A Classification of Newcomb Problems and Decision Theories

Kenny Easwaran

March 24, 2018

1 Introduction

Consider the following two decision problems.

Example 1: The Newcomb Problem

You walk into a booth at a carnival and find a strange game being offered. There is a mad scientist, who offers you the option of taking just the contents of an opaque box, or the contents plus another box containing a thousand dollars. But as you walked in, her machines scanned your brain and body, and using the laws of physics and neurology, her computers predicted what you were going to choose. If they predicted that you would choose just the contents of the opaque box, she put a million dollars in it. But if they predicted that you would take both boxes, or if they couldn't predict what you would do, then she put nothing in. While you are deliberating, you learn that her predictions in the past have been moderately reliable, so that a significant majority (though not all) of her predictions of each type have been correct, conditioning either on her prediction or on the choice made by past participants. Should you just take the contents of the one opaque box, or take both boxes?

Example 2: The Smoking Lesion

Scientists discover that the statistician R.A. Fisher's work on behalf of the tobacco companies was actually correct. Although there is a strong correlation between smoking and lung cancer, it turns out that this is not because of any mechanism by means of which the nicotine, tar, or anything else affects human bodies. Rather, it is because of a previously unknown physiological condition that both has a tendency to cause cancerous growths in the lungs (and is the primary source of these growths), and seems to behaviorally lead to smoking. (This condition is caused by environmental factors that happen to have been cleaned up greatly in North America over the

past few decades, but are still quite high in developing countries.) You have had some cigarettes in the past, and have moderately enjoyed them. Should you take up smoking?

Classic "evidential decision theory" notes that in each case, the fact that an agent takes one of the options (taking both boxes, smoking) is good evidence that she will receive quite a bad outcome (failing to get her million dollars, cancer), and thus says that a rational agent who knows the setup will choose the other option. Classic "causal decision theory" notes that in each case, at the moment of decision, the features of the world that give rise to the bad outcome (prediction, physiological condition) are already fixed and outside of the agent's control, and instead focuses on the fact that the option that is correlated with this outcome will definitely cause a moderately good outcome (getting a bonus thousand, the enjoyment of a cigarette), and thus says that a rational agent who knows the setup will choose this option. Many people have expressed intuitions that cross-cut these cases, and thus some theorists have gone to great lengths to try to define versions of decision theory that recommend taking both boxes in The Newcomb Problem, but smoking in The Smoking Lesion. (See, among others, the "ratifiability" of Jeffrey (1981), the "tickle defense" of Eells (1982),the attack on causal decision theory by Price (1986), the "cohesive decision theory" of Meacham (2010), and the "timeless decision theory" of Yudkowsky (2010).)

This paper does not seek to adjudicate between these theories, but instead attempts to give a structure for classifying problems of the relevant sort, and to use this structure to give some insight into the classification of possible decision theories for dealing with them. These two classification systems then suggest some ways of generating new cases of the relevant sort, and new theories as well. The last section of the paper considers some interactions of the concepts of free will, determinism, and rationality that are suggested by these classifications, though there are no definitive claims about these issues.

2 Classifying Decision Problems

One of the puzzling features of these cases is the fact that the "states of the world" are not probabilistically independent of the agent's choice. Thus, these decisions can't be properly analyzed by the methods proposed by Savage (1954) without having vastly more information than is available. In fact, in Savage's framework, the proper analysis of a decision problem can only proceed when one has made a maximally fine-grained analysis of all possible uncertainty about the world, and all the possible actions that one can take over the course of one's whole life. Without this extremely fine-grained analysis, one can only create a "small world" representation of a decision problem. He gives some conditions under which a small world representation of a problem will make the same recommendation as the detailed "grand world" representation, but these conditions are at best suggestive. Much of the development of decision theory since Savage (and especially the development of evidential and causal

decision theory, and their competitors and modifications) has been driven by the search for principles for setting up these small world representations in cases like these, where there are natural notions of "act" and "state" that are not probabilistically independent.

For simplicity, I will focus on decision problems that are naturally described in terms of two possible acts and two possible states, which have some sort of correlation with each other rather than being independent as required by Savage. My proposed classification of these problems consists of two parts. First, I will characterize the value of the possible outcomes to the agent. Second, I will describe the nature of the correlation between the states and the acts. Although these parts leave out most of the real-world texture and background of the problems, I suspect that they are largely adequate for the purposes of decision theory.

