What decision theory provides the best procedure for identifying the best action available to a given artificially intelligent system?

S. A. Barnett

MMathPhil Philosophy Thesis

Abstract

Decision theory has had a long-standing history in the behavioural and social sciences as a tool for constructing good approximations of human behaviour. Yet as artificially intelligent systems (AIs) grow in intellectual capacity and eventually outpace humans, decision theory becomes evermore important as a model of AI behaviour. What sort of decision procedure might an AI employ? In this work, I propose that *policy-based causal decision theory* (PCDT), which places a primacy on the decision-relevance of *predictors* and *simulations* of agent behaviour, may be such a procedure. I compare this account to the recently-developed *functional decision theory* (FDT), which is motivated by similar concerns. I also address potentially counterintuitive features of PCDT, such as its refusal to condition on observations made at certain times.

Acknowledgements

This work was originally presented for the Part C Philosophy Thesis in the MMath-Phil Mathematics and Philosophy course at the University of Oxford. As such, I would like to first and foremost thank my supervisor Hilary Greaves for her constructive feedback and support in writing this. I would also like to thank my tutor Steven Methven, who up until last year was responsible for teaching me all of the philosophy I know. Finally, I would like to thank Abram Demski and the rest of the participants of the 2017 MIRI/CFAR AI Summer Fellows Program for their stimulating discussions in the early stages of this project.

Contents

1 Decision-Relevant Dependence **5**

1.1 Formalising Decision Theory . . . 5

1.2 Causal Decision Theory (CDT) . . . 7

1.2.1 Formalising CDT . . . 8

1.2.2 Causal Counterfactuals as Interventions . . . 10

1.3 Newcomblike Scenarios . . . 11

1.3.1 If You're So Smart, Why Ain'cha Rich? . . . 14

1.3.2 Predictors and Simulations . . . 15

1.4 Evidential Decision Theory (EDT) . . . 17

1.4.1 Managing the News . . . 19

1.4.2 The Smoking Lesion Problem . . . 22

1.5 Functional Decision Theory (FDT) . . . 23

1.5.1 Subjunctive Dependence . . . 24

1.5.2 FDT in Action . . . 26

1.5.3 The Problem of Counterlogicals . . . 27

2 Policy-Based Causal Decision Theory **31**

2.1 Policy Selection . . . 31

2.1.1 Three Problems for Policy Selection . . . 32

2.1.2 Decision-Relevant Dependence for Policy Selection . . . 40

2.2 Formalising PCDT . . . 40

2.3 Comparison with Related Accounts . . . 43

2.3.1 Causal Decision Theory 43
2.3.2 Spohn's CDT 45
2.3.3 Meacham's Cohesive Expected Utility Theory 47
2.3.4 Graphical Updateless Decision Theory 48
2.4 What a Policy Node Represents 52

3 The Veil of Ignorance 56

3.1 The Curious Benefactor 56
3.1.1 Generalising the Problem in CB 58
3.2 Making Reliable Assurances 60

4 Conclusion and Future Work 62

A A Review of Causal Bayesian Networks 64

A.1 Bayesian Networks 64
A.1.1 Adjusting for Evidence 65
A.1.2 d-separation 66
A.2 Learning Bayesian Networks 67
A.3 Causality 67

Introduction

The purpose of decision theory is to determine what principles of instrumental rationality govern choice. These principles are commonly encapsulated into a formula for the *expected utility* of a given act: an agent adopting the principles will seek an action that maximises her expected utility.[1]

Decision theory has had a long-standing history in the behavioural and social sciences as a tool for constructing good approximations of human behaviour. Yet as artificially intelligent systems (AIs) grow in intellectual capacity and eventually outpace humans, decision theory becomes evermore important as a model of AI behaviour. What sort of decision procedure might an AI employ? What epistemic perspective does an AI take when evaluating the actions it may take?

In this work, I propose a decision theory encapsulating our principles of instrumental rationality that an AI may employ. This decision theory places a primacy on the decision-relevance of *predictors* and *simulations* of agent behaviour. While the same considerations are in principle decision-relevant for humans, in practice they are predominately pertinent to an AI, which is far more likely to find itself in a world in which there exist copies employing the same decision procedure, and in which the source code which governs that decision procedure is openly available for others to see and potentially exploit. Since predictors and simulations base their decisions on how an agent may act in a given scenario, I call such a decision theory *policy-based*.

This work will not demand persistent reference to AIs and their architecture -

[1] I shall hereafter refer to the agent using female pronouns. Any other decision-making agent that is present within a decision problem will take male pronouns.

AIs and humans alike benefit from adopting the principles of instrumental rationality, which are agnostic about the recipient of utility. However, framing certain decision problems *as if they are faced by an AI* helps to tease out certain intuitions with regards to rational behaviour. The question of what decision theory it would be ideal to program *into* an AI, distinct from the question addressed here,[2] is of independent interest.

Chapter 1 provides expository accounts of causal, evidential, and functional decision theory (CDT, EDT, and FDT, respectively), each proposals for the ideal normative theory of instrumental rationality. I take the crux of the disagreement between these theories as determining what facts of the universe ought to be held fixed in evaluating different actions, which I call finding the *decision-relevant notion of dependence*. While I take CDT to possess the right such notion, I disagree with its prescription in *Newcomblike problems*, in which predictors and simulations are not taken to be decision-relevant.

