Understanding the Tickle Defense in Decision Theory

Caspar Oesterheld, Duke University
caspar.oesterheld@duke.edu

first written in 2018, last updated January 11, 2022

1 Introduction

Following the publication of Newcomb's problem (Nozick, 1969), decision theorists have started to differentiate between two forms of expected utility maximization – evidential and causal decision theory (EDT and CDT) – though in recent years many other theories have been proposed[1]. Roughly speaking, EDT recommends to choose the action that, given all our knowledge, provides the highest "news value", i.e., to choose based on the expected utility conditional on our beliefs and our action. In contrast, CDT only considers what our action brings about causally given our beliefs. In Newcomb's problem, EDT one-boxes, reasoning that if the agent one-boxes, box B must contain a million dollars. CDT two-boxes reasoning that it cannot causally affect the content of box B.

EDT is conceptually simpler than CDT and its decision in Newcomb's problem seems defensible[2]. However, it has been alleged that EDT makes unsound recommendations in other decision problems, most notably so-called medical Newcomb problems. Roughly speaking, critics have argued that if a disease causes some symptomatic behavior (like stroking cats, or going to the doctor), EDT irrationally refrains from exhibiting that behavior as doing so would be evidence of having the disease. In contrast, CDT reasons that, because disease causes symptom and not vice versa, one should exhibit the symptom whenever doing so conveys other benefits. For instance, if stroking a cat were a symptom of a painful and deadly disease

[1] I attempt to give a comprehensive list at https://casparoesterheld.com/a-comprehensive-list-of-decision-theories/. It is also worth noting that even prior to Nozick's 1969 paper, two different theories of Bayesian decision making had been developed: Savage's theory and the theory developed by Jeffrey and Bolker (Ahmed, 2014, Sections 1 and 2; Steele and Stefánsson, 2020, Section 3). The Jeffrey-Bolker theory is clearly the archetypical form of EDT. Savage's theory (which is much more widely known) uses a very causalist formalism, although EDT (and other theories) can also be formalized as instances of Savage's general theory.

[2] Newcomb's problem famously divides people, with one-boxing being sligthly more popular than two-boxing, see https://casparoesterheld.com/2017/06/27/a-survey-of-polls-on-newcombs-problem/ for an overview of surveys and polls on Newcomb's problem. Among professional philosophers and decision theorists, two-boxing has become significantly more popular than one-boxing. However, this is presumably in great part due to what is seen as the result of generalizing one-boxing to (e.g., via EDT) the medical Newcomb-like problems discussed in this paper (cf. Talbott, 1987, p. 452f.). A testament to this hypothesis is that many authors have proposed variants of causal and evidential decision theory that one-box in Newcomb's problem but behave like CDT in medical Newcomb-like problems (Spohn, 2003; Spohn, 2012; Poellinger, 2013; Yudkowsky and Soares, 2018). If the tickle defense argument as defended in this paper goes through, then this reason against one-boxing disappears. For a general discussion of EDT versus CDT and whether to one-box in Newcomb's problem, see Ahmed (2014). I also have my own draft paper arguing against CDT (and other theories that two-box), see Oesterheld and Conitzer (2019).

and you had some small desire to stroke a cat, CDT would stroke the cat, whereas EDT recommends refraining to obtain evidence of being healthy – or so critics have argued.

In this paper, I give an introduction to the primary counter-argument to this argument against EDT: the tickle defense. Roughly speaking, the tickle defense presented here goes as follows: If, say, stroking cats is a symptom of some disease, then this must be because the disease causes some kind of desire or "tickle" to stroke cats. A rational agent should know of such tickles influencing her decision. But once the tickle is observed, stroking the cat provides no further evidence that the agent has contracted a cat-stroke-desire-causing disease. Hence, EDT concurs with CDT in recommending to stroke cats.

The rest of this paper is structured as follows:

- In Section 2, I introduce causal graphs as described by Pearl (2009), which I will use as models of decision problems. Causal graphs are non-standard in philosophical decision theory but allow us to represented the internal structure of medical Newcomb problems in an intuitive, graphical way.
- In Section 3, I define causal and evidential decision theory on the basis of causal graphs.
- In Section 4, I introduce medical Newcomb problems and the problem they allegedly pose for evidential decision theory.
- In Section 5, I describe the tickle defense. I argue that it applies to so-called medical Newcomb-like problems but not to problems like Newcomb's problem itself or the prisoner's dilemma against a copy.
- Finally, I give an overview of alternative versions and critiques of the tickle defense (Section 6).

If you have any questions or comments, feel free to contact me!

2 Causal graphs

In this paper, I will represent decision problems by causal graphs as described and extensively discussed by Pearl (2009) (though I will introduce a few modifications). In this section I give a very brief introduction to the most important ideas behind causal graphs for the purpose of this paper. For instance, I will not talk about how causal relationships can be inferred or about causality's metaphysical status. For broader discussions of causal graphs, see Pearl's book. For a more general philosophical discussion of causality and its metaphysical status, see, e.g., Price and Corry (2007) or Schaffer (2016). (Pearl's causal graphs and the *do*-calculus are an example of an interventionist account of causality.)

Causal graphs are basically what their name suggests: graphs of variables representing events in the world with arrows between variables representing causal relationships. As an example, consider the causal graph in Figure 1, which states that which season it is causally affects whether the sprinkler is on and whether it is raining, both of which causally affect whether the grass is wet, which in turn affects whether the grass is slippery.

Each node of the causal graph comes with a probability distribution giving the probability of each value of that node for each combination of the values of the predecessors of that node. For example, in the causal graph in Figure 1 the "wetness" node comes with values for the following probabilities:

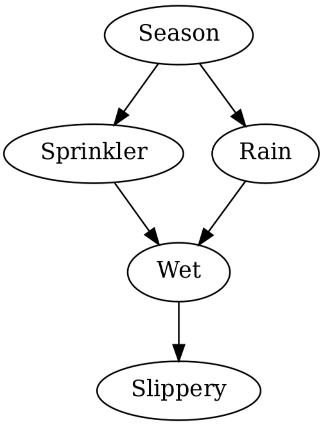

Figure 1: An example of a causal graph from `http://lesswrong.com/lw/ev3/causal_diagrams_and_causal_models/`, which in turn is based on Pearl (2009, fig. 1.2).

$$\begin{gathered} P(\text{dry} \mid \text{no rain}, \text{sprinkler off}) \\ P(\text{dry} \mid \text{no rain}, \text{sprinkler on}) \\ P(\text{dry} \mid \text{rain}, \text{sprinkler off}) \\ P(\text{dry} \mid \text{rain}, \text{sprinkler on}) \end{gathered}$$

Nodes without a predecessor – in this case only Season – come with a prior probability distribution over the possible values of that node. These probabilities make causal graphs a special case of Bayesian network.

From a causal graph (or any other Bayesian network for that matter) we can infer any of the regular conditional probabilities we might be interested in. For example,

$$\begin{aligned} P(\text{rain} \mid \text{sprinkler on}) &= P(\text{winter} \mid \text{sprinkler on})P(\text{rain} \mid \text{winter}) \\ &\quad +P(\text{summer} \mid \text{sprinkler on})P(\text{rain} \mid \text{summer}) \\ &= \frac{P(\text{sprinkler on} \mid \text{winter})P(\text{winter})}{P(\text{sprinkler on})}P(\text{rain} \mid \text{winter}) \\ &\quad +\frac{P(\text{sprinkler on} \mid \text{summer})P(\text{summer})}{P(\text{sprinkler on})}P(\text{rain} \mid \text{summer}), \end{aligned}$$

where the last line is an application of Bayes' theorem and

$$P(\text{sprinkler on}) = P(\text{sprinkler on} \mid \text{winter})P(\text{winter}) + P(\text{sprinkler on} \mid \text{summer})P(\text{summer}).$$

Intuitively speaking, if we learn that the sprinkler is on, this updates our beliefs about about what season it is, which in turn updates our beliefs about whether it is raining. Note that we are hereby making inferences in the opposite direction of causal flow. If we learn a value of a node, this tells us something about that node's causal predecessors. For the general (fairly simple) rules for making inferences from observed values of variables in Bayesian networks, see, e.g., Pearl (2009, Section 1.2), Russell and Norvig (2010, ch. 14) or Ben-Gal (2007). Note that the rules are no different between causal versus other Bayesian networks.