2.1 Value structures

I will represent the values of the possible outcomes of an act for an agent in a 2×2 matrix like $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$. Each row of the matrix corresponds to one of the possible actions the agent can take and each column corresponds to one of the states. The contents of each cell then represents the value of the outcome that would be achieved by the corresponding combination of act and state. In most cases, the precise value of the outcome isn't relevant, so we will only be concerned with the ordinal structure of the value of these four outcomes. Thus, I will generally represent the values with the numerals 1-4, with 1 representing the best outcome and 4 the worst. When relevant, I will supplement this with some more detailed comparisons — for instance, $1 > 2 \gg 3 \sim 4$ will represent the fact that for the analysis of the relevant decision problem, it matters that the best option is better than the second best, which is much better than the third and fourth, which could be equal or even slightly reverse the comparison without changing things much. When merely ordinal information is relevant, there are 24 possible value matrices, though there are more if these magnitudes of comparison are considered.

Since switching the two rows just corresponds to relabeling the actions the agent can take, and switching the two columns just corresponds to relabeling the states, each decision problem has four possible representations. I will cut this to two by following the convention that the act corresponding to the top row is positively correlated with the state corresponding to the left column, and the act corresponding to the bottom row is positively correlated with the state corresponding to the right column. (When we look at the structure of the correlations in section 2.2, it will be clear that there are further complications that I am ignoring right now.) To remove this last degree of freedom, I will choose the representation where the upper left correlated outcome is better than the lower right one. In addition to giving each decision problem a unique matrix representation (at least, when the two correlated outcomes have distinct values), this eliminates half of the possible value matrices, so that there are only 12 when merely ordinal information is represented.

The first thing to note is that both The Newcomb Problem and The Smoking Lesion are represented by the matrix $\left(\begin{smallmatrix} 2 & 4 \\ 1 & 3 \end{smallmatrix}\right)$, with $1 > 2 \gg 3 > 4$. Each outcome of the lower action (taking two boxes, smoking) is slightly better than the outcome of the upper action (taking one box, not smoking), but the outcomes of the correlated left state (million dollars, no cancer) are much better than the outcomes of the right state (empty box, cancer). Several other puzzles that are prominent in the literature can be represented with the same payoff table.

Example 3: The Toxin Puzzle

> (Kavka, 1983, p. 33-34) You have just been approached by an eccentric billionaire who has offered you the following deal. He places before you a vial of toxin that, if you drink it, will make you painfully ill for a day, but will not threaten your life or have any lasting effects. (Your spouse, a crack biochemist, confirms the properties of the toxin.) The billionaire will pay you one million dollars tomorrow morning if, at midnight tonight, you *intend* to drink the toxin tomorrow afternoon. He emphasizes that you need not drink the toxin to receive the money; in fact, the money will already be in your bank account hours before the time for drinking it arrives, if you succeed. (This is confirmed by your daughter, a lawyer, after she examines the legal and financial documents that the billionaire has signed.) All you have to do is sign the agreement and then intend at midnight tonight to drink the stuff tomorrow afternoon. You are perfectly free to change your mind after receiving the money and not drink the toxin. (The presence or absence of the intention is to be determined by the latest 'mind-reading' brain scanner and computing device designed by the great Doctor X. As a cognitive scientist, materialist, and faithful former student of Doctor X, you have no doubt that the machine will correctly detect the presence or absence of the relevant intention.)

Ignoring the central issue that Kavka uses this puzzle to discuss (namely, the question of whether it is even possible to form the relevant intention, knowing that your future self will have some incentive to go back on it), the payoff structure is quite similar. Regardless of whether or not you formed the intention tonight at midnight, not drinking the toxin tomorrow will be somewhat better than drinking the toxin. However, regardless of whether or not you drink the toxin tomorrow, forming the intention tonight results in a substantially better outcome than not forming the intention. There is presumably some sort of correlation between forming the intention and drinking the toxin, and it's better to form the intention and drink the toxin than to do neither.

Example 4: Parfit's Hitchhiker

> (Parfit, 1984, p. 7) Suppose that I am driving at midnight through some desert. My car breaks down. You are a stranger and the only other driver near. I manage to stop you, and I offer you a great

> reward if you rescue me. I cannot reward you now, but I promise to do so when we reach my home. Suppose next that I am transparent, unable to deceive others. I cannot lie convincingly. Either a blush, or my tone of voice, always gives me away.

Here again we have an issue for planning. If I sincerely plan now to give you the reward when I get home, then regardless of whether or not I actually do give you the reward, I will get the major benefit of a ride out of the desert. However, actually giving you the reward is a cost to me, and it is correlated (in some sense to be discussed later) with the sincere plan. (In this case, if I don't plan to give you the reward, I may not even get the option to give it, because I'm stuck in the desert, so we have $3 \sim 4$ rather than $3 > 4$.)

