, other humans in general, other organic beings, other intelligent beings, etc.

The degree of similarity between members of the reference class is clearly going to be an issue. You should expect more correlation between a copy of you that was made ten minutes ago, complete with a copy of your brain structures, than you would expect with a member of some unknown alien species or a hypothetical artificial intelligence.

Another issue, and one that will be significant in everyday situations, is that your knowledge of the reference class does not necessarily just come from the single event corresponding to your own decision. If you know about any other events in the reference class, then by definition, these will contribute to your knowledge of the reference class. The more knowledge you already have about events in the reference class before making your decision, the less that you can gain from the outcome of your decision: Your belief about how other events in the reference class should occur will be almost completely based on what you already know about the reference class, and your own decision will have very little to do with it.

I will give an example of this.

Suppose that today, as part of the research for this article, I arrange to play some variation of the prisoner's dilemma game against an opponent. Suppose that I assume that there will be a strong correlation between my behavior and my opponent's, because as humans we both work in the same kind of way, and we are both trying to be rational. The only reasonable justification for playing in this way would be that, after I have made my decision, my knowledge that I decided in a certain way will also give me knowledge that there is some probability that my opponent decided in the same way, so my own decision will be telling me something about my opponent's actions, meaning that I "choose" my opponent's actions when I make my decision.

This all falls apart when we consider the fact that my knowledge of how human beings behave does not just come from my own decision in this game. Over my lifetime, I have observed a vast number of decisions being made by humans. I already have a lot of knowledge of the reference class. Any idea that my own decision at this moment is going to tell me anything *significant* about how humans behave in situations like this is naïve. It would, admittedly, give me a small amount of extra information about the reference class, but I should expect the correlation between my behavior and the other player's, and any apparent "control" that I have, to be extremely weak.

It is for this reason that the scenario at the start of this article involved each of the players being raised in a cave by machines and, as has been discussed, even that can cause issues with prior knowledge of the reference class. There is one way in which this issue may be less significant: Your choices *might* be considered good predictors of your own future decisions even when they are not telling you much about the decisions of others.

I should point out, however, that these are just decisions made by individual people in everyday life that we are talking about. There may be situations in which some entity has little knowledge of a reference class, and in which its own decisions therefore provide a significant amount of information about the reference class, therefore providing apparent, non-causal "control" over events in the reference class.

### 4.5.5 The Strangeness of the Idea that *Lack* of Knowledge is Power

I have argued that the apparent "control" you have over reality, outside the normal "forward causation" control that you have, is greatest when you know least. When you are faced with a decision, if you do not know much about reality, your decision will tell you a lot about it, thereby allowing you to *decide what you want to learn about reality*. This may seem to be implying that we should try not to acquire knowledge! The idea would be that when we do not know much, our decisions have enormous power and we are able to reach out and "control" the world non-causally, but when we find out more about the world, our decisions tell us less about reality, so we have less opportunity to find ourselves having significantly different knowledge about the world as a result of making different choices. Why should we want knowledge if it is going to take away our power like this?

Thinking like this would be a fallacy and it would result from thinking of things in terms of causality rather than correlation. If, in advance of some decision, you obtain information that reduces the "power" of that decision, the fact that you received that information has not drastically altered reality. The information that you gained must have been about features of reality that were correlated with your decision anyway, so if what you found out took away a lot of power from your decision, the correlation means that there is a good chance that it is consistent with what you would have decided anyway.

We can see this if we consider the time-delayed control group version of the scenario introduced at the start of this article. In that scenario, the deciding group of ten people actually chose whether to press the button in the past, and these people will decide your fate. When you are deciding whether to press, you should assume some degree of correlation with these people, even if it is very weak. Everyone who enters the room should act as if he/she has some control over the earlier decision of the deciding group. Now, suppose that *only the deciding group will be put through the process of being asked for a decision*. For the other players, things will be different now. When they enter

the room, they are immediately told how many of the deciding group pressed their buttons. They then go into the five minute decision period and have an opportunity to press the button or not press it.

Suppose that you walk into the room and are immediately told how the deciding group voted. You now know what your probability of death is in Step 1. The only reason for not pressing the button would have been to use the correlation to get a reduced probability of death in Step 1, but any "power" that your decision might have had in this respect has been taken away from you by the knowledge that you were just given. Not pressing the button therefore achieves nothing, and the sensible action is now always to press the button to avoid having to go into Step 2. (Needless to say, the fact that you are having a very different experience to that of the deciding group ruins the correlation, anyway – not that it matters now.)