For the tickle defense, we will need to consider some rules under which learning about one node does not provide any information about some other node. The general criterion – again in both causal graphs and more generally Bayesian networks – is called *d-separation* (see Pearl, 2009, Section 1.2.3). For the purpose of understanding the tickle defense, we only need some basic understanding of this concept. Intuitively, if learning the value of one node is to provide information about the value of another there has to be a path between the two nodes that is, in some sense, "unblocked". For example, there is a path Sprinkler ← Season → Rain that allows us to draw an inference from whether the sprinkler is on about whether it is raining. But imagine we already knew which season it was and were only then told that the sprinkler is on. Then knowledge of the season would block the path from Sprinkler to Rain and so the sprinkler being on would provide no further evidence on whether it's raining.[3]

We now come to a feature that is exclusive to causal graphs (i.e., one that does not make sense in all Bayesian networks): Pearl's *do*-calculus. (Since we will mainly be concerned with EDT's choices, we will only cover it briefly.) The *do*-calculus asks the question: how do the probabilities of nodes in the causal graph change if we intervene from the outside to set the value of a particular node? For example, let us say the probability distribution attached to the Sprinkler node in Figure 1 describes the behavior of some automated system. Then we could ask questions such as: what is the probability that it's winter given that I come from outside the system and manually turn on the sprinkler? In Pearl's formalism, what is $P(\text{winter} \mid do(\text{sprinkler on}))$?

Pearl proposes to assign probabilities conditional on $do(x)$ for a variable X according to the following procedure: remove all ingoing arrows of X and replace the probability distribution attached to X with one where $P(x) = 1$. Then assign probabilities according to this new causal graph. For example, in the causal graph of Figure 1 it is

$$P(\text{winter} \mid do(\text{sprinkler on})) = P(\text{winter}).$$

This illustrates the main point of the do-calculus: avoiding inference about predecessors – turning on the sprinkler gives you very different information about the world than observing passively that the sprinkler is on. In particular, if you turn the sprinkler on yourself, you learn nothing about the other potential causes contributing to whether the sprinkler is on.

3 Evidential and causal decision theory

Using causal graphs and the *do*-calculus we can now define evidential and causal decision theory (EDT and CDT), the two main contenders for normative theories of rational choice in Newcomb's problem.

Let's say we are given some causal graph. For the graph to represent a decision problem it has to contain an action node A and a payoff node U.[4] Then EDT chooses

$$\arg\max_a \mathbb{E}\left[U \mid A = a\right], \tag{1}$$

where the arg max is over possible values of A. Evidential decision theory evaluates an action a just as it would update its expectation of U upon learning that action a was taken. Hence it

[3] Somewhat counterintuitively, the *lack* of knowledge can block paths, too. For instance, the reason why the path Sprinkler → Wet ← Rain does not render Sprinkler evidentially relevant to Rain is that not knowing the value of Wet blocks this path. If it were known, for instance, that the grass is wet, then even if it is also known that it is, say, summer, the sprinkler being on would provide evidence on whether it's raining.

[4] Of course, in real world decision problems the payoff will rarely be some node in our model of the world. Instead, it is something that we assign to some set of nodes. However, for the purposes of this paper it does not matter whether the utility is a node caused by some set of nodes or a value assigned based on the values of this set of nodes. We here opt for the former because it is notationally simpler.

is said that EDT evaluates the "news value" of a. This also means that to use EDT we do not need to represent our decision problem as a causal graph. Any Bayesian network (or any other representation of a joint probability distribution) will do.

CDT as described by Pearl (2009, Section 4.1.1) chooses

$$\arg\max_{a} \mathbb{E}\left[U \mid do(A=a)\right]. \tag{2}$$

CDT evaluates an action as one would evaluate that action if it was taken as the result of outside intervention. That is, for any action a, CDT asks: what would happen if some outside force where to ensure that I take action a.

To demonstrate EDT's and CDT's behavior, I would like to use the original Newcomb's problem as an example. Unfortunately, causal graphs are limited in what they can model and these limitations become especially obvious in the context of Newcomb-like problems. The problem is that causal graphs can only express dependences between two variables that arise from some causal relationship (either one causally affecting the other or a third causally affecting both). Hence, their ability to model all situations hinges on Reichenbach's common cause principle, which states that *all* dependences result from causal relationships (Arntzenius, 2010).

Newcomb's problem and many problems like it are examples where Reichenbach's common cause principle fails. Here, from the perspective of the agent, the prediction and the agent's decision are dependent; learning about one provides information about the other. (For example, two-boxing provides information that box B is probably empty.) But this dependence need not arise from a causal relationship. Of course, neither of the two causes the other. But they are not solely the result of a common cause either. Imagine that all the causes of your action and all the causes of the prediction were known, e.g., the state of your brain before making a decision and the model of your brain used by the predictor. To an outside, logically omniscient entity, prediction and action could then be predicted with certainty and would hence be independent. But if you are, say, a human facing the choice you will usually not know the result of your reasoning or the result of the prediction – you are logically uncertain (see Garrabrant et al., 2016, and the references in its Section 1.2) about both, but believe that they will likely be the same.[5] This means action and prediction will remain dependent because your action provides evidence on what the predictor's model implies.[6]

There are a few ways to respond to this problem. For instance, we could extend causal graphs with a new kind of arrow that represents non-causal dependence (see, e.g., Poellinger,

[5] Yudkowsky (2010, Chapter 11) uses a similar example:

> Suppose that I place, in Mongolia and Neptune, two calculators programmed to calculate the result of 678 × 987 and then display the result. As before, the timing is such that the events will be spacelike separated – both events occur at 5PM on Tuesday in Earth's space of simultaneity. Before 5PM on Tuesday, you travel to the location of both calculators, inspect them transistor by transistor, and confirm to your satisfaction that both calculators are physical processes poised to implement the process of multiplication and that the multiplicands are 678 and 987. You do not actually calculate out the answer, so you remain uncertain of which number shall flash on the calculator screens. As the calculators are spacelike separated, it is physically impossible for a signal to travel from one calculator to another. Nonetheless you expect the same signs to flash on both calculator screens, even though you are uncertain which signs will flash.

Here, too, the two calculation outcomes are dependent for a logically non-omniscient agent, but this dependence does not result from a common cause.

[6] Perhaps *all* non-causal dependences are logical dependences? Ahmed and Caulton (2014) have argued that under some interpretations of quantum mechanics and some views about causality, there are other cases of non-causal dependence involving quantum entanglement (cf. Ahmed, 2014, ch. 6).

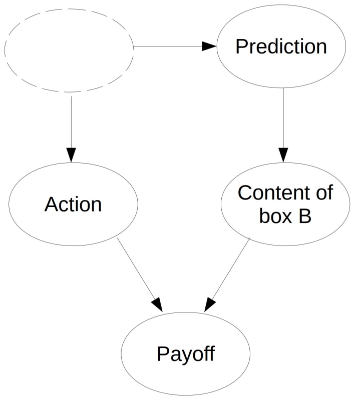

Figure 2: A pseudo-causal graph representing Newcomb's problem.

2013). But since we can use proper causal graphs for understanding the tickle defense itself, we will use a more ad hoc solution. That is, we will simply represent non-causal dependences by introducing an imaginary "logical" common cause. As an example, a representation of Newcomb's problem is given in Figure 2. We may even interpret this common cause as the logical fact about the output of our decision mechanism, although logical facts are, of course, usually not regarded as things that can be causes (see Schaffer, 2016, Section 1).[7]

Let us briefly describe how evidential and causal decision theory reason about this problem in this causal graph. EDT reasons that if the agent one-boxes the common cause of action and prediction will (likely) be such that the prediction will (likely) also be that the agent one-boxes. Thus, if the agent one-boxes, box B will (likely) contain a million dollars and the payoff will (likely) also be a million. By the same line of reasoning, EDT estimates the payoff of two-boxing to be lower than a million. Thus, EDT one-boxes.