Example 5: Prisoner's Dilemma with a Twin

> You and your twin sibling have robbed a bank. However, you were caught and are now being held in separate cells. The police have enough evidence currently to convict each of you of conspiracy. However, they don't have enough evidence to convict either of you of actually robbing the bank. The police say that if you testify against your sibling (giving them enough evidence to convict), they'll drop your conspiracy charge. They're also making the same offer to your sibling in the other cell. Conspiracy carries a sentence of a year in jail, but bank robbery carries a sentence of five years.

In this case, regardless of whether or not your sibling testifies, if you testify, you'll spend one less year in jail. However, because your twin is very much like you, your testifying is strongly correlated with your twin testifying, which would add five years to your jail sentence whether you testify or not.

Further analysis of the similarities and differences among these decision puzzles will wait until section 2.2.

The distinct, but related, matrix $\left(\begin{smallmatrix} 3 & 2 \\ 1 & 4 \end{smallmatrix}\right)$ also corresponds to a few prominent puzzles in the literature.

Example 6: Death in Damascus

> (Gibbard and Harper, 1978, pp. 185-6) Death works from an appointment book which states time and place; a person dies if and only if the book correctly states in what city he will be at the stated time. The book is made up weeks in advance on the basis of highly reliable predictions. An appointment on the next day has been inscribed for him. Suppose, on this basis, the man would take his being in Damascus the next day as strong evidence that his appointment with death is in Damascus, and would take his being in Aleppo the next day as strong evidence that his appointment is in Aleppo. Two acts are open to him: A, go to Aleppo, and D, stay in Damascus. There are two possibilities: SA, death will seek him in Aleppo, and SD, death will seek him in Damascus. He knows that death will find him if and only if death looks for him in the right city.

In this case, the rows correspond to the actions of going to Aleppo and staying in Damascus. Two states of the world are possible, death seeking him in Aleppo and death seeking him in Damascus. The best outcomes ($1 \sim 2$) occur if the man goes where death isn't seeking him, while the worst ($3 \sim 4$) occur if he goes where death is seeking him.

Example 7: The Psychopath Button

> (Egan, 2007, p. 97) Paul is debating whether to press the "kill all psychopaths" button. It would, he thinks, be much better to live in a world with no psychopaths. Unfortunately, Paul is quite confident that only a psychopath would press such a button. Paul very strongly prefers living in a world *with* psychopaths to dying.

In this case, the upper row is the act of not pressing the button while the lower row is pressing, and the left column is the state of not being a psychopath while the right column is being one. The best outcome is pressing the button while not being a psychopath, and the worst is pressing while being a psychopath. Not pressing is in between, so we have $1 > 2 \sim 3 \gg 4$.

Both of these puzzles have been proposed to cause various problems for causal decision theory. Causal decision theory leads to a kind of instability in Death in Damascus (whatever decision one makes, one will then think that the other would have been better). But it actually seems to lead to the intuitively wrong recommendation in The Psychopath Button (if Paul's prior probability that he is a psychopath is relatively low, then causal decision theory tells him to press the button, which seems like a really bad idea). The important difference here is primarily due to the fact that for Death in Damascus the ordinal structure is really $1 \sim 2 \gg 3 \sim 4$, while for The Psychopath Button the ordinal structure is really $1 > 2 \sim 3 \gg 4$. As a result, Death in Damascus might actually be better represented with $\left(\begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix}\right)$, while The Psychopath Button might be better represented with $\left(\begin{smallmatrix} 2 & 2 \\ 1 & 3 \end{smallmatrix}\right)$.

Some of the other possible payoff tables have been investigated in the context of game theory, though not so much in the context of evidential or causal decision theories. In the game theory literature, $\left(\begin{smallmatrix} 1 & 4 \\ 2 & 3 \end{smallmatrix}\right)$ corresponds to the "Stag Hunt", while $\left(\begin{smallmatrix} 1 & 3 \\ 4 & 2 \end{smallmatrix}\right)$ corresponds to a "coordination game".

A few other payoff tables are unlikely to yield any interesting cases. With $\left(\begin{smallmatrix} 1 & 2 \\ 3 & 4 \end{smallmatrix}\right)$ and $\left(\begin{smallmatrix} 1 & 3 \\ 2 & 4 \end{smallmatrix}\right)$, the top action is preferred to the bottom one regardless of state, and the left state is preferred to the right one regardless of act, so considerations of the correlation don't seem to change the preference for the top action. However, $\left(\begin{smallmatrix} 2 & 1 \\ 4 & 3 \end{smallmatrix}\right)$ may yield some interesting considerations, if $1 \gg 2 > 3 \gg 4$. Even though there is a sort of "superdominance" of the top action over the bottom action (in that any outcome of the top action is better than any outcome of the bottom action), the strong interest in getting the much better outcomes of the right rather than left column may motivate some attempt to take advantage of the correlation of the right column with the bottom row, even at the cost of running the risk of actually choosing the bottom action. Describing a

thought experiment that motivates this reasoning may be an interesting project for future work.