Chapter 2 motivates the policy-based account and develops it formally as policy-based causal decision theory (PCDT). In addition to giving expository comparisons to similar decision theories, I also discuss the interpretation of the novel components within the formalism of PCDT; in particular, what the nature and temporal location of a *policy* is.

Chapter 3 discusses a counter-intuitive feature of PCDT - its refusal to condition on the observations made in its lifetime. I argue that this feature is essential in order to accommodate the decision-relevance of predictors and simulations, and that it becomes a virtue when the agent needs to issue reliable assurances.

[2] In answering this question, claiming that a particular decision problem would not in practice be faced by an AI would be legitimate grounds for the dismissing that problem. However, a decision theory intended to capture ideal instrumental rationality could not avail itself of such a defence.

Chapter 1

Decision-Relevant Dependence

This chapter sets out in detail the problem to be addressed in the thesis: that of finding the ideal decision theory. Once this problem has been specified, I review three decision theories that have been proposed as candidates for the ideal decision theory: causal, evidential, and functional. I argue that all of them fail to capture the correct notion of a decision-relevant link between choice and outcome, which I claim can be seen by considering so-called *Newcomblike scenarios* involving predictors and simulations. In explaining *where* I believe each of these three decision theories fails in this regard, I provide a general motivation towards my proposed decision theory, which I shall develop in subsequent chapters.

1.1 Formalising Decision Theory

In this section, I shall highlight some of the core assumptions that are made by the most prevalent decision theories. In particular, I shall adopt Joyce's account[3] of decision problems, which I briefly review here. The purpose of doing so is twofold: first, in order to highlight some of the key common features; and second, in order to avoid misunderstandings regarding the basic task of decision theory.

The ideal decision theory is the formal decision procedure that best captures our ideal normative theory of instrumental rationality. One suitable definition

[3] Joyce (1999, Chapter 2).

of an ideal normative theory of instrumental rationality is what means one must take to achieve certain outcomes given credences, beliefs, or information about the probabilities that various means will achieve those ends.[4]

More formally, we can use Joyce's definition of a *decision problem* as a common framework for the decision theories to be discussed. According to this framework, an agent chooses among *acts*, whose *outcomes* depend on the *state of the world*. The *decision problem* is described by specifying a partition $\mathcal{A}$ of possible acts among which the agent must decide, a partition $\mathcal{O}$ of outcomes that provides the list of desirable and undesirable things that could occur as a result of the agent's choice, and a partition $\mathcal{S}$ of states that each describe possible external conditions that determine what outcome each act in $\mathcal{A}$ will produce.[5] Notably, states describe "descriptions of aspects of the world that lie outside the decision maker's control",[6] where the scope of the agent's *control* is a matter on which different decision theories disagree.

We seek only one such act, state, and outcome to obtain for a decision problem in actuality. Consequently, we take the propositions of $\mathcal{O}, \mathcal{S}, \mathcal{A}$ to be mutually disjoint within the sets themselves, and we take these sets of propositions to be embedded within a larger set of propositions Ω that has the structure of a Boolean σ-algebra. In particular, Ω is the smallest collection of propositions that includes the partitions $\mathcal{O}, \mathcal{S}, \mathcal{A}$, and is closed under negation and countable disjunction. A decision problem is thereby formally given as $\mathcal{D} = (\Omega, \mathcal{O}, \mathcal{S}, \mathcal{A})$. Furthermore, to account for the views of Savage[7] and Jeffrey[8] that each act/state conjunction $A\&S$ will entail exactly one outcome, we stipulate $\mathcal{O}$ to be a coarsening of the partition set $\mathcal{W} = \{A\&S \mid A \in \mathcal{A}, S \in \mathcal{S}\}$.

We represent the beliefs and desires of the agent by the functions $Cr \colon \Omega \to [0, 1]$ and $\mathcal{U} \colon \Omega \to \mathbb{R}$, referred to as the agent's *credence* and *utility* functions.

[4] Kolodny and Brunero (2016).

[5] Joyce (1999, p. 48).

[6] Joyce (1999, p. 61).

[7] Savage (1972).

[8] Jeffrey (1990).

Equipped with a credence and utility function, an agent, given a set of options constituting a decision problem, ought to use her decision procedure to recommend the option that maximises utility. Her own uncertainty about her situation means that, for each option, the agent is in fact evaluating her *expected* utility: the utility of the outcome weighted by her credence that the outcome will occur on the option being taken.

The claim that an agent should choose an option with maximal expected utility is known as the *expected utility principle*.[9] The decision theories that I shall consider in this thesis all agree with this principle: however, what they disagree on is how one ought to *define* expected utility.[10] This chapter will discuss approaches to the definition that differ with respect to the nature of the dependence between option and outcome that a decision-maker ought to take into account. Call such a dependence a *decision-relevant dependence*. Equivalently, the definitions of expected utility disagree with respect to what varies with the option that the agent takes, or what can be taken to be held *fixed* while the agent evaluates the outcome on taking a certain option.