CDT, on the other hand, reasons that by whatever action the agent takes, all ingoing arrows are severed. So, neither the intervention to one-box nor the intervention to two-box tells the agent anything about the prediction or the content of box B. Since two-boxing is better regardless of the content of box B, two-boxing is also the better intervention for any distribution over contents of box B. Thus, CDT two-boxes.

4 Medical Newcomb-like problems

We are now ready to introduce medical Newcomb-like problems – a class of decision problems that allegedly pose a devastating problem to EDT while being satisfactorily solved by CDT. We here use the classic problem discussed in this context: the Smoking Lesion (Egan, 2007, Section 1) (to my knowledge first introduced as Fisher's problem by Jeffrey (1965/1983, p.

[7] Essentially our approach could be described as modifying our conception of causality to make Reichenbach's principle true. Several decision theorists have proposed similar modifications. But where we have tried to mimic standard CDT in causal graphs, they have tried to generate alternatives to CDT. For instance, Yudkowsky (2010) proposes that the agent should view itself as choosing the logical fact that causes the physical action. His variant of CDT thus one-boxes in Newcomb's problem. Spohn (2012), on the other hand, argues that non-causal dependences should be represented as outgoing (rather than ingoing) arrows.

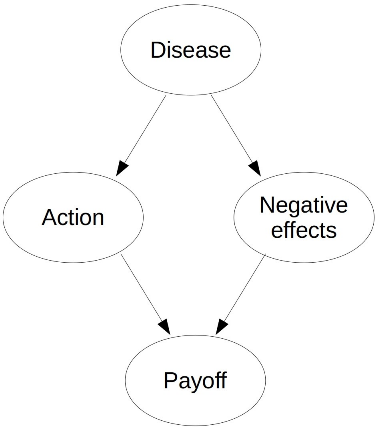

Figure 3: A causal graph representing medical Newcomb-like problems.

15).[8] Imagine that contrary to fact, smoking doesn't cause cancer. Instead, smoking and cancer have some common cause. For instance, there might be a disease (or gene) which both causes people to smoke and to get cancer. Smoking itself is assumed to have no negative effects. Thus, once you know whether someone has cancer or merely the smoking/cancer gene, it would not be bad news to see the person smoking. We also assume that smoking provides some amount of pleasure. A causal graph representing this type of decision problem is given in Figure 3. In this hypothetical situation, should you smoke or not?

When deciding based on this causal graph whether to smoke, EDT recommends reasoning as follows: if I smoke, then this gives me evidence that I have the disease. This increase in the probability of having the disease outweighs – via the negative effects of the disease – any pleasure I might derive from smoking. Thus, EDT recommends refraining from smoking in order to produce evidence that one doesn't get cancer.

CDT, on the other hand, reasons that it cannot causally affect whether it has the disease and smokes since this brings some additional pleasure regardless of the probability that one has the disease.

Arguably, EDT's behavior is problematic. After all, the reasoning can be generalized from cancer and smoking (in this hypothetical scenario) to all diseases with some harmless symptoms or other signs like going to the doctor. And in most of these cases, people agree that one should not try to refrain from showing symptoms.

Medical Newcomb-like problems have led some – particularly people who aren't familiar with the decision theory literature – to conclude with high confidence that EDT cannot be used to guide rational decision making. For example, Pearl (2009, Section 4.1.1) writes (Pearl, 2021):

> The paradoxes that emerge from this fallacy [EDT] are obvious: patients should

[8] Pedagogically, the Smoking Lesion is suboptimal because it asks us to make contrary-to-fact assumptions about a real-world decision problem. Many of the other problems of this type that have been proposed are superior in this regard. (Indeed, many of these other problems have been proposed to overcome the issues of the Smoking Lesion, see, e.g., Altair (2013, Section 3).) However, I chose to discuss the Smoking Lesion because it will better enable the reader to understand other papers, in which the Smoking Lesion will be the most likely point of reference.

> avoid going to the doctor "to reduce the probability that one is seriously ill" (Skyrms, 1980, p. 130); workers should never hurry to work, to reduce the probability of having overslept; students should not prepare for exams, lest this would prove them behind in their studies; and so on. In short, all remedial actions should be banished lest they increase the probability that a remedy is indeed needed.

And:

> I purposely avoid the common title "causal decision theory" in order to suppress even the slightest hint that any alternative, noncausal theory can be used to guide decisions.

While most people avoid such provocative rhetoric, many people whom I have talked to (again, mainly those with a more casual interest in Newcomb's problem) have expressed similar sentiments. Can EDT be defended?

5 The tickle defense

The tickle defense is an attempt to argue that EDT (like CDT) recommends smoking in the Smoking Lesion if it is supplied with all the knowledge available. Many versions of the defense have been proposed. They mainly differ in terms of the extent to which they claim to align EDT's with CDT's decisions. The present version is similar to that given by Ahmed (2014, Section 4.3). See Section 6 for more details and references.

Assuming introspection The tickle defense is based on the following premise: the agent knows all the immediate causes of its choice. (Here, by "choice" I mean that which is chosen as determined by the decision theory. Of course, the action that the agent ends up physically implementing may additionally be affected by uncontrollable parts of its brain, hardware malfunctions (in case of a robot), etc. that the agent is not aware of, see Figure 4. From now on, we will ignore this distrinction and use all of "choice", "decision" and "action" to refer to what is fully under control of the agent's wills.)

Why might this premise be true? Consider that Bayesian decision theories like EDT and CDT make their decisions solely based on their beliefs and their preferences. Thus, the only causes of a Bayesian decision theorist's choice are the agent's beliefs, the agent's preferences, and the agent's decision theory. Knowing these three, the agent's choice is uniquely determined by lines 1 or 2 and therefore cannot have other immediate causes, see Figure 5.

For the same reasons that beliefs and preferences are causes of the agent's decision, the agent also has to have some access to these beliefs and preferences. Otherwise, it could not perform expected value calculations. It also seems plausible that the agent knows its decision theory. For instance, if humans try to follow the reasoning of some Bayesian decision theory, they, of course, have to know which theory they are using. Furthermore, if we were to build an agent ourselves according to, say, EDT, it seems useful to provide that piece of additional information to the agent.

More generally, we could consider any other agent design that makes decisions based on some class of "reasons" and then argue that since the agent decides based on these reasons, she knows the reasons influencing her.

Despite the plausibility of this *introspection assumption*, it should nevertheless be noted that it poses an assumption that not every conceivable reasoner satisfies. Looking at Figure 5,

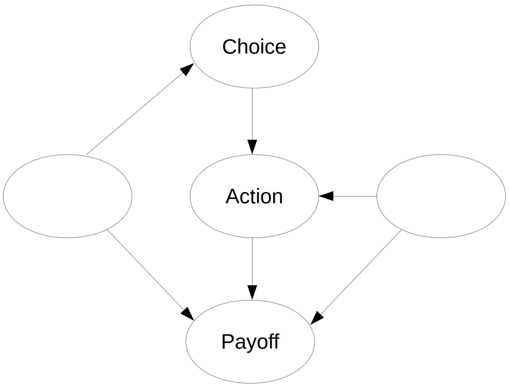

Figure 4: A decision problem in which the agent's final action is determined not only by the agent's choice but additionally by some outside influence.