2.2 Causal structures

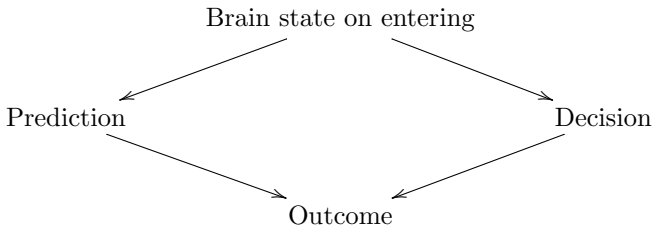
Having considered the payoff tables, it is useful to investigate the nature of the correlation between action and state. My overarching methodological assumption is a version of Reichenbach's Common Cause Principle — if two things are correlated, then either one must be a cause of the other, or there must be some common cause. (See Arntzenius (2010) for careful formulations of the principle.) My approach is to analyze the causal structure of the situation using one of the "causal modeling" frameworks of Pearl (2000) or Spirtes et al. (2000). The differences between these frameworks might be relevant for detailed analysis of some cases, but for present purposes, an outline of the shared idea may be sufficient.

A causal model represents the causal relations among various variables of interest. These variables might be binary (the money is in the box or not; the agent takes one box or two) or might have multiple possible values (the amount of fertilizer added to a field; the height of the crops in the field). Each variable is represented by one node in a network, with an arrow going from one variable to another if the values of the first play a direct role in determining the values of the other. (We can describe the first as the "parent" and the second as the "child".) A central assumption is that causation is non-circular — there is no path from a variable to itself that follows arrows only in the direction they point. Thus, for any two variables, there is a definite notion of whether one is a causal ancestor or descendant of the other, or if there is no causal chain running from one to the other.

Once this network is set, the model can be specified by providing, for each variable, a probability distribution over all its values *conditional* on any particular setting of values for its parents. These probability distributions for each variable individually then combine to define a joint distribution for all the values of all the variables collectively, by working through the network iteratively from the variables earliest in the chain to those latest in the chain. This joint probability distribution then validates Reichenbach's Common Cause Principle — two variables will have a probabilistic correlation only if one is a causal ancestor of the other, or they share a causal ancestor in common. At the level of analysis I'll be using, it won't be essential to give the detailed probability distributions — I'll just provide the network diagrams. But to give a properly detailed decision-theoretic analysis, the specific probabilities will be needed (just as the specific utility values for outcomes will too).

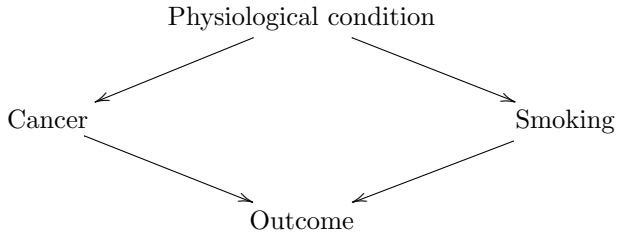
For The Newcomb Problem, the causal network diagram might look some-

thing like this:



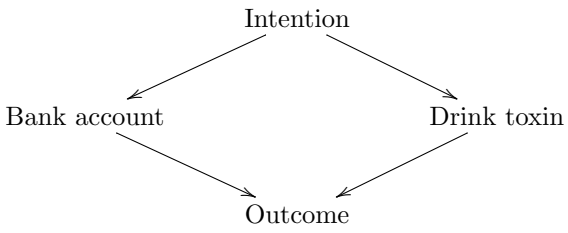
The outcome of the problem is most directly determined by whether or not the predictor put the money in, as well as the agent's decision whether to take one or two boxes. The prediction and the decision are correlated, because both are in some sense determined by the agent's state on entering the booth. This determination may be probabilistic — perhaps the brain states that the predictor is able to read merely make certain decisions on the part of the agent very likely, or perhaps the predictor's scanner is not 100% accurate.