Finally, it is assumed that our decision-theoretic agent is *idealised*, in the sense that she has unlimited computational resources to evaluate her credence and utility functions, and perform calculations of expected utility.[11]

1.2 Causal Decision Theory (CDT)

In this section, I shall set out **causal decision theory** (CDT). In the terms of the previous section: causal decision theory takes the decision-relevant dependence between option and outcome to be *causal*.[12] Therefore, a CDT agent evaluates

[9] Nozick (1969, p. 118). See, e.g., Joyce (1999, Ch. 1) and Briggs (2017) for discussions of this principle.

[10] Joyce (1999) proves a general representation theorem that, in his words, ties the "'global" requirement to maximize expected utility [with] the "local" [...] constraints on individual beliefs and desires' (Joyce, 1999, p. 224).

[11] Weirich (2004) presents precise methods of evaluating decisions when idealisations concerning an agent's resources and decision problems are relaxed. Morton (2004) discusses the effects of computational limits on decision procedures with specific reference to computational complexity.

[12] Weirich (2016).

the expected utility of an option by weighting the utility of each outcome by her credence that taking that option will effect the outcome.

1.2.1 Formalising CDT

Within the literature, there are a number of ways[13] to specify an agent's credence that choosing an option will causally effect an outcome, which in turn leads to a number of different formalisations of CDT. I favour the formalisation developed by Pearl[14] (amongst others[15]), which I shall informally review here and in further detail in the Appendix.

Pearl's account gives a clear depiction of the causal structure of the universe in which the agent finds herself, and provides the clearest procedure for turning an ordinary-language decision situation into a decision problem $\mathcal{D}$. However, as Joyce[16] and Pearl[17] argue, once one has described the essential features of the causal situation correctly, then the incremental evidence that a cause provides for its effect in virtue of being its cause will always work out to have the same probabilistic value on any of the prominent formalisations of CDT. In particular, one can reduce a decision problem in Pearl's framework to a decision problem $\mathcal{D} = (\Omega, \mathcal{O}, \mathcal{S}, \mathcal{A})$ as specified by Joyce, and vice versa. Therefore, my preference for Pearl's account is mainly æsthetic.

In Pearl's account, the agent possesses a *directed acyclic graph* (DAG), together with an associated conditional probability distribution over the nodes of these graphs, in order to evaluate her causal probabilities. The nodes of these graphs are variables representing, formally, sets of propositions in Ω. For instance, in a decision problem where the agent decides between acts a and b, the agent's

[13] Joyce (1999, Chs. 5-6) reviews approaches to calculating efficacy values through *imaging*, a means of transforming a credence function that shifts subjective probabilistic weight into possible worlds in which the act in question has occurred. Alternatively, one may calculate these values according to the *probabilistic accounts of causation* defended by Skyrms (1980), Suppes (1970), and Eells (1991).

[14] Pearl (1996, 2009).

[15] Spirtes et al. (2000).

[16] Joyce (2010).

[17] Pearl (2010).

performed action will be represented by a variable $\textsc{Act}$, whose value lies in $\{a, b\}$.

The connections between these nodes represent a causal dependence between these outcomes. In particular, each value that a node N may take is determined according to a conditional probability distribution by the values taken by the directly preceding nodes. We call the set of such nodes that contain edges whose target is the node N the set of *parents* of N, denoted $pa(N)$. Thinking of a DAG as a family tree, the *(causal) ancestors* and *(causal) descendants* are defined as expected. A node that is not a target for any edge is called *exogenous*: often such nodes are used as 'error terms' that serve as a catchall for all relevant causal factors left out of the model.

The graphs we use in a decision-theoretic context are characterised by two conditions:[18]

Sufficiency If X and Y are correlated given a variable Z that is not among the descendants of either X or Y, and neither X nor Y is a descendant of the other, then there exists some variable C that is a common cause of both X and Y.

Causal Markov Conditional on values for its causal parents, any variable X is independent of all others, save its descendants. Therefore, if $V = pa(X)$, there exist values of X, Y, and V such that $Cr(y \mid x \& v) \neq Cr(y \mid v)$ only if Y is a descendant of X.[19]

Informally, these two conditions require that every correlation is explained by the existence of a common cause, and that a complete specification of the common causes of two variables will screen off any correlations between them, except those generated by common descendants.

[18] Stated as in Joyce (2010, p. 144).

[19] We denote variables and sets of variables in uppercase, and the values of variables and sets of variables in lowercase.

1.2.2 Causal Counterfactuals as Interventions

CDT requires us to evaluate the agent's credence that taking some act a would effect an outcome o. To do so with the graphical apparatus, we imagine that our graph contains a node representing the action we take, as well as a node representing the outcome that occurs, whose utility we are ultimately trying to evaluate and optimise for. Informally, we evaluate the aforementioned credence as follows:

1. We 'intervene'[20] on the action node by fixing its value to the action a.
2. We 'propagate the value forwards' by sequentially evaluating all of the node's descendants.
3. If propagating action a forwards determines that the outcome node is evaluated as o, then the credence that action a causally effects outcome o is 1. If the outcome node has ancestors that are not themselves descendants of the action node, we consider each possible evaluation of these ancestors, and evaluate the outcome in each case. This gives us a distribution over a causally effecting each outcome by making use of our prior credence over the evaluations of the exogenous variables that are ancestors of the outcome node.