When you were told how the deciding group voted, were you exposed to extra risk, on account of losing the chance to save your life by making the right decision to reduce your risk in Step 1? Thinking like this would be a fallacy, because what was going to happen to you was based on what the deciding group did all along, and nothing has changed that. Whatever degree of correlation would have existed between the decision you would have made and the decisions of the deciding group, that correlation would have been an indication of how you would have decided anyway, if the game had proceeded normally. Revealing the decisions of the deciding group did not change the situation: It merely "fast-forwarded" things. Fast-forwarding like this reduces the power of your decision – in this case to nothing – but tends to leave you in the kind of situation in which you would have put yourself anyway.

## 4.6 Analogy with Correlation in Everyday Life and Special Relativity

I have been arguing that the correct approach to decision-making is an evidential one, in which our own decisions provide as much evidence about reality as anything else. The intuition of many people will tell them that our "normal", everyday approach of just using forward causation is the correct one and that thinking in terms of correlation is unsound. Your own decision is only one source of information. In everyday life, we have a huge amount of information about what we expect reality to be like, and this makes any information that we obtain from our decisions less important, hiding the need for an evidential approach.

This is somewhat similar to the way in which special relativity is hidden from us under everyday conditions, and the way that special relativity disagrees with many people's intuition, because everyday life presents us with a special-case situation in which Newtonian physics applies and special relativity is not obvious. In more extreme conditions, special relativity starts to become apparent, and similarly the difference between a conventional approach to decision-making and an evidential one will tend to

be more obvious in extreme situations in which the information that you are obtaining from your decision is important because you are not getting much more information from everywhere else.

As I have said previously, use of your choices as evidence of your future choices may be a special case: Your choices *might* be considered better predictors of your own future decisions than they are for the decisions of others.

## 4.7 What about knowledge of your own cognitive processes?

One awkward issue that will appear when considering evidential decision theory is the information that you have from your own cognitive processes. I have said that you should not give special status to your own decisions just because you "own" them, but at the same time this does not imply that you should ignore the fact that you have the inside track on your own cognitive processes: You can observe the cognition that leads up to a decision. When I said that you should treat the issue of predicting the consequences of a choice as an outside observer would treat it, that does not mean that you should ignore this information: Rather, you should imagine an outside observer who somehow has all this, but is not you.

Clearly, the knowledge of your own cognitive processes will count as information: That is unavoidable, and from what has already been said, it will be obvious that it is likely to reduce the value of the information from your own choices to some degree. That is not a problem, in itself, however: I have admitted that evidential decision theory will be affected by information that you have.

It could be a bigger problem if we think that the information about your cognitive processes leading up to a particular decision gradually informs you about that decision, so that as you decide, you learn progressively more about the cognitive processes underlying your decision, and at some stage you learn so much that the decision itself is irrelevant: You can then choose as you wish, on a causal basis, and rely on the idea that you already know all about the cognition behind such a decision. Should this not make the choice itself irrelevant as a source of information?

This line of reasoning would be a fallacy. The problem with it is that any knowledge that you gain from your own decision-making process that tells you about how other entities are likely to decide is also going to tell you about how you are going to decide: It is going to reduce your uncertainty in your own decision. Suppose there is a situation where you have not yet decided, yet you are 100% certain what your choice is going to be. It should be obvious that this is a contradiction. If you know what your choice is going to be, you have *decided*: It makes no sense to separate knowledge about the decision that you are going to make from the decision itself. This may seem strange to some people who will insist that knowing how you will choose in future is not the same as choosing, but let us

consider an example. Suppose a man is 100% certain that he will *decide* to take his umbrella when he goes out, yet he says to you that he has not yet decided to take his umbrella, but he is sure that when he does decide he will decide to take it. You should wonder, here, what exactly is supposed to be involved in this "decision". Why even bother going through the "deciding" process? His mind already knows what he is going to do, so the "decision" is redundant: It has already been made.