For The Smoking Lesion, the diagram is similar:

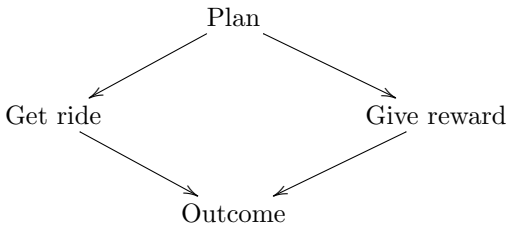


In this case, "Outcome" may not be a distinct event that is *caused* by the fact of whether the agent has cancer or not and smokes or not, but may rather be *constituted* by it. But in order to see how the overall satisfaction of values for the agent is brought about by the various factors involved, we need a single node to represent it.

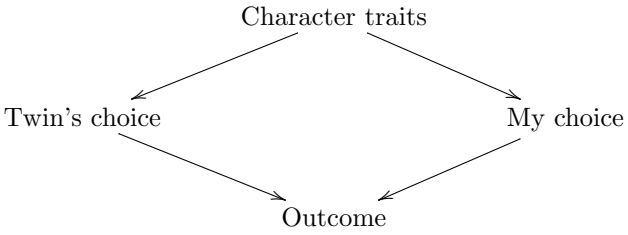
The diagrams are also very similar for The Toxin Puzzle:



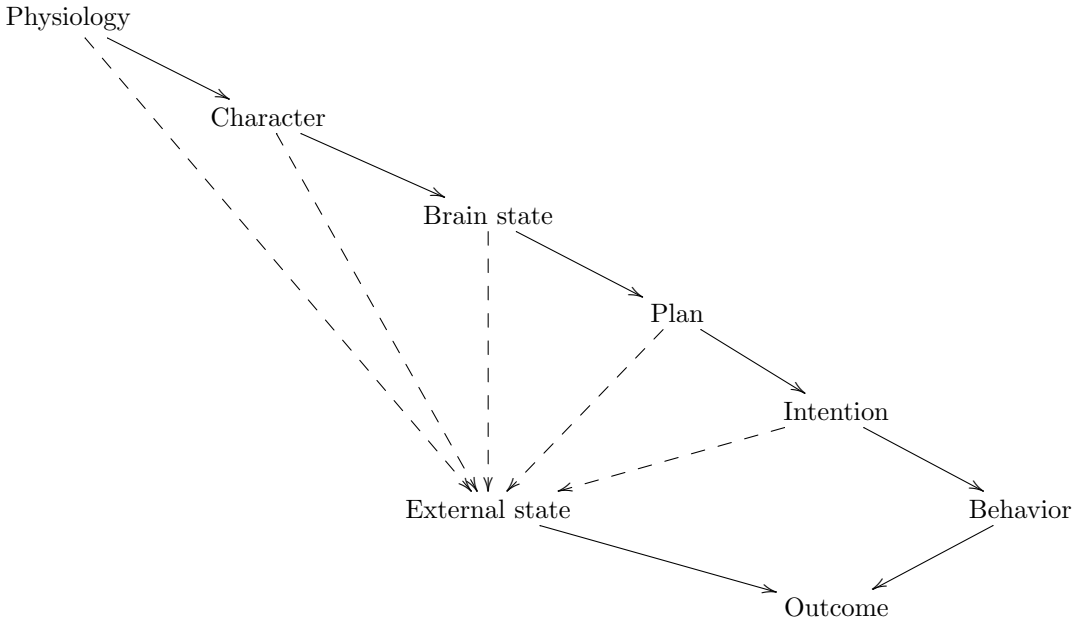
and Parfit's Hitchhiker:



and Prisoner's Dilemma with a Twin:

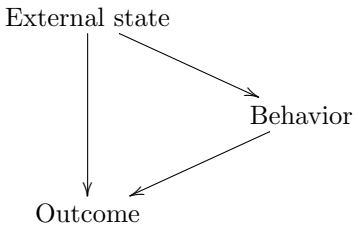


All of these puzzles have been considered Newcomb-like, because all involve some common cause of the external state and the agent's choice, inducing the correlation that underlies the problem. When we just include the one common cause, the diagrams look structurally identical. But when we make the structure more detailed, we can see that all of the variables might fit into different locations in a single larger structure:



The difference between the causal structures of the different examples then comes down to the question of which of the dashed arrows is relevant. There may also be differences in the particular probability distributions by which the dashed arrow explains the causal antecedents of the external state (some of these might be more reliable than others), and the particular causal structure among these various antecedences of the behavior might be more complex than I have indicated here.

Once we consider this causal structure, we can develop other cases where the correlation takes a different form. For instance, one might take the external state to be correlated with the behavior not because of a common cause, but because the external state is itself a cause of the behavior.