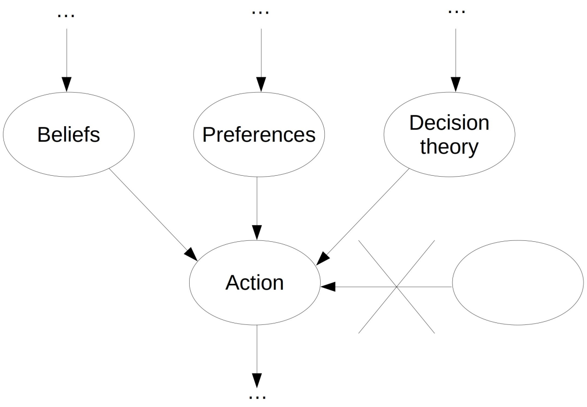

Figure 5: A Bayesian decision theorist's action can only have three (classes of) causes: her beliefs, her preferences and her decision theory.

it is perfectly possible that when computing the decision theory's choice the agent has access to its beliefs, preferences and decision theory, but that at the same time the beliefs themselves do not describe the beliefs, preferences and decision theory. Indeed, specifying beliefs that describe themselves in addition to other parts of the world is an additional challenge, albeit one that is easily solved for artificial agents, who can simply be enabled to open their own source file.[9]

Assuming knowledge of the mechanism Let us now return to medical Newcomb problems and in particular the Smoking Lesion. Since an EDT agent's choice is determined by its beliefs, preferences and decision theory, the common cause of cancer and smoking – e.g., the smoking/cancer gene – must cause this through one of those three. Besides the introspection assumption, we assume that the agent knows how the cancer gene causes smoking or at least has strong suspicions. For example, in the Smoking Lesion it seems most plausible that there is a gene which causes both cancer and a (stronger than usual) desire to smoke. In Pearl's (not-)going-to-the-doctor case, it is even clearer what the causal mechanism must be: being ill causes one to believe that one is ill; believing one is ill causes one to go to the doctor. In Section 6.5, we will discuss medical-like cases in which the agent does *not* know (or have strong beliefs about) how the causal mechanism works.

The tickle defense Given the introspection assumption and assuming knowledge of the causal mechanism of the Smoking Lesion, our agent will not use a model like that in Figure 3 but that of Figure 6. In this new problem it becomes clear that if the agent already knows its preferences, EDT recommends smoking – knowledge of the preference blocks the path from action to disease and so the action can provide no further evidence on whether the agent has the disease or not. We say that knowledge of the preferences *screens off* the action from whether the agent has the disease.

As another example, consider again the case of (not) going to the doctor. Imagine that you are generally healthy, but this morning you woke up with abdominal pain. Assuming that your condition is sufficiently severe, should you make an appointment with the doctor? Common sense dictates that you should and Pearl alleges (see Section 4) that EDT sometimes prescribes that you should avoid going to the doctor because going to the doctor gives you evidence that you are ill. There is some truth to the claim: if you called your friend and told her that you were going to the doctor without telling them about your abdominal pain, they would certainly take your trip to the doctor as evidence that you are sick. In this sense it is true that going to the doctor is evidence that you are sick. However, if you first told your friend in detail about your symptoms (and they believe that you are relating them truthfully), they wouldn't take your going to the doctor (or not) as further evidence of whether you are sick. Similarly, when you decide whether to go to the doctor, you already know what symptoms you have and therefore your decision as to whether to go to the doctor doesn't give you any evidence about whether you are sick. Hence, EDT recommends going to the doctor. As in the Smoking Lesion, being sick causes you to go to the doctor. This time, it does so by changing your beliefs (rather than your desires) – being sick causes you to know that you have symptoms, in this case abdominal pain. Thus, knowledge of your symptoms screens off your decision to go to the doctor from whether you are healthy.

[9] Treutlein (2018) has argued that through its reliance on the introspection assumption the present version of the tickle defense creates a difference between decision theory for humans and decision theory for decision-making machines. The latter can easily be built to satisfy the introspection assumption, while the former may not always satisfy it.

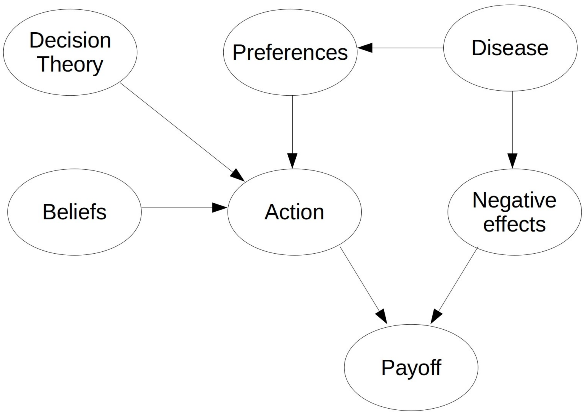

Figure 6: A medical Newcomb-like problem in which the agent's model explicitly contains her beliefs, preferences and decision theory.

We can generalize the argument. If action and payoff have a common cause, then this common cause must affect the action via some "tickle", e.g., an urge to smoke. For that tickle to be effective, the agent has to be aware of whether the tickle is present (this is the introspection assumption). But once this is the case, the agent's action can provide no further evidence on the common cause. Hence, if the decision problem is modeled by a causal graph and if the agent knows all causes of its actions (or at least the ones that share a common cause with the payoff), then EDT and CDT assign the same expected values to each action. I hope the intuitive reason for this is now clear – formally it follows directly from Theorem 3.4.1, Rule 2 of Pearl (2009, Section 3.4.2).

The tickle defense, as presented here, does not apply to Newcomb's problem, the Prisoner's Dilemma against a copy and so forth. Let's again model Newcomb's problem with causal graphs that allow for logical facts as causes and let's add beliefs, preferences and decision theory to our model to obtain the model of Newcomb's problem given in Figure 7. Here, the correlation between prediction and action is explained by both being influenced by the logical fact as to what particular decision theories do in particular situations. For instance, if the agent uses EDT and the standard beliefs and preferences about Newcomb's problem, then these together with the logical fact that EDT one-boxes cause the action to be one-boxing. The predictor's prediction about the EDT agent, too, depends on this logical fact.

In this model, the introspection assumption is false, in the sense that not all causes of the agent's choice are known. When an agent decides whether to one- or two-box, it cannot already know that in this particular situation its decision theory implies one-boxing (or two-boxing). Otherwise, the agent would not face a choice at all and at least EDT could not make any recommendations since it cannot assign expected utility to actions that have a probability of 0.

In sum, the tickle defense presented here is an argument that with a sufficient but plausible degree of introspection, EDT takes the intuitively plausible action and the one that CDT recommends in "medical" Newcomb-like problems. It does not show that EDT and CDT agree in all cases. Hence, the tickle defense has no bearing on whether EDT's recommendation in logical dependence-based Newcomb-like problems are correct. If you find one-boxing in Newcomb's problem and cooperating against a copy appealing (or at least don't see these as

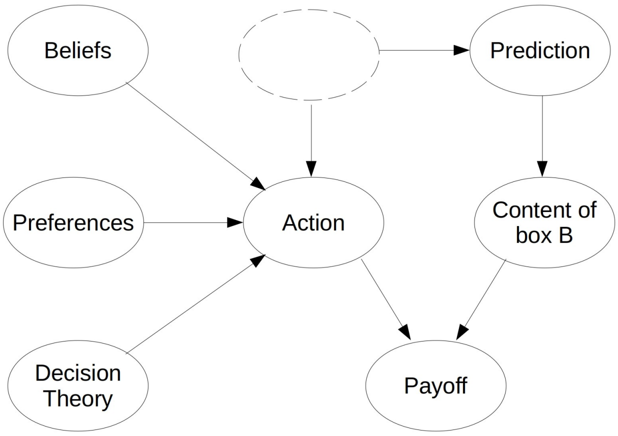

Figure 7: A pseudo-causal graph representing a Newcomb-like problems in which the dependence between action and prediction is logical. That is, even if the agent knows her beliefs, preferences and decision theory, taking one one action or another still provides her with information about what these beliefs, preferences and decision theory imply and hence about what the predictor will predict the agent to do.

refutations), but do not buy into refraining from smoking lest you gain evidence of having the smoking lession, then the tickle defense presented here is a strong argument in favor of EDT.[10]

6 Bibliographical notes and further considerations

As noted above, this version of the tickle defense is based on Ahmed (2014, Section 4.3). Over the years, very similar versions have also been proposed by Skyrms (1980, pp. 130-132), Horgan (1981, Section VIII), Jeffrey (1983, p. 25), Eells (see Section 6.2), Albert and Heiner (2001, Appendix B), Almond (2010, Section 2.8), Horwich (1987, Section 11.2), Yudkowsky and Soares (2017, Section "The Smoking Lesion Problem") and arguably even Nozick (1969, fooootnote 11). I will close by giving an overview of the other literature on the tickle defense. In particular, I will try to explain how other accounts and critiques of the tickle defense differ from the one given in this paper and why I favor the tickle defense as presented here.