Suppose there is a situation in which you are 80% certain about the decision that you are going to make, but have not yet decided: By the same reasoning, it is incoherent to deny that you are 80% decided. Whatever cognition you have been doing to put your mind in this state, it is 80% of the way towards knowing what your choice will be, and when you know what your choice will be, you have decided. This means that any information that is generated by your own cognitive processes that informs you about your own tendency towards a particular choice, while making it, is actually part of the act of making that choice itself. Now, suppose we imagine a situation just before you have made a choice, in which you are supposed to know all about the cognition underlying that choice. For your knowledge to be of any use at all in determining the behavior of others, it would have to tell you something about your own tendency to choose a particular way, so it would have to be some degree of certainty about the choice you face. Let us say that "just before you have made a choice" means 99% certainty. This means that, when you are 99% certain that you will choose a particular way, you are supposed to be able to use this knowledge to predict what other people will do, while somehow heading off into causal decision theory and doing what you want. The problem here is that *you are 99% of the way through the deciding process*: You have already forced your decision in a particular direction. Suppose we imagine you escaping this: You learn a lot about deciding process, and then you choose decide in a way that has nothing to do with what you have learned. This would mean that what you just found out about the deciding process was not even good enough to say anything worthwhile about your own imminent decision: It would be absurd to expect it to tell you about other people's.

To put this another way:

*To choose is to reduce your uncertainty in some future choice, and to reduce your uncertainty in some future choice is to choose.*

This means that there is no "get out" like this. Trying to find out about your tendency to choose a particular way exacts a price: It forces you to choose.

## 4.8 General Artificial Intelligence

The issues being discussed here are relevant in attempts at general artificial intelligence: We should want an intelligent agent to make its decisions rationally. If we are using some approaches to artificial intelligence, this should not be an issue. We should expect

an evidential approach to decision-making to operate naturally in a system if it is functioning in an ideal way and basing its decisions on models of reality that it makes from observations, and if those models are made without giving any special status to itself as the “owner” of the system. Note that I said “ideal way”, however: A real-world system is likely to be taking short-cuts and approximating the world, and this might cause something more like causal decision theory to result.

# 5 What is a "decision"?

## 5.1 "Controlling" Reality

I have argued for evidential decision theory, and this inevitably means thinking in terms of "controlling" reality non-causally. This is clearly going to be a problematic concept and we need to be careful with the semantics and to think about what we are really saying. This is why, so far, I have tended to put the word "control" in quotes: to make it clear that the word is not really being used rigorously. At some point, however, we have to deal with the issue of what we really mean, and we will do that now.

Suppose some decision faces you, with a number of possible choices. Each choice leaves you with a different description of reality. Part of this difference comes from the fact that each choice tells you something about yourself, and therefore about reality as a whole, through correlation or extending causal patterns backwards. You can consider each possible choice, and what it tells you about reality, and "choose" the reality that you want. From this, it may seem that you are able to "control" reality non-causally.

Can we really use the word "control" in this sense? Although our intuition may be much against it, I suggest that it is hard to avoid. The main point here is that each choice you make has specific features of reality associated with it, and *your act of choosing is about which reality you want*. If you know that reality is going to be one way after making once choice, and one way after making another, it is not coherent to think of justifying the decision in terms other than those of "control". Whether or not the control has any causal mechanism behind it should be irrelevant here. From your point of view, you get one reality if you do something, and you get another reality if you do something else: *For all practical purposes you are controlling*.

This, of course, will not persuade some people. An objection will be that you can only be considered to be "controlling" something if there is a clear causal chain of events between your decision and that thing and that "correlation is not causation". The issue here would be just one about the semantics of the word "control". Suppose we take an extreme example. A decision faces you in which if you make one choice, you will almost certainly live, and if you make another choice you will almost certainly die. However, your knowledge of this does not come from any direct causal sequence following on from your choice: It comes from some correlation. Someone might tell you that you have no "control" here – that "control" is only concerned with causation – but that would not change the fact that it would still be really stupid to make the choice associated with almost certain death. Regardless of the semantics, choosing that way means you are almost certainly going to die.

Now, suppose we admit that the word "control" is inappropriate for this: that we should reserve it for causation. We might make another word to describe how various outcomes are associated with a decision we make, but *not necessarily* in a non-causal

way. Suppose we way that this word is “zontrol”. We might say that when you are making this decision you “zontrol” whether you live or die – which only means that with one choice you expect almost certainly to live and with the other choice you expect almost certainly to die – without any assumptions about why this is the case.

Now, when a decision is facing you, you should not just look at what you “control”. That is too restrictive. You need to look at what you “zontrol”, which takes care of everything. This is particularly obvious with by the example we are using here, where if you look at what you “zontrol” you will see that one choice is associated with almost certain survival and the other choice is associated with almost certain death.