This procedure of intervention and propagation is captured by the do operator, which takes evaluations of sets of variables as its input and outputs a modified graph and credence function. Specifically, one can derive[21] the post-intervention credence to be

$$Cr(x, y, z \mid \texttt{do}(X = x^*)) = \begin{cases} Cr(x \& y \& z)/Cr(x \mid z) & \text{if } x = x^* \\ 0 & \text{if } x \neq x^*. \end{cases} \tag{1.2.1}$$

[20] As Joyce (2010, pp. 146-147) points out, the talk of *interventions* is not meant to imply a necessary connection to human purposes and activities, or import an interventionist metaphysics of causation. Pearl-style interventions are best thought of as *suppositions* that stem from a particular act.

[21] Pearl (2009, pp. 72-73).

Here, z stands for any realisation of the set Z of non-descendants of X, and y is any realisation of the set Y of descendants of X.

Letting our OUTCOME and ACT variables take values in $\mathcal{O}$ and $\mathcal{A}$ respectively, the expected utility of an action a under this formalisation of CDT is given by

$$\mathcal{EU}_{\text{CDT}}(a) := \sum_{o \in \mathcal{O}} Cr(o \mid \texttt{do}(\text{ACT} = a)) \cdot \mathcal{U}(o). \tag{1.2.2}$$

A precise account of the do-calculus is given in the Appendix. Importantly, Pearl[22] provides a clear method to infer a causal model of a given state of affairs that accurately reflects its causal structure, given the Sufficiency and Causal Markov conditions. In many of the decision problems considered, drawing the causal graph allows us to clarify the epistemic perspective of the agent. Moreover, using this method, we are capable of evaluating the relevant causal credences in a way that is both systematic, and provably well-defined.

1.3 Newcomblike Scenarios

In this section, I shall consider decision problems[23] in which I regard CDT to fail to recommend the rational course of action. The next chapter will address how CDT can be modified to address these problems, while still maintaining the parts of CDT that do capture the ideal normative theory of instrumental rationality. In the remainder of this chapter, I shall identify the failure, argue for why I do think it is indeed a failure, and consider some of the unsuccessful proposals for decision-relevant dependence that aim to resolve this failure.

The problem, I claim, is that CDT fails to take into account certain decision-relevant dependencies that involve predictions of your action, or simulations of you whose actions will reflect your own. Decision problems involving these dependencies are called *Newcomblike*, after the following problem:[24]

[22] Pearl (2009).

[23] In all of these problems, utility is taken to be linear in money.

[24] Nozick (1969). This version is due to Sobel, and is quoted by Joyce (1999, p. 146).

Newcomb's Problem (NP) *Suppose there is a brilliant (and very rich) psychologist who knows you so well that he can predict your choices with a high degree of accuracy. One Monday as you are on the way to the bank he stops you, holds out a thousand dollar bill, and says: "You may take this if you like, but I must warn you that there is a catch. This past Friday I made a prediction about what your decision would be. I deposited $1,000,000 into your bank account on that day if I thought you would refuse my offer, but I deposited nothing if I thought you would accept. The money is already either in the bank or not, and nothing you now do can change the fact. Do you want the extra $1,000?" You have seen the psychologist carry out this experiment on two hundred people, one hundred of whom took the cash and one hundred of whom did not, and he correctly forecast all but one choice.*[25] *There is no magic in this. He does not, for instance, have a crystal ball that allows him to "foresee" what you choose. All his predictions were made solely on the basis of knowledge of facts about the history of the world up to Friday. He may know that you have a gene that predetermines your choice, or he may base his conclusions on a detailed study of your childhood, your responses to Rorschach tests, or whatever. The main point is that you now have no causal influence over what he did on Friday; his prediction is a fixed part of the fabric of the past. Do you want the money?*

The causal decision theorist reasons as follows: the psychologist's prediction of my choice of action is causally independent from the action that I in fact take. In this particular scenario, it is fixed in the past, over which I have no causal influence. I must therefore evaluate the outcome under two different possible state of affairs: either the psychologist deposited $1,000,000 in my bank account, or he didn't. Crucially, regardless of which of these two states of affairs holds, I gain $1,000 by accepting the psychologist's offer. Hence, I, a CDT agent, should accept the offer and take the $1,000.

[25] All but one choice, that is, for each of the *hundred*. In other words, the empirical probability of the psychologist making an incorrect prediction is 0.99. This is not made clear from the original text, but it simplifies any calculations without loss of generality.