Whereas I believe that the previous sections of this paper are of interest to anyone curious about decision theory and Newcomb's problem, the target audience of this section is much more limited.

6.1 Skyrms

Skyrms (1980, p. 130-132) – who, to my knowledge, was the first to discuss the tickle defense in published writing – argues that a decision maker cannot know all the causes of her action:

> If the utilities, probabilities, and decision rule determine the action, then it is tantamount to taking which action will be performed as a *datum* in the evaluation

[10] Cf. Yudkowsky (2010, Section 9.1) who writes: "What we would ideally like is a version of the tickle defense that lets an evidential theorist [smoke], and also take only box B in Newcomb's Problem."

> of the alternative courses of action. The whole decision process then threatens to become dangerously self-referential.

So Skyrms says that if we already know our beliefs, preferences and decision rule, then we would automatically also know what action we will choose. But if an agent already knows which action she will take – Skyrms continues – she cannot make decisions at all, anymore. Hence, an agent cannot know all causes of her actions, which in turn means that the tickle defense presented in this paper fails to apply because the introspection assumption isn't satisfied.

I disagree with this part of Skyrms' analysis. If a reasoner knows his beliefs, preferences and decision rule, he still does not know which decision is implied by them (cf. Eells, 1982, p. 150, as discussed in Section 6.2; also Eells, 1984b, p. 87). He has *logical uncertainty* (see, e.g., Garrabrant et al., 2016, and references given in Section 1.2) over what decision he will make and this should be enough to evaluate different alternative outputs. Consider Newcomb's problem as an example and let's assume that your decision will be the one resulting from EDT applied to some known beliefs and preferences. Then you might still not know what decision you will make. After all, an EDT agent cannot simply evaluate EDT applied to her beliefs and preferences and expect an answer, since that instance of EDT would do then also try to evaluate itself, *ad infinitum* (cf. Dennett, 1984, p. 112). What you can easily infer is that if you one-box, the predictor will probably predict that you one-box and put money under the box. For EDT, this is good enough reason to one-box.

This proof of concept notwithstanding, I believe that decision making under knowledge of some full description of one's decision (e.g., EDT applied to some beliefs and preferences) poses theoretical problems that have received little attention in the literature so far. For instance, some have alleged that such an agent could choose any action a with the justification, "If I choose a, then this implies that a is the action recommended by my decision rule applied to my beliefs and desires. Hence, if I take a, a is the action that I should take." (Drescher, 2006, Section 5.5; cf. Lewis, 1981, footnote 8; Jeffrey, 1977)

6.2 Eells

The version of the tickle defense put forth by Eells (1982, ch. 6-8; cf. 1984a, Section 2, pp. 181f.) is perhaps the best known one. His version resembles ours in that it is based on viewing the beliefs, preferences and decision rule as the exhaustive list of causes of a decision. On p. 138f., he writes (cf. p. 146):

> [I]t is natural to think that a rational person's act – and, indeed, the rational act – is determined by the body of information the person has at hand (by his subjective beliefs and desires)[...].
>
> [...]
>
> I propose that the only way in which a common cause can cause an agent-like person to perform the symptomatic act is by suitably affecting his body of possessed information. I shall assume that the way in which a common cause causes a rational person to perform a symptomatic act is by causing him to have such beliefs and desires leads to the conclusion that the symptomatic act is the best act. And I shall assume that our agent believes this hypothesis about how the common cause causes the symptomatic act.

In response to Skyrms' critique of the beliefs-desires-decision-rule version (see Section 6.1), Eells (1982, p. 150) notes that

> even if [the agent] does know [which decision rule he will use], there may still be some uncertainty about which act he will choose, for there may still be some calculation left to do. He may have chosen but not yet applied the rule. For instance, even if the agent knows that he will [use EDT], he may still not know which act he will perform, for even though he knows his probabilities, desirabilities and rule, he must still *calculate* the [expected utilities according to EDT] of the available acts.

Two further points of resemblance are Eells' use of extended causal structures (pp. 139, 193) and his discussion of higher-order beliefs and their necessity for the tickle defense to work (pp. 144f.).

However, Eells' and my view of the tickle defense come apart when it comes to the scope of the tickle defense. On my view, the tickle defense only implies agreement between CDT and EDT in medical Newcomb-like problems. Eells, on the other hand, uses the tickle defense to argue for agreement between CDT and EDT in all decision problems. He seems to simply miss the possibility of Newcomb-like problems from logical dependences that persist when all common causes are known.

What's interesting about this is that Eells refers to logical dependences and logical uncertainty in multiple places. Most importantly, on p. 111 he defines the "subjective common cause principle":

> Two logically independent things, X and Y, can be subjectively probabilistically relevant to each other only to the extent that the agent believes one or more of the following: X is causally relevant to Y, Y is causally relevant to X, and there is some other factor which is commonly causally relevant to both X and Y.

This is essentially a version of Reichenbach's common cause principle that is weakened by explicitly acknowledging the possibility of logical dependences. Unfortunately, Eells seems to forget about the possibility of logical dependences when he applies the subjective common cause principle to show that CDT and EDT are equivalent (see, e.g., his discussion of Newcomb's problem in chapter 8).

6.3 On the introspection assumption

The tickle defense presented in this paper (and most other versions of it) assume that the agent knows which beliefs, preferences and decision theories it has. Some have argued that this assumes too much, i.e., that EDT should (and does not) make the right recommendations when the agent does not have introspective access to the causes of her decisions (e.g., Lewis, 1981, pp. 10f.). Let us call this the *lack of introspection critique* of the tickle defense. Counter-arguments against this critique of the tickle defense have to assert one of the following:

1. For a decision theory to be plausible it need not make reasonable recommendations in situations in which the agent does not know its beliefs, preferences and decision theory (Jackson and Pargetter, 1983, p. 296f.).

2. EDT makes the correct recommendation, even when it does not possess such introspective knowledge. This in turn, can be asserted in two different ways.

a EDT agrees with causal decision theory on common cause Newcomb-like problems, even if it does not have the introspective knowledge necessary for the tickle defense presented in Section 5 to work (see Section 6.4).

b EDT does not always agree with causal decision theory on common cause Newcomb-like problems, but when they do disagree EDT's choice is preferable (or at least acceptable).

Both 1 and 2b seem acceptable to me.

A notable variation of the lack of introspection critique is due to Jackson and Pargetter (1983). They first offer arguments to the effect of counter-argument 1 above (p. 296f.), but then apply the lack of introspection critique to a different target. Roughly, their argument is the following:

1. A good decision theory should enable us to assess *other* agents' options relative to *their* goals and *our* beliefs.
2. If someone else is facing, say, the Smoking Lesion problem, we should assess smoking to be better for them than not smoking.
3. If we use EDT to assess their options, then it will assess smoking to be worse than not smoking.
4. Hence, EDT is not a good decision theory.

The tickle defense of Section 5 does not apply as a rebuttal of step 3, because we often do not know whether the other agent has the tickle (the desire to smoke) or not.

Nevertheless, I disagree with Jackson and Pargetter's argument. In particular, I find claims 1 and 2 implausible. There are various ways in which a decision theory or system of epistemology can make recommendations to other agents. Here are the ones that one can uncontroversially demand from such theories:

- If you could (and had to) force the other person do take a particular action, which action would you force them to take?
- Which action of the other agent would make you, as their friend, "happiest"? I.e., which one makes you update to the highest expected utility for them?