All this would mean that the concept of “control” would be irrelevant in decision-making. We would have moved on to the concept of “zontrol”. Any situation in which “control” would be of interest is dealt with by the more general idea of “zontrol” anyway. The word “control” has now become useless in as much as it relates to planning our actions and making decisions. We are using “zontrol” for what we previously used the word “control”. Changing the word should seem pointless. It makes more sense just to keep the word “control”, and to apply it to everything that is associated with making some decision, whether by causation or otherwise.

## 5.2 Wanting a Special Status for Causal Relationships

Some people will object to evidential decision theory on the grounds that causal relationships have some special status: The idea would be that if one event causes another then they are linked in a more “real” way than if the two events are merely correlated. We should, however, try to look at things from the point of view of an observer looking at the system from outside. Such an observer would just see relationships of various kinds. Some would be causal, linking events together over time in an unbroken chain. Some would be non-causal patterns. There would be nothing profoundly different about the causal relationships. Causal relationships are merely a specific type of pattern which describes how events are arranged over time. All the relationships would be usable for extending a description of reality, made with limited observations.

## 5.3 Concerns About “Free Will”

Some people will object to all this by saying that the only reason evidential decision theory can work in the first place is that your decisions are not “free”, but are controlled by what you are, and that, therefore, if we adopt an evidential approach we are saying that *we cannot even make decisions*. For example, in the scenario at the start of this article, the justification for not pressing the button would not be that your decision is “free” and you can then cause other people not to press it: It would be that your

decision is not “free”, and is caused in a similar way to the decisions of the other players, and is therefore correlated with them.

This objection is based on flawed thinking, because if there is a problem here, then it arises whether we use evidential decision theory or not. If we think that our decisions are caused by the states of our brains, or even just “our nature” if we do not want to adopt a particular physical model, then our decisions are dependent on things and every time we are “making a decision”, some chain of causality, started back in the past, is simply running through to a conclusion, with our “decision” having its place in the chain. If the reality of this concerns you when using evidential decision theory, then it should concern you just as much when using causal decision theory: It is not as if the causality underlying your own decision-making processes suddenly goes away when you adopt causal decision theory.

There is an inconsistency here in how people are treating the word “decision”. When we are following a “conventional”, causal approach to decision theory, few people are going to get bothered about use of words such as “decision” or “choice”, and start objecting that our choices or decisions are caused by things, and so are not real. Anyone who really thinks about what is going on, and it is not embracing various mystical notions, knows this anyway. Philosophically, we deal with it by making limited demands of words like “decision” or “choice”. We may know that decisions are caused, but we are not expecting more from the concept of “decision” when we use the word: We are not expecting it to confer free will, in the strong sense, on things. Some people who would be quite happy operating like this will take an entirely different position when we start talking about evidential decision theory, and will start objecting that the decisions which correlate with other things out there in the world cannot be decisions as they are just caused: Doing that would be expecting much more from the concepts of “decision” and “choice” in evidential decision theory, and it would be inconsistent.

What we call a “decision” is not some “source” of causality. Instead, our decision is in a sequence of causal events. Causality acts *through* our decisions when we perceive that our decisions are causing things. Evidential decision theory does nothing to change this.

For some reason, this seems to become an issue with people when evidential decision theory is proposed, even with people who have no problem with the idea that underlying physical processes are causing their thoughts. This is probably because evidential decision theory relies on this causal underpinning, and confronts you directly with it, whereas causal decision theory does not.

## 5.4 The Language of “Decisions”

Some people will still object to the idea that we can “decide” or “make choices” if our thoughts are themselves caused by other things. As I have said, this is not a problem for evidential decision theory specifically. Nevertheless, I will try to answer it.

The problem here is one of semantics. We experience a process that we call “making a decision”. It is when we have a number of alternatives, and we weigh them up to determine which is preferable, and then select that one. If we look deeper into what is going on we will find underlying physical processes: neurons firing and so on. We will see a chain of causality: What we thought of as our “decision” was really one thing causing another thing, which in turn causes another thing and so on. The “decision” or “choice” disappears when we look closely. Why should this surprise us, though? *Everything* disappears when we look closely enough and we just see its component parts or the underlying processes. For some reason, people expect “decisions” or “choices” to be different.