This is a plausible description of the structure in The Psychopath Button, and also in the following case from Christopher Hitchcock, whose payoff table is more like the original Newcomb case:

Example 8: Banana

> (Hitchcock, 2016) Suppose that on some mornings, you eat a banana. You don't put any thought into this, nor do you experience any phenomenologically accessible banana-cravings. Your mornings run on auto-pilot, and sometimes you just find yourself eating a banana. You notice that on days when you eat a banana, you often suffer from a painful, migraine-like headache. You occasionally suffer from these headaches when you don't eat a banana, and on those days, the headache is even worse.
>
> You believe the following about the causal structure of this phenomenon: 10% of the time, you wake up suffering from a potassium deficiency. ... You are much more likely to eat a banana on days when you suffer from a potassium deficiency. ... If you suffer from a potassium deficiency, you will suffer from a headache of intensity 10. Eating a banana reduces the intensity by one, and no headache occurs if there is no potassium deficiency.

More complex cases might involve multiple causal arrows that all partly contribute to the correlation. (This is perhaps a way to explain what is going on in the "Meta-Newcomb" puzzle of Bostrom (2001).) As noted by Cartwright (1979), some of these cases might even involve no apparent correlation between act and state, because multiple causal pathways could cancel out (for instance,

the extent to which bananas prevent headaches might exactly counteract the correlation with headaches produced by the potential common cause of potassium deficiancy). There is an interesting distinction between these cases (where act and state have multiple causal pathways connecting them that cancel out) and cases in which there are no causal pathways between the acts and states. The former are violations of the "faithfulness" assumption of Spirtes et al. (2000), and thus are expected to be quite rare (because usually the multiple causal pathways will fail to exactly cancel out in their influence), while the latter can be analyzed by traditional decision-theoretic methods.

Combining a payoff table and a causal structure, we can generate further cases with aspects of the Newcomb structure. Some of these cases may generate different intuitions than the ones considered so far, and some may be useful for testing various decision theories. So at this point it will be useful to turn to the theories.

3 Classifying Decision Theories

Once we have the causal structure, the natural idea suggested by Cartwright (1979) and Hitchcock (2016) is to use this causal structure to understand rationality in terms of effective strategies for getting good outcomes. This represents a rejection of classical "evidential decision theory", which ignores the distinction between conditioning and intervening that Meek and Glymour (1994) use to motivate the structure of causal modeling. However, as Hitchcock notes, thinking in terms of effective interventions doesn't yet fix a unique decision theory. Following Eberhart and Scheines (2007), he notes that interventions on a variable might either be seen as *breaking* the incoming causal links to a variable (which they call "structural" and he calls "hard"), or instead be seen as just *another* causal factor contributing to the probability distribution (which they call "parametric" and he calls "soft"). Furthermore, Hitchcock notes that there is controversy in the Newcomb problem over whether interventions really are *possible* at the level of behavior, given the way the problem is set up.

Once we have looked at the structures underlying the various cases described above, we see a range of variables that could all be plausible sites of intervention in the larger causal model. A person's character, brain state, plans, intentions, and behaviors are all to one degree or another at least partially under her control. However, in many cases, the control an agent has of one of these features goes by means of her control of the others.

Furthermore, the control an agent has of any of these features is often quite far from perfect. Much literature in epistemology questions the extent to which decision-theoretic models could be appropriate for studying belief, because belief is in some sense not under our voluntary control. Consideration of these Newcomb-type cases suggests that perhaps this issue of voluntarism should be central to certain parts of decision theory as well — what sort of voluntary control is needed for our actions to be subject to normative theories, and do we have this sort of control over actions, intentions, plans, mental states, or any of

the other possibly relevant features? Pamela Hieronymi (2006) argues that in fact beliefs and intentions are both out of our control in the same sort of way (distinct from the control that we have over our actions), but she also argues (2008) that this sort of lack of voluntariness is a hallmark of things for which we bear a certain sort of normative responsibility. We control our beliefs and our intentions only *by* directly reasoning about the facts or actions that they are supposed to correspond to in the world. I don't know if this is the right way to think about the sort of control we have over our moral or practical character, or virtues, or other features that are relevant for some of the problem cases. But there is some notion of control here and this might suggest multiple points of intervention in these problems.

3.1 Pluralism versus unity

Once we consider personality, character, mental state, intention, plan, and action all to be sites of potential intervention, there's a question about which, if any, of these is the proper locus of rationality. As long as it is clear that one's genes are not a site of intervention, it will be clear that one should end up smoking in The Smoking Lesion. But for most of the other decision problems, what rationality recommends depends on whether the intervention that is relevant is upstream or downstream from the point of correlation.

A very plausible-seeming "enkrasia" principle is the following. It is rational to plan to do X iff it is rational to do X. Similarly, one might say that the rational character traits to have are the traits that lead one to do rational acts. The rational kind of mental state to have is the kind that leads one to make rational plans. This kind of principle speaks to a kind of unity of rationality. In order to figure out what rationality means in general, we need to figure out which site of intervention is fundamental, and rationality for the others follows from that.