But EDT gets the first one right and is no obstacle to getting the second one right as well. By making recommendations for other agents Jackson and Pargetter must mean something other than these and I do not see why decision theories should provide such other recommendations (see Dufner and Schmidtz, 1988, for essentially the same line of reasoning). In fact, I do not even see a reason why decision theories would need to model other things in the environment as agents facing decisions.[11]

Of course, there are other ways to argue against Jackson and Pargetter's version of the lack of introspection critique. If you believe that they successfully repair the lack of introspection critique in face of objection 1 above, you could still defend the tickle defense using some version of objection 2 (cf. Eells, 1985).

[11] Eells (1985) argues that CDT sometimes gives unsatisfactory assessments of other agents' actions as well. We could take this as an argument that making recommendations for other agents is not generally a demand that decision theories can fulfill.

6.4 Eells redux

In response to the aforementioned limitations of the tickle defense of this paper – inapplicability in cases of imperfect introspection and what we call logical dependence – Eells (1984b; 1984a, pp. 185-191) has proposed a second version of the tickle defense (also see Skyrms, 1986).

To understand his tickle defense, we first have to slightly broaden our perspective of how an EDT agent operates. Usually, we view actions as things that directly affect the outside world, such as saying "two boxes please" or moving an arm. But Eells notes that there can also be *deliberative acts*. For instance, calculating π up to the 10th digit is a deliberative act and an agent may decide whether to perform that act or do something else, such as performing some other calculation.

Similarly, available evidence to condition one's probability distribution on is usually understood to consist of things received from the external world, such as seeing the finger prints on a murder weapon. But a rational agent should also consider "mental" evidence, such as the output of calculations it has conducted.

I believe that this view of an EDT agent is perfectly legitimate. Indeed, it is necessary to correctly deal with problems in which, e.g., the agent's computations have some (causal or non-causal) effects on the world that is not mediated by the agent's action.

Let's consider an agent facing Newcomb's problem. Now, that agent could decide between computing the expected value of one- and two-boxing and taking an action at random before calculating any expected value. (Of course, the agent may also perform a number of other computations, but calculating the expected values seems most relevant.) Let's say the agent decides, as seems reasonable under most circumstances, to compute the expected values. The expected value of one-boxing is higher than that of two-boxing. However, that doesn't force the agent to one-box immediately. Instead, it could perform further computations. In particular, it could re-compute the expected values, now conditioning on the additional mental piece of evidence e_1, representing the results of the first expected value calculations. After that, it could perform another round of expected value calculations, now conditioning on the mental evidence e_1, e_2 from the first two rounds, and so forth. Let's say that the agent is forced to make a decision at some point, e.g., because the predictor's offer threatens to expire (leaving the agent with no money at all). Then at this point, the agent will make a decision based on the latest pair of expected values $\mathbb{E}\left[\text{payoff} \mid \text{one/two-box}, e_1, ..., e_n\right]$.[12]

So, what does this have to do with the tickle defense? Eells argues that the mental evidence from the previous expected value calculations $e_1, ..., e_n$ enables a form of screen-off (similar to how knowledge of the agent's beliefs and desires screens off the action from its causal predecessors in our version of the tickle defense). For example, in Newcomb's problem, so Eells argues, if you know that your past expected values favor one-boxing, this already gives you such strong evidence that the predictor will put money in both boxes that

$$\begin{aligned} & P(\text{money in both boxes} \mid \text{one-box}, e_1, ..., e_n) \\ \approx\ & P(\text{money in both boxes} \mid e_1, ..., e_n) \\ \approx\ & P(\text{money in both boxes} \mid \text{two-box}, e_1, ..., e_n). \end{aligned}$$

That is, Eells claims that once you have calculated a series of expected values (each conditioning on the result of the previous), your actual action provides no or little further evidence on what the prediction of your decision will be. Thus, according to this new probability distribution, two-boxing is better.

[12] This iterative evaluation is not specific to EDT. We could similarly consider CDT agents that iteratively compute causal expected utility, conditioning on past expected utility (Armendt, 2019).

I will argue that this tickle defense does not work in what I take to be the most relevant cases. First, consider common cause Newcomb problems. Here, the more standard tickle defense described in Section 5 already establishes that the action has only causal bearing on the payoff and hence that EDT and CDT give identical recommendations. That said, remember that our tickle defense requires that the causes of our decision are not only accessible for expected value calculations but described in the agent's beliefs. Were this not the case, then $e_1, ..., e_n$ could provide evidence about these causes and thus establish a new way to screen off the action from its causal predecessors.

Second, consider a case like Newcomb's problem, in which the difference between EDT and CDT results from a logical dependence of the payoff on the agent's decision. Here, our tickle defense is not applicable. However, I see no reason why Eells' (1984b) second tickle defense would establish the convergence of EDT and CDT, either. In Newcomb's problem we know that the predictor predicts whether we eventually one- or two-box. In particular, we have no reason – other than mere speculation – to believe that Omega bases the content of box B (directly) on our initial expected value calculations that we do not (directly) derive our action from. Hence,

$$\begin{aligned} & P(\text{money in both boxes} \mid \text{one-box}, e_1, ..., e_n) \\ \approx \quad & P(\text{money in both boxes} \mid \text{one-box}) \end{aligned}$$

and

$$\begin{aligned} & P(\text{money under both boxes} \mid \text{two-box}, e_1, ..., e_n) \\ \approx \quad & P(\text{money under both boxes} \mid \text{two-box}). \end{aligned}$$

Thus, the more elaborate EDT agent described by Eells evaluates one- and two-boxing in a very similar way as the EDT agent who performs only one expected value calculation (Horwich, 1987, p. 185; Joyce, 1999, p. 158f.).

This is not to say that agents should never perform such "screen offs by inclination" (Ahmed, 2014, p. 105). Imagine you faced a real-world implementation of Newcomb's problem, in which the prediction of your decision is made by a panel of 10 leading psychologists. You may hold the plausible belief that these psychologists are good at predicting your initial inclinations or intuitions (such as thinking that this is fishy or considering that $1000 isn't all that much). However, you may at the same time believe that the psychologists have close to no ability to predict your thinking as it develops from these inclinations and that therefore their decision is based on their predictions about your initial inclinations. If this is true, then once you have observed your own initial inclination, your actual decision can provide you with no further evidence about the psychologist's prediction. Similarly, if you play a prisoner's dilemma against an identical twin who has received no training in decision theory, your initial inclination probably provides you with strong evidence about the twin's inclinations. But after that, the decision you make from a trained decision-theoretical perspective doesn't give you any further evidence about whether your twin cooperates or not (Ahmed, 2014, Section 4.5, 4.6; cf. Horwich, 1987, Section 11.2).

That said, even in cases where the psychologists can only reliably predict one's first, say, 10 expected utility calculations (or perhaps one's first minute of deliberation), an EDT-based case for one-boxing could be made: if you are aware of the psychologist's limited ability to predict, EDT might recommend to resolve to one-box quickly in order to make it more likely that you end up with the million dollars. Of course, this requires that the psychologists can predict whether you make an early decision (as opposed to engaging in a long series of deliberative acts).

Eells presents his argument as being about orthodox EDT. Although his argument fails, we could modify EDT in such a way that Eells' argument does apply (Joyce, 1999, Section 5.2). That is, we could let the agent *pretend* that a tentative decision to one- or two-box screens off the final decision from, say, the box in Newcomb's problem even if it does not. In particular, the structure of Eells argument is the same as the structure of reasoning in so-called ratificationist versions of EDT.[13] Intuitively speaking, ratificationism states that one should only take an action if the action is optimal given that one decides to take that action. In Newcomb's problem, a ratificationist version of EDT is supposed to reject one-boxing based on the following line of reasoning: if I decided to one-box, then this would mean that the $1M are probably in box B. But given that there is $1M under the box, I should two-box after all. Hence, one-boxing is not ratifiable. In this way, ratificationist EDT is supposed to yield CDT-like behavior, potentially without having to rely on a notion of causality. For an overview on ratificationism, see Weirich (2016, Section 3.6).