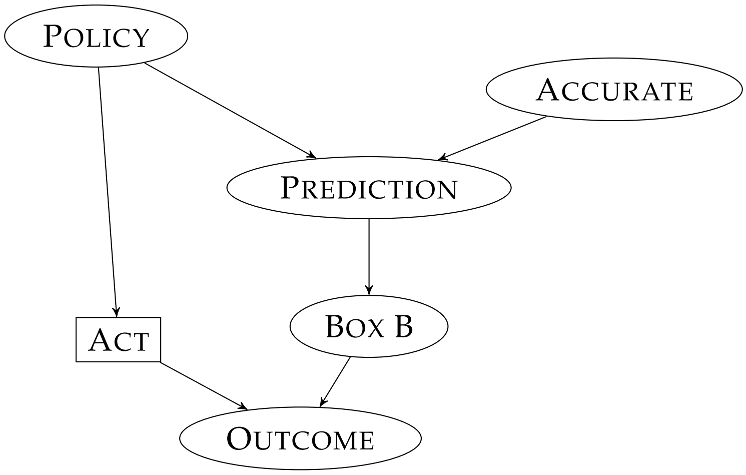

Figure 1.1: A causal graph for Newcomb's Problem. The rectangular border for the ACT node represents that it is the node upon which we intervene. Since interventions on ACT do not propagate to BOX B or any of its ancestors, the content of the box is beyond our control, and so we hold its value fixed, leading to a two-boxing recommendation.

In the original description of NP, there are two boxes. Box A, which is transparent, contains $1,000. Box B, which is opaque, contains $1,000,000 if the predictor has predicted you will take only Box A, and is empty if the predictor has predicted you will take both boxes. The two descriptions of NP are structurally equivalent, though I take the former description to be more sympathetic to the causal decision theorist's argument that you should effectively ignore the psychologist. This is since it specifically places the prediction in the *temporal past*, which implies but is not implied by the more general stipulation that the prediction is causally independent from our action. Yet, despite the structural equivalence, our intuitions about the agent's powerlessness over the psychologist are stronger when we place the prediction in the past.[26] By convention, I use the terms *one-boxing* and *two-boxing*, taken from the original description, to refer to the equivalent of these actions in all Newcomblike scenarios.

[26] Lewis (1979, pp. 236-237) makes a similar point. Gibbard and Harper (1978, pp. 181-182) propose a case, *Transparent Newcomb's Problem* (TNP), in which the contents of Box B are visible to the agent, making the intuition for two-boxing *stronger still*. The decision theory I propose advocates one-boxing in TNP, for reasons I shall later justify when considering a similar case, the Curious Benefactor.

1.3.1 If You're So Smart, Why Ain'cha Rich?

The following argument has been given by a number of decision theorists[27] in favour of one-boxing, *pace* the recommendation of the CDT agent: in NP, the agents that one-box will reliably end up richer than the agents that two-box. All decision theorists acknowledge this. Even a CDT agent who does not currently find herself in a Newcomblike situation acknowledges this. Given that it is presupposed by a theory of instrumental rationality that utility is what we value in decision problems, and utility is taken to be linear in money, it seems hard maintaining that agents that two-box are acting rationally.

This argument is referred to as the 'Why Ain'cha Rich?' (WAR) argument. Although I endorse one-boxing in NP, I believe WAR fails as a reason for making such an endorsement. I give a popular reason for rejecting WAR below, which I endorse. Subsequently, I motivate my intuition that one-boxing is the rational choice in NP.

Joyce responds that WAR fails to address the charge that the one-boxer is being irrational. As Joyce observes,[28] it is part of the definition of NP that the agent must believe that what she does will *not* affect what the psychologist has predicted. Hence, it is irrational for the agent to ignore her belief and one-box, even *if* one-boxers end up richer than two-boxers. Of course, the CDT agent holds that it would be rational to become the type of agent that one-boxes in NP if one knows one is about to face such a situation. However, Joyce claims[29] that if one already finds oneself in a Newcomblike situation, then regardless what type one is, the prediction is fixed, and so two-boxing is in fact the better option. Joyce adds that jointly endorsing two-boxing and being the *type* of person of one-boxes are not inconsistent.

In other words, Joyce's response is: Yes, it is true that a normative theory of instrumental rationality seeks to maximise one's utility. However, the WAR ar-

[27] See, e.g., Hargreaves Heap et al. (1992, p. 342). The argument is also discussed by Lewis (1981a), Arntzenius (2008), and Ahmed (2014, pp. 181-194).

[28] Joyce (1999, p. 152).

[29] Joyce (1999, pp. 153-154).

gument misidentifies the epistemic perspective from which this utility ought to be evaluated in a decision theory. By taking an evaluation of utility which poses the two choices as yielding either $1,000 (for being a two-boxer) or $1,000,000 (for being a one-boxer), it ignores the fact that a decision theory is supposed to evaluate expected utility from the perspective of the agent herself. The agent's beliefs, in particular her belief that she cannot influence the psychologist's already-made prediction, means that the choices available yield from the agent's perspective either $\$(N + 1000)$ or $\$N$, where the value of N is *fixed*. Under this perspective, the two-boxing agent *is* choosing the option that makes her richer, *pace* WAR.

1.3.2 Predictors and Simulations

I agree with Joyce that the WAR argument makes a mistake about the epistemic perspective from which an agent's actions are evaluated. However, I believe further that two-boxers *are* behaving irrationally in NP. To wit: there intuitively exist decision-relevant dependencies stemming from the agent's actions in NP that CDT fails to capture.