Traditional causal decision theory is often presented in a way that presupposes that rationality for plans and character traits and so on is best evaluated by seeing the momentary actions that they lead to, with only those momentary actions evaluated by the causal interventionist model suggested above. Two-boxing is the rational act in The Newcomb Problem, because an intervention at the last moment to two-box would result in the best outcome. Since two-boxing is rational, this means that the rational plan to make is to two-box, and a rational person is one who two-boxes, and so on. The primary bearer of rationality is the act. The intention, the psychology, the character, etc., are only rational in a sense that is parasitic on the acts that they tend to produce.

There is a kind of tragedy of rationality for this view. Rational people end up with only \$1,000 in The Newcomb Problem, and they end up in a situation of mutual defection in Prisoner's Dilemma with a Twin. They can't get the prize in The Toxin Puzzle and they get abandoned in the desert in Parfit's Hitchhiker. This is all unfortunate, because irrational people (according to the traditional causal decision theorist) often end up much better than the rational ones do. David Lewis (1981) argues that this tragedy is just part of life in an imperfect

world. Sometimes the world rewards the irrational, but that doesn't mean that they're rational.

If the moment of action were the only conceivable point of intervention, this would seem very plausible. But because other points of intervention are possible (and because many of our individual actions seem to be controlled at least in part by habit or other long-term states rather than by direct deliberation), some theorists have taken this tragedy to be a sign that traditional causal decision theory has things backwards. Rather than looking at the most momentary and specific point of intervention, we should look at the farthest *upstream* point of intervention, and define rationality from that.

This is often cashed out through a metaphor of taking the goal of rationality to be the design of the best overall algorithm for reacting to decision problems, treating the self almost as a robot. (Yudkowsky, 2010) Where the traditional causal decision theorist says that rationality is about being the kind of person that does in the moment, the best thing to do in that moment, this sort of view says that rationality is about doing in the moment what the most successful kind of person does. One should be the kind of person that gets the prize if one can, even if that means leaving money on the table (literally, in the case of The Newcomb Problem). If one rationally formed plans or intentions or character traits in the past, one should go along with them in the present. Even if one *hadn't* formed these plans or intentions (perhaps one finds oneself in a situation analogous to one of the ones above, without any advance warning), one should act *as if* one had.

This sort of impulse exists in many related philosophical domains. Process reliabilism (Goldman, 1979), virtue epistemology (Turri et al., 2017), Kantian ethics, and rule utilitarianism (Hooker, 2015) all take this kind of form. A belief is evaluated for rationality not by evaluating the belief itself for truth, but instead by evaluating the process or epistemic virtues that gave rise to it for their truth-aptness. An act is evaluated for morality not by evaluating its own consequences, but by evaluating the consequences of the maxim or rule that gave rise to it when considered as a guide in general. Although it is still the consequences at the end of the stream that guide the evaluation, the act or belief in the moment is evaluated on the basis of having been generated by some upstream feature that could give rise to good consequences if generally followed.

However, I am not convinced that rationality has the kind of unity assumed by each of these views. Each seems to ignore the fact that there are *many* possible points of intervention, and that even the person who has made the right intervention at one point could still do better by also intervening at another. There is no coherent way to be such a person in general (a general inclination to intervene at the further downstream points would itself be a character trait that contradicts the upstream trait). Instead, I suspect there is just a pluralism of rationalities. There is one notion of rational act, and another notion of rational plan, and another notion of rational character.

The pluralist way sees a kind of dilemma, or even essential tragedy, for rationality. The rational act may not be the one that would be performed by

an agent with the rational character. A rational agent should choose to *become* such that she can do no other than pick one box in The Newcomb Problem, pay up in Parfit's Hitchhiker, and drink the toxin in The Toxin Puzzle. But it would be rational to *violate* this character trait if she could. There is one question of what kind of person to be, and another question of what to do, and the answers to these questions just unfortunately don't cohere in Newcomb-like cases. If we had perfect control over momentary actions, then maybe we should agree with the traditional causal decision theorist about rationality — but then there would be no such thing as general character traits that we had any control over. If we had perfect control over the kind of person we are, then maybe we should agree with Yudkowski and others — but then there would be no possibility of changing one's mind at the last moment. These problems arise only when there are multiple points that all have some claim of being the point at which we can intervene.