6.5 Pseudo-medical Newcomb-like problems

As noted in Section 5, I don't believe the tickle defense applies to Newcomb's problem. More generally, I claim that CDT and EDT can disagree if, even under the introspection assumption, if there is some logical (and thus non-causal) connection between the agent's choice and its payoff. The difference can be subtle, however.

As an example, consider Ahmed's (2014, Section 5.1) Betting on the Past case:

> Betting on the Past: In my pocket (says Bob) I have a slip of paper on which is written a proposition P. You must choose between two bets. Bet 1 is a bet on P at 10:1 for a stake of one dollar. Bet 2 is a bet on P at 1:10 for a stake of ten dollars. So your pay-offs are as in Table 1. Before you choose whether to take Bet 1 or Bet 2 I should tell you what P is. It is the proposition that the past state of the world was such as to cause you now to take Bet 2.

(The way P is stated assumes that the world is deterministic to the extent that the past state uniquely determines which bet you accept. But it's easy to come up with non-deterministic versions.) Note that Bet 1 and Bet 2 are bets on the same proposition, but Bet 1 offers has the better odds. Also, your choice does not causally affect whether P is true.[14] For these reasons, CDT always prefers Bet 1 over Bet 2. Of course, if the agent accepts Bet 1, then P is false and so the agent loses the dollar. If, on the other hand, the agent accepts Bet 2, then P is true and the agent wins a dollar. EDT recommends accepting Bet 2 based on this reasoning.

What is interesting about this case is that at first sight it may seem to be a medical or common-cause type problem and therefore covered by the tickle defense: The past state causes the agent's decision and it has to do so via the agent's beliefs, preferences and decision-theoretical attitudes, see Figure 8. However, as in Newcomb's problem, this causal model doesn't express how one would naturally reason about what evidence accepting Bet 2 provides. Elaborating on the above, the way EDT arrives at the conclusion that Bet 2 is better is

[13] This similarity has been noted by a number of authors, including Price (1986, Section 3), Horwich (1987, Section 11.3) and Ledwig (2000, p. 153).

[14] Of course, for this argument to go through we need to assume that the bet is not resolved by looking which bet you in fact choose. Because if that were the case, your choice would causally affect how the bets are resolved.

	P	¬P
Take Bet 1	10	−1
Take Bet 2	1	−10

Table 1: Payoffs in Betting on the Past.

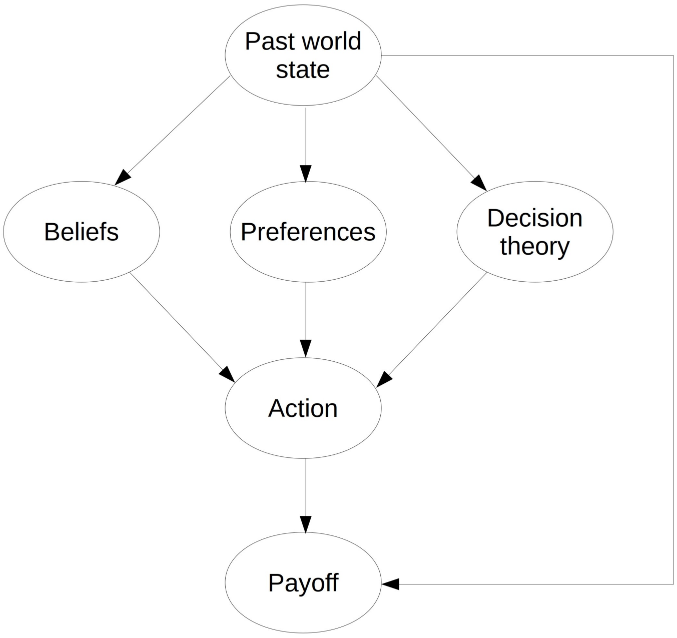

Figure 8: A representation of Betting on the Past as a causal graph from the outside, logically omniscient perspective.

roughly as follows. If I choose Bet 2, then I learn the logical fact that every agent with beliefs, preferences and decision theory such as mine chooses Bet 2. Hence, if I accept Bet 2 I learn that all possible past states that lead me to have these beliefs, preferences and decision theory, lead to my accepting Bet 2 and therefore that P is true. As in Newcomb's problem, we see that the inference from the agent accepting Bet 2 to P being true is via a logical fact, see Figure 9.

In principle, the Smoking Lesion could have a similar structure as Betting on the Past. That is, the gene that causes cancers also causes smoking in complicated, unknown ways such that even looking at your beliefs, preferences and decision theory and being aware of the study, you cannot tell – until you have made your decision – whether you have the gene or not (Eells, 1984a, Section 2; Levi, 1985, p. 239; Price, 1986, Section 3). For example, imagine that hundreds of studies have been conducted and each of them has shown a strong correlation between choosing to smoke and getting cancer. In many of these studies, the participants were aware of all previous studies and had introspective access to their beliefs, preferences and decision theories. Overall, these participants were in analogous situations as you. Even among

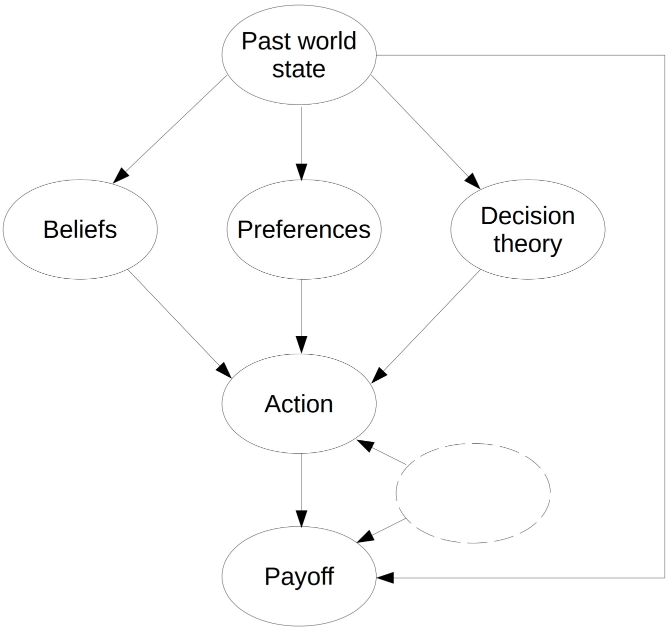

Figure 9: An alternative representation of Betting on the Past as a pseudo-causal graph. The dashed node represents the logical fact as to whether past states satisfy P if they lead to the agent's beliefs, preferences and decision theory.

these participants, the correlation persisted.

What physical circumstances could bring about such a situation? Imagine that there is a gene that is widespread in the population (including people who don't get cancer and don't smoke) and has the following function. This gene is translated into proteins that in turn analyze other genes, in particular the genes that determine whether an individual will decide to smoke, including whether the individual will decide to smoke upon being exposed to studies about how smoking doesn't cause cancer, and so on. The protein then causes cancer if the individual is predisposed to smoke, even after being exposed to these studies.

In this version of the Smoking Lesion, the tickle defense does not apply and EDT chooses to smoke, even if randomized controlled trials show that smoking doesn't cause cancer. I think in this account of the Smoking Lesion, it is rational to refrain from smoking (as I find it rational to prefer Bet 2 over Bet 1 in Betting on the Past). However, this account of the Smoking Lesion is widely implausible. Although medical Newcomb-like cases are usually underspecified, I believe that when people read them they usually interpret it in such a way that the tickle defense applies.

References

Ahmed, Arif (2014). *Evidence, Decision and Causality.* Cambridge University Press.

Ahmed, Arif and Adam Caulton (Dec. 2014). "Causal Decision Theory and EPR correlations". In: *Synthese* 191.18, pp. 4315–4352.