If someone thinks he was making a decision, and then finds out that his brain was following some causal path, he may think that his decision was not a decision, but it was a decision really: There just happened to be a causal path underlying it. A problem would only arise if he thought things could ever have been otherwise: that there was “free will” in some strong sense of the term.

Faced with this, some people may object that it is pointless to talk about what a “correct” decision theory is: how people should choose. They may say that it is pointless, in the scenario at the start of this article, to talk about whether you *should* or *should not* press the button, as if there is any real “choice”, because you are going to “decide” what the causal processes underlying your “decision” are going to make you decide. I have always found this view of decision-making strange. The idea seems to be that, when you realize your decisions are causally determined, you should despair of making decisions at all and ignore any evidence, arguments or justifications that might be available in the decision-making process. This ignores the fact that, even if your decisions are causally determined, consideration of such things is part of that same causal process. Some people say that if their decisions are causally determined, there is no point in trying to decide anything, saying things like, “Oh well, if that’s the case I may as well not bother doing anything then.” This, however, would be inconsistent, because it seems to imply that *you can choose to give in and not do anything*. People saying things like this are arguing for the futility of making any decisions at all, yet that is a decision they do not seem to mind making, and they are arguing for the futility of making arguments to justify actions, yet that is an argument they do not seem to mind accepting. You might say that they know that what seems to be a “decision”, when they do this, is merely causal processes in the brain, but the same can be said of all the other “decisions” they are abandoning as futile! Again, this problem seems to arise from people expecting too much of words like “decision” or “choice”. Whatever decisions might or might not be a low level, your *experience* is one of deciding – and it is *you*, not your individual neurons, that arguments like this one are addressing.

We might side-step the whole issue, here, by not using words such as “decision” or “choice”. As an example, suppose you are in the room in the scenario at the start of this article. You can press the button or not. Without referring to “choice” or “decision”, I

can just say that an argument can be made that *not* pressing the button may give you a better chance of survival, because of what it says about you, and therefore about other people. There would be no claim, here, that you have any "free will", and I would be staying away from the whole semantic problem around words like "decision" or "choice". You can then consider what I have said, and it may affect your behavior. You may then, due to the causal processes occurring in your brain, desist from pressing the button. All of this can occur causally, without any reliance on any idea of "deciding". It therefore does not invalidate an argument like this to say that "free" decisions are taken away from us by the argument, while at the same time it urges us to decide in a particular way.

## 5.5 Meta-Causation

I have said that we should view choosing between different descriptions of reality, whether they differ due to causation or not, as controlling reality. This raises the issue of whether we should treat the concept of "causation" in the same way. The problem, here, is that "causation" is a specific concept: People understand it as referring to *causal* relationships.

Nevertheless, when we are faced by decisions in which we can control the outcome non-causally, it makes sense to say that we are doing something similar to "causing" a particular outcome if we choose a particular way. I suggest that to deal with this we use the term "meta-causation". Meta-causation would be the relationship that some event has and the choice that correlates with it. We would say that making that choice *meta-causes* that event, yet we would not necessarily be making any claim of conventional causality. Causality would be a special case of meta-causation, in which the correlation happens to involve causal relationships.

# 6 Conclusion

There are two main, and conflicting, approaches to making decisions: *evidential decision theory* and *causal decision theory*. In evidential decision theory, the decision itself provides information about the expected state of the world, because it tells us something about the decider, the decider's past and the behavior of other entities with some degree of similarity with the decider, which may be expected to have behavior which correlates with that of the decider. Many people reject evidential decision theory, insisting that causal decision theory – which just involves assessing a decision in terms of the causal consequences that follow on from it – is the only valid approach. A justification has been given for using *evidential decision theory* as opposed to *causal decision theory*.

The justification of evidential decision theory has involved a scenario in which the reader has been urged to agree with the idea of not pressing a button, even though causal decision theory would tell the reader to press it. Part of the justification for doing this has been based on limiting the amount of information available to the player in the scenario. Even if there is significant correlation between your decisions and events in the world, if you have a lot of background knowledge about the world, independently of the knowledge that you will gain from your decision, your decision becomes less important in evidential terms.

Some people are perturbed by the idea that evidential decision theory can have us deciding as if we are assuming backwards causation: as if we are choosing previous states of reality. This is particularly apparent in Newcomb's paradox, which causes many people to reject any idea of deciding on an evidential basis. However, any decision we make gives us information about a previous state of reality, even if it is only that it is one of a set of states leading us to that decision, and therefore any decision means acting as if there is backward causality. Nor does the problem go away if we assume non-determinism: It merely gets expressed statistically.