4 Bigger Questions

The viewpoint I have described on Newcomb-type problems categorizes them by the payoff table of potential outcomes, and the causal structure behind the correlation between act and state. It categorizes the decision theories based on which one or more theoretically possible points of intervention are relevant for evaluation of rationality. I have not addressed the question of how to deal with uncertainty about the causal structure. (Most traditional problems stipulate that the relevant agent knows for sure what is going on.) And I have not addressed the question of how to determine what the causal structure is. But this is a major part of the general problem of induction, and it is a lot to ask of a decision theory to answer that question!

The various understandings of "intervention" in these causal models raises some further questions for the notion of rationality in these cases. On one picture (often associated with Pearl (2000)), intervention must come from outside the model and break the structure that exists. This is a natural point of view for models that are inherently incomplete, representing the causal structure of only part of the world. This is often appropriate for an understanding of experimental design, where the experimenter can be treated as an aspect of the universe that is outside the scope of the model, and can break its relations in some cases. But it may not be appropriate for a theory that is supposed to tell us something about the normative character of rationality, particularly if the cases (like The Newcomb Problem, The Toxin Puzzle, Prisoner's Dilemma with a Twin, and even Banana) are ones where it is assumed that *all* relevant causal structures are included, *including* ones that act through the mental states of the agent.

On another picture (often associated with Spirtes et al. (2000)), an intervention is itself represented by a node within the causal model, that has the property of screening off its children from their other parents in the settings in which it is active. However, if we get a complete causal model of the world, it seems likely that any node will fail to perfectly exemplify these properties, and

thus will only at best approximately count as an intervention. When we combine this with the fact that none of the sites of intervention are totally effective (particularly in Newcomb-like problems), it looks like some of the foundations of a traditional conception of rationality start to crumble.

On either picture, the apparent adequacy of a conception of intervention requires the model to be incomplete. But a truly normative picture of rationality that can deal with cases stipulated as precisely as some philosophers like to discuss, seems like it should work best when we have a complete model of the world. If this causal modeling structure is the right way to proceed, then the Newcomb-type problems may point to an essential incompleteness or approximation in our concept of rationality, even beyond the tragedy of the incompatible requirements at different levels that I mentioned above.

References

Arntzenius, F. (2010). Reichenbach's common cause principle. *Stanford Encyclopedia of Philosophy*.

Bostrom, N. (2001). The meta-Newcomb problem. *Analysis*, 61(4):309–310.

Cartwright, N. (1979). Causal laws and effective strategies. *Noûs*, 13(4):419–437.

Eberhart, F. and Scheines, R. (2007). Interventions and causal inference. *Philosophy of Science*, 74(5):981–995.

Eells, E. (1982). *Rational Decision and Causality*. Cambridge University Press.

Egan, A. (2007). Some counterexamples to causal decision theory. *Philosophical Review*, 116(1):93–114.

Gibbard, A. and Harper, W. (1978). Counterfactuals and two kinds of expected utility. In Harper, W., Stalnaker, R., and Pearce, G., editors, *Ifs*, pages 153–190. Dordrecht: Reidel.

Goldman, A. (1979). What is justified belief? In Pappas, G. S., editor, *Justification and Knowledge*, pages 1–25. Reidel.

Hieronymi, P. (2006). Controlling attitudes. *Pacific Philosophical Quarterly*, 87(1):45–74.

Hieronymi, P. (2008). Responsibility for believing. *Synthese*, 161(3):357–373.

Hitchcock, C. (2016). Conditioning, intervening, and decision. *Synthese*, 193(4):1157–1176.

Hooker, B. (2015). Rule consequentialism. *Stanford Encyclopedia of Philosophy*.

Jeffrey, R. (1981). The logic of decision defended. *Synthese*, 48(3):473–492.

Kavka, G. (1983). The toxin puzzle. *Analysis*, 43:33–36.

Lewis, D. (1981). Why ain'cha rich? *Noûs*, 15(3):377–380.

Meacham, C. (2010). Binding and its consequences. *Philosophical Studies*, 149(1):49–71.

Meek, C. and Glymour, C. (1994). Conditioning and intervening. *British Journal for the Philosophy of Science*, 45:1001–1021.

Parfit, D. (1984). *Reasons and Persons*. Oxford University Press.

Pearl, J. (2000). *Causality: models, reasoning, and inference*. Cambridge University Press.

Price, H. (1986). Against causal decision theory. *Synthese*, 67:195–212.

Savage, L. J. (1954). *The Foundations of Statistics*. Dover.

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, prediction and search*. MIT Press.

Turri, J., Alfano, M., and Greco, J. (2017). Virtue epistemology. *Stanford Encyclopedia of Philosophy*.

Yudkowsky, E. (2010). Timeless decision theory. Technical report, The Singularity Institute.