Albert, Max and Ronald Asher Heiner (2001). *An Indirect-Evolution Approach to Newcomb's Problem.* CSLE Discussion Paper, No. 2001-01. URL: https://www.econstor.eu/bitstream/10419/23110/1/2001-01_newc.pdf.

Almond, Paul (Sept. 2010). *On Causation and Correlation Part 1: Evidential decision theory is correct.* URL: https://casparoesterheld.files.wordpress.com/2016/12/almond_edt_1.pdf.

Altair, Alex (2013). *A Comparison of Decision Algorithms on Newcomblike Problems.* Machine Intelligence Research Institute. URL: http://intelligence.org/files/Comparison.pdf.

Armendt, Brad (May 2019). "Causal Decision Theory and Decision Instability". In: *The Journal of Philosophy* 116.5, pp. 263–277. DOI: 10.5840/jphil2019116517.

Arntzenius, Frank (2010). "Reichenbach's Common Cause Principle". In: *The Stanford Encyclopedia of Philosophy.* Ed. by Edward N. Zalta. Fall 2010. Metaphysics Research Lab, Stanford University.

Ben-Gal, Irad (2007). "Bayesian Networks". In: *Encyclopedia of Statistics in Quality and Reliability.* Ed. by Fabrizio Ruggeri, Ron S. Kenett, and Frederick W. Faltin. Wiley & Sons. URL: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470061572.eqr089.

Dennett, Daniel C. (1984). *Elbow Room – The Varieties of Free Will Worth Having.* Oxford University Press.

Drescher, Gary L. (2006). *Good and Real – Demystifying Paradoxes from Physics to Ethics.* MIT Press. URL: https://www.gwern.net/docs/statistics/decision/2006-drescher-goodandreal.pdf.

Dufner, Thomas and David Schmidtz (Nov. 1988). "The 'Tickle Defense' defense". In: *Philosophical Studies* 54.3, pp. 383–386.

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

— (1984a). "Causal decision theory". In: *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, pp. 177–200.

Eells, Ellery (July 1984b). “Metatickles and the dynamics of deliberation”. In: *Theory and Decision* 17.1, pp. 71–95.

— (1985). “Reply to Jackson and Pargetter”. In: *Paradoxes of Rationality and Cooperation: Prisoner’s Dilemma and Newcomb’s Problem.* Ed. by Richmond Campbell and Lanning Sowden. University of British Columbia Press. Chap. 11, pp. 219–223.

Egan, Andy (Jan. 2007). “Some Counterexamples to Causal Decision Theory”. In: *The Philosophical Review* 116.1, pp. 93–114. URL: http://www.jstor.org/stable/20446939.

Garrabrant, Scott et al. (Sept. 2016). *Logical Induction.* Machine Intelligence Research Institute. URL: https://intelligence.org/files/LogicalInduction.pdf.

Horgan, Terence (June 1981). “Counterfactuals and Newcomb’s Problem”. In: *The Journal of Philosophy* 78.6, pp. 331–356.

Horwich, Paul (1987). *Asymmetries in Time – Problems in the Philosophy of Science.* MIT press.

Jackson, Frank and Robert Pargetter (Sept. 1983). “Where the tickle defence goes wrong”. In: *Australasian Journal of Philosophy* 61.3.

Jeffrey, Richard C. (1977). “A Note on the Kinematics of Preference”. In: *Erkenntnis* 11, pp. 135–141.

— (1983). *The Logic of Decision.* 2nd ed. The University of Chicago Press.

Joyce, James M. (1999). *The Foundations of Causal Decision Theory.* Cambridge Studies in Probability, Induction, and Decision Theory. Cambridge University Press.

Ledwig, Marion (Jan. 2000). “Newcomb’s Problem”. PhD thesis. University of Constance. URL: https://kops.uni-konstanz.de/bitstream/handle/123456789/3451/ledwig.pdf.

Levi, Isaac (1985). “Common Causes, Smoking, and Lung Cancer”. In: *Paradoxes of Rationality and Cooperation: Prisoner’s Dilemma and Newcomb’s Problem.* Ed. by Richmond Campbell and Lanning Sowden. University of British Columbia Press. Chap. 13, pp. 234–247.

Lewis, David (1981). “Causal Decision Theory”. In: *Australasian Journal of Philosophy* 59.1, pp. 5–30.

Nozick, Robert (1969). “Newcomb’s Problem and Two Principles of Choice”. In: *Essays in Honor of Carl G. Hempel.* Ed. by Nicholas Rescher et al. Springer, pp. 114–146. URL: http://faculty.arts.ubc.ca/rjohns/nozick_newcomb.pdf.

Oesterheld, Caspar and Vincent Conitzer (2019). “Extracting Money from Causal Decision Theorists”. URL: https://users.cs.duke.edu/~ocaspar/CDTMoneyPump.pdf.

Pearl, Judea (2009). *Causality. Models, Reasoning, and Inference.* 2nd ed. Cambridge University Press.

— (Nov. 2021). *Causation and Decision Theory – An Introduction.* Tech. rep. R-512. Computer Science Department, University of California, Los Angeles. URL: https://ftp.cs.ucla.edu/pub/stat_ser/r512.pdf.

Poellinger, Roland (2013). “Unboxing the Concepts in Newcomb’s Paradox: Causation, Prediction, Decision”. URL: http://philsci-archive.pitt.edu/9887/7/newcomb_in_ckps.pdf.

Price, Huw (1986). “Against Causal Decision Theory”. In: *Synthese* 67, pp. 195–212.

Price, Huw and Richard Corry, eds. (2007). *Causation, Physics, and the Constitution of Reality: Russell’s Republic Revisited.* Oxford University Press.

Russell, Stuart and Peter Norvig (2010). *Artificial Intelligence. A modern approach.* 3rd ed. Pearson Education, Inc.

Schaffer, Jonathan (2016). “The Metaphysics of Causation”. In: *The Stanford Encyclopedia of Philosophy.* Ed. by Edward N. Zalta. Fall 2016. Metaphysics Research Lab, Stanford University.

Skyrms, Brian (1980). *Causal Necessity. A pragmatic investigation of the necessity of laws.* Yale University Press.

— (1986). “Deliberational Equilibria”. In: *Topoi* 5, pp. 59–67.

Spohn, Wolfgang (2003). “Dependency Equilibria and the Causal Structure of Decision and Game Situation”. In: *Homo Oeconomicus* 20, pp. 195–255.

— (Jan. 2012). “Reversing 30 years of discussion: why causal decision theorists should one-box”. In: *Synthese* 187.1, pp. 95–122.

Steele, Katie and H. Orri Stefánsson (2020). “Decision Theory”. In: *The Stanford Encyclopedia of Philosophy.* Ed. by Edward N. Zalta. Winter 2020. Metaphysics Research Lab, Stanford University. URL: `https://plato.stanford.edu/archives/win2020/entries/decision-theory/`.

Talbott, W. J. (Mar. 1987). “Standard and non-standard Newcomb Problems”. In: *Synthese* 70.3, pp. 415–458.

Treutlein, Johannes (July 2018). *How the decision theory of Newcomblike problems differs between humans and machines.* Talk at the 2nd Workshop on Decision Theory & the Future of Artificial Intelligence in Munich, Germany.

Weirich, Paul (2016). “Causal Decision Theory”. In: *The Stanford Encyclopedia of Philosophy.* Spring 2016. URL: `https://plato.stanford.edu/archives/spr2016/entries/decision-causal/`.

Yudkowsky, Eliezer (2010). *Timeless Decision Theory.* The Singularity Institute. URL: `http://intelligence.org/files/TDT.pdf`.

Yudkowsky, Eliezer and Nate Soares (2017). *Functional Decision Theory: A New Theory of Instrumental Rationality.* URL: `https://arxiv.org/abs/1710.05060`.

— (May 2018). *Functional Decision Theory: A New Theory of Instrumental Rationality.* URL: `https://arxiv.org/abs/1710.05060v2`.