The agent rightly believes that the psychologist's prediction is a fixed part of the past. In graphical terms, this means that we cannot draw a direct causal link between the agent's action and the psychologist's prediction or the contents of the agent's bank account (see Fig. 1.1). Nonetheless, the agent also believes, importantly, that the psychologist is a highly reliable predictor of her choices. Consequently, the agent is aware of the fact that the psychologist's prediction reliably tracks the actions of any agent in a Newcomblike scenario. Even though the past is fixed, the agent knows that she can reliably infer the content of this fixed past through the choice that she makes. This is an intuitive dependency between our choice and the outcome, acknowledgeable to any agent who finds herself within a Newcomblike problem, that I take to arise from describing Newcomblike problems as involving predictor-like entities.

As a justification for one-boxing, this is distinct to the appeal to 'a view-fromnowhere' or 'long-run utility'[30] à la WAR. My justification for one-boxing respects that the agent must reason from her epistemic perspective alone, and in such an epistemic perspective she already finds herself within a Newcomblike scenario. My intuition is that, if the agent is rational, she believes that whatever she chooses, this will be reliably and relevantly tracked by the prediction of the psychologist, even if this tracking does not constitute a direct causal link between action and prediction. Therefore, the non-causal dependency between the agent's action and the psychologist's prediction is decision-relevant, and so we cannot hold the prediction fixed while varying the agent's action. CDT fails to account for all of the decision-relevant dependencies that I claim are epistemically available to an agent in Newcomblike scenarios.

There is a another Newcomblike scenario worth mentioning. It does not involve predictors,[31] though I shall show that it is equivalent to a scenario that does. Moreover, such a scenario is important in understanding how near-perfect prediction may be physically possible, particularly if the agent is an AI.

Psychological Twin Prisoner's Dilemma (TPD) *An agent and her psychological twin must both choose to either 'cooperate' or 'defect'. If both cooperate, they each receive $1,000,000. If both defect, they each receive $1,000. If one cooperates and the other defects, the defector gets $1,001,000 and the cooperator gets nothing. The agent and the psychological twin know that they reason the same way, using the same considerations to come to their conclusions. However, the agent's twin has already made her decision in a separate room, without communication. Should the agent cooperate with her twin?*

Lewis[32] argues that this sort of problem is a Newcomblike scenario. To show this, he claims that in NP, all that is required of a prediction is that some 'potentially predictive process' should go on, which yields the outcome that constitutes

[30] See Yudkowsky and Soares (2017, p. 4) for an example of such talk.

[31] The dilemma is due to Nozick (1969, pp. 130-131). This description of the dilemma is adapted from Yudkowsky and Soares (2017, p. 2).

[32] Lewis (1979).

the psychologist's prediction. A clear example of a potentially predictive process is a *simulation*: to make a prediction, the psychologist simply has to make a replica of the agent, put the replica in the agent's predicament, and see whether the replica takes her $1,000. A psychological twin is itself a clear example of a replica. At this stage, we have arrived at TPD from NP.

It is also obvious that we can go in the other direction: just as we can regard the choice of a twin in an identical predicament as the reliable prediction of a psychologist, we can treat the reliable predictions of a psychologist as the choices that constitute the twin. Consequently, any decision-theoretic account should treat dependencies involving predictors as decision-relevant if and only if it also treats dependencies involving simulations as decision-relevant.

Advocates of CDT believe that two-boxing in NP and defecting in TPD are the rational choices. I, on the other hand, believe that one-boxing in NP and cooperating in TPD are the rational choices. In doing so, I am holding that, even though the actions of predictors and simulations may be causally independent from our own actions, the dependencies between our actions and those of our predictors and simulations are *decision-relevant*. My goal in this thesis is to develop a decision theory that is able to coherently incorporate such dependencies, while at the same time avoid incorporating any spurious and irrational dependencies. First, I shall consider two popular decision theories that fail to strike this balance.

1.4 Evidential Decision Theory (EDT)

Jeffrey's[33] **evidential decision theory** (EDT) treats the decision-relevant dependence between choices and outcomes as being *evidential*: roughly, we ought to make the choice that gives us the highest evidence of the outcome with the highest utility (also known as the *auspicious* choice). In order to do so, rather than weighting the utility of an outcome by the agent's credence that her choice will causally effect the outcome, the agent instead weights the utility by her credence

[33] Jeffrey (1990).

that the outcome will obtain, conditioned on the fact that the choice has been made. Formally, with $\mathcal{O}$ any partition of outcomes,[34] an EDT agent evaluates the expected utility of an action A as

$$\mathcal{EU}_{\text{EDT}}(a) := \sum_{o \in \mathcal{O}} Cr(o \mid a) \cdot \mathcal{U}(o). \tag{1.4.1}$$

Here, the conditional credence is defined as

$$Cr(o \mid a) := \frac{Cr(o\&a)}{Cr(a)}. \tag{1.4.2}$$

For any outcome, this conditional credence is only well-defined if $Cr(a) > 0$. Hence, for an agent to consider a, she must have some non-zero credence that she will take that action. If this not the case, then by convention EDT dismisses a as an available option. As such, it is implicit in the cases in which a is a relevant option that the inequality holds, so the conditional credences are well-defined.