Using light cones as a simplification, there can be events which are just "on the edge" of being controllable with a causal approach, assuming the decision occurs at an instant, but nothing profound happens when the transition to an evidential approach is made.

Evidential decision theory should seem justified if we view our decisions from the perspective of an outside observer who knows what we know. To such an observer, our decisions would give information about the world in general, and it would be natural to take an evidential approach with regard to them. Likewise, we should not regard our own decisions as having any special status due to being owned by us. In fact, doing this violates the Copernican principle.

Advocates of causal decision theory may wish to restrict our projections of the future to those which follow causally from our actions, but this raises the issue of how we stop our choice contaminating the model without limit. A choice may have a causal effect, which has a causal effect, which implies a correlation with something else and which ultimately implies a correlation with the actions of various agents.

Another argument for evidential decision theory is based on degrees of similarity. If playing a game involving identical players, we would clearly be foolish to assume that our actions did not correlate with those of others, yet the situation can be practically the same if the players are *almost* identical. This argument becomes more powerful still if we imagine that you have previously observed many such games from outside and are familiar with the time taken for players with varying degrees of similarity to diverge.

The degree of correlation is important in evidential decision theory, but the amount of information you are obtaining from sources independent of your choice is also important. In general, the less the amount of such information that you have, the more important is the knowledge gained from your choice. When you have a lot of background knowledge, such as knowledge about how people typically behave or your own past, the information that you gain from your choice, other than with regards to forward causation, will be less important, and is less likely to describe a dramatic change in your situation. An analogy has been made between this and the way that special relativity is not apparent in everyday life, with Newtonian physics being adequate for everyday situations: Evidential decision theory should become more relevant in situations where you know less. When there is a lot of background information, evidential decision theory resembles causal decision theory. This means that, although I am trying to justify evidential decision theory, I am in no way arguing that it has any place in, for example, standard economics: Economic dealings between people will involve so much background information that a single choice by an individual will tell that individual nothing of any consequence. Evidential decision theory is more suitable for “extreme situations” in philosophical arguments. There is one respect in which evidential decision theory may be relevant in everyday life: We might consider our choices as evidence of our future choices, and therefore each choice we make might be regarded as implying future choices.

It has been argued that we should view the correlation in evidential decision theory as corresponding to real control over reality: that we should act as if we can control other people’s behavior, or manipulate the past. This does not mean that we are assuming any magic causation. Instead, the term “meta-causation” has been proposed. A choice *meta-causes* an event if it corresponds with occurrence of that event irrespective of whether or not the event causally follows from it.

In this article, I have put forward reasons for favoring evidential decision theory, while admitting that it has little significance in many everyday situations in which we have a lot of knowledge. In the next article, I will look at some extreme situations where it

*could* be relevant. Evidential decision theory will be more important in situations in which we have very limited background knowledge, and situations in which this applies will be discussed. One such situation could be when the entity making the decisions is not a single human, but *all of human civilization*. As we lack knowledge of the behavior of any non-human civilizations, our knowledge of the reference class in which our civilization, and other possible civilizations, appear, might be viewed as being minimal.

# 7 Acknowledgements

Yvonne Deborah Finch, Joseph Kenneth Frantz and Michael Fridman have been helpful in discussions of the scenario that is introduced at the start of this article, similar scenarios, and general issues surrounding the arguments given here. In the cases of Joseph Frantz and Michael Fridman, this was in e-mail exchanges. With Yvonne Finch some of it was while eating a very nice curry outside. With zucchini.

# 8 Bibliography

Everett, H., 1957. Relative State Formulation of Quantum Mechanics. *Reviews of Modern Physics*, 29, pp.454-462.

Hofstadter, D.R., 1986. *Metamagical Themas: Questing for the Essence of Mind and Pattern*. London: Penguin Books. pp.748-750, pp.752-755, p.758, pp.763-766. (Originally published: 1985. New York: Basic Books).

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.

Salgado, R., 1996. *A more illuminating look at The Light Cone.* [Online] The Light Cone: an illuminating introduction to relativity. Available at: http://www.phy.syr.edu/courses/modules/LIGHTCONE/lightcone.html [Accessed 13 September 2010].

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]).