EDT is usually taken to prescribe one-boxing in NP and cooperation in TPD.[35] In the former case, recall that the agent believes that the psychologist is 99% reliable in making his predictions. Write our partition of outcomes in NP as the possible amounts of money the agent can end up with. We arrive at the following conditional probabilities:

$$Cr(\$1,000,000 \mid onebox) = 0.99, \quad Cr(\$0 \mid onebox) = 0.01,$$
$$Cr(\$1,000 \mid twobox) = 0.99, \quad Cr(\$1,001,000 \mid twobox) = 0.01.$$

From these values, the reader may verify that $\mathcal{EU}_{\text{EDT}}(onebox) > \mathcal{EU}_{\text{EDT}}(twobox)$, from which it follows that EDT prescribes one-boxing.[36]

[34] EDT is *partition-invariant*. This is true of some forms of CDT (Joyce, 1999), but not others (Lewis, 1981b).

[35] Eells (1984), amongst others, argues otherwise. Such arguments lie beyond the scope of this thesis.

[36] Notably, for EDT to prescribe one-boxing, it is in fact sufficient (and necessary) for the psychologist to be at least 50.05% reliable! In other words, the psychologist only has to perform slightly better than chance for EDT to prescribe one-boxing, given the monetary values at stake in this version of NP.

1.4.1 Managing the News

I do not believe that auspiciousness is the correct property to capture the decision-relevant dependence between action and prediction/simulation. As Lewis argues,[37] EDT 'commends an irrational policy of managing the news so as to get good news about matters which you have no control over'. To see this, we require a decision problem that demonstrates *to the one-boxing advocate* that 'managing the news' is irrational. Given that Lewis uses NP as his demonstration of the irrationality of EDT, we require a separate, non-Newcomblike decision problem.

The following decision problem[38] meets this requirement.

XOR Blackmail (XOR-B) *An agent has been alerted to a rumour that her house has a terrible termite infestation that would cost her $1,000,000 in damages. She doesn't know whether this rumour is true. A greedy psychologist with a strong reputation for honesty learns whether or not the rumour is true, and drafts a letter:*

> "I know whether or not you have termites, and I have sent you this letter if and only if exactly one of the following is true: (i) the rumour is false, and you are going to pay me $1,000 upon receiving this letter; or (ii) the rumour is true, and you will not pay me upon receiving this letter."
>
> *The psychologist then predicts what the agent would do upon receiving the letter, and sends the agent the letter if and only if exactly one of (i) or (ii) is true. We assume that the psychologist is, in fact, infallible, in that his probability of making a mistake is 0. Thus, the claim made by the letter is true. Assume the agent receives the letter. Should she pay up?*

In this case, the agent knows that the psychologist is honest and infallible. Let us suppose our agent operates on EDT. She then knows that, conditional on paying, (i) must be true, so that the rumour is false. However, conditional on

[37] Lewis (1981b, p. 5).

[38] Originally given as the 'Evidential Blackmail Problem' in Soares and Fallenstein (2015, p. 2). This version is from Yudkowsky and Soares (2017, p. 24).

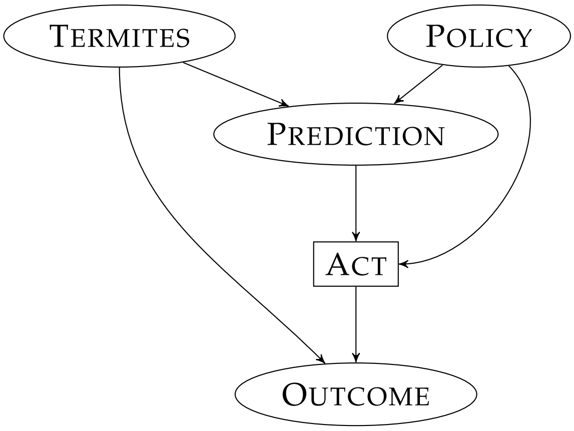

Figure 1.2: A causal graph for XOR-Blackmail. Observe that TERMITES is a non-descendant of PREDICTION, meaning that the psychologist has *no* control on whether termites are present. This is unlike in NP, where the psychologist has the choice of whether to put the money in the bank account or box.

not paying, (ii) must be true, so that the rumour is true. In the latter case, the agent stands to lose more money. Therefore, she decides to pay up. This is mistaken: intuitively speaking, our choice to pay will have no effect on the presence of termites, and so we should not be beholden to such blackmails.

XOR-B is structurally distinct from NP. The key difference is that, in XOR-B, the psychologist learns of the presence or absence of termites, and then makes his prediction. On the other hand, in NP, it is the psychologist's prediction that determines whether he puts the money in the agent's bank account, which is the 'big win' that is analogous to the 'big loss' of having termites.

I share the intuition that, while we ought to treat the psychologist's prediction as varying with the agent's choice, this co-variation does not extend to other events that are beyond the psychologist's control. The relevant distinction here is that, while the psychologist makes his prediction based on a reflection of the agent's decision procedure (perhaps running a simulation of the agent in order to do so), the presence of termites holds no such connection to the agent. This is in spite of the fact that the psychologist makes his prediction *before* the agent makes her choice, and we take the agent's choice to be a genuine exercise of her free will.

The CDT advocate will respond that we are mistaken to hold this distinction as being decision-relevant. Indeed, Joyce's wording of NP above makes a clear attempt to describe the psychologist in a way that downplays his role as a predictor, stressing that the prediction is something the agent now ought to hold *fixed* in her deliberations. Yet despite such pleas, there remains in the philosophical community a vocal minority of inveterate one-boxers[39] whose intuitions side with the decision-relevance of predictors. It is my task in this thesis to develop a decision theory catered to such intuitions that retains the strengths of CDT.

By treating *correlation* as the decision-relevant notion of dependence between choice and outcome, we treat auspiciousness as inherently valuable. In XOR-B, we see how this results in the agent irrationally giving in to the blackmail. Therefore, EDT does commend an irrational policy of managing the news, though we do not see this in Newcomblike scenarios.

A defender of EDT may respond by arguing that, even if the psychologist does not determine through his action whether there are termites, the infallibility of the psychologist together with the fact that the letter has been sent in effect determines whether there are termites or not. To develop the defence further: the agent knows that the letter has been sent, and takes the content of the letter to be true. Therefore, the agent ought to deduce that, if she chooses not to pay, then she will have termites. Why, then, should the agent treat the presence of termites as beyond her influence, but not the psychologist's prediction, which is also in the past and so causally independent from the agent's action?

My response, which shall be the basis of my proposed decision theory in the following chapter, is that while the decision-relevant notion of influence we should employ is a *causal* one, the decision-relevant causal influence stems not from one's action, but from one's having been the kind of agent who takes that action in the decision problem at hand. Under this categorisation of decision-relevant influence, we do have influence over the psychologist's prediction. Yet we do not have influence over just anything that is correlated with our action.

[39] E.g., Ahmed (2014), Meacham (2010), and Spohn (2012).

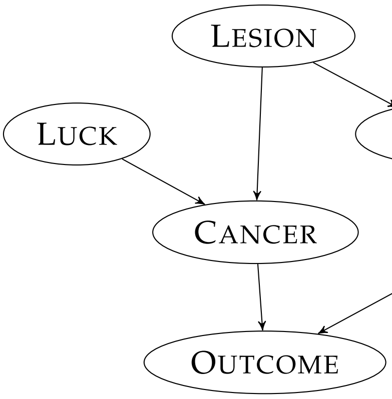

Figure 1.3: A causal graph for Smoking Lesion. Note the assumption that the presence or absence of LESION is screened off from ACT by POLICY. Here, LUCK signifies the possibility that one has the lesion but has avoided getting cancer, playing a similar role to that of ACCURATE in NP.

1.4.2 The Smoking Lesion Problem

'Medical Newcomb Problems' are another category of decision problems that are cited in order to underscore the irrationality of EDT. Consider the following:[40]

Smoking Lesion (SL) *Susan is debating whether not to smoke or to smoke. She believes that smoking is strongly correlated with lung cancer, but only because there is a common cause - a lesion that tends to cause both smoking and cancer. Once we fix the absence or presence of this lesion, there is no additional correlation between smoking and cancer. Susan prefers smoking without cancer to not smoking without cancer; and she prefers smoking with cancer to not smoking with cancer. Should Susan smoke?*

Neither EDT nor CDT agents differ from their reasoning in NP.[41] The EDT agent recognises the auspiciousness of refraining from smoking, and therefore refrains, despite losing the small pleasure of smoking. The CDT agent, by contrast, notes that her actions cannot have any causal effect on whether she has lung cancer (see Fig. 1.3). As a result, she accepts smoking and its utility, and hopes that she does not have a lesion whose presence is beyond her control.

[40] Originally given by Stalnaker (1980). This version comes from Ahmed (2014, p. 90).

[41] Eells (1982, Chapter 7) argues that EDT should in fact endorse smoking, in an argument known as the *Tickle Defence*. For a comprehensive account of this defence, refer to Ahmed (2014, Section 4.3). Note that the EDT agent cannot avail herself of the tickle defence in XOR Blackmail Yudkowsky and Soares (2017, p. 25).

SL is not Newcomblike, in the same way that XOR-B is not Newcomblike. To wit: the event that causally effects an outcome with large negative utility (the presence of termites, or of the lesion) is not causally dependent on the nature of the agent's decisions. In XOR-B, this arises from the fact that the presence of termites is also beyond the influence of the predictor. In SL, this arises from the fact that there is no such predictor: though the lesion functions as a common cause of smoking and lung cancer (according to the agent), this alone is not sufficient for the lesion to be a predictor. Of course, the presence of the lesion *does* causally influence the nature of the agent by affecting the utility she attaches to smoking. In the causal framework discussed, however, this does not entail that in varying the nature of the agent we should also vary the presence of the lesion. The lesion is a causal non-descendant of the nature of the agent, and so its presence or absence is held fixed. For this reason, I take EDT to be mistaken and CDT to be correct on SL.

What the agent chooses in a given decision situation causally determines (up to a reliability constant) what prediction a predictor makes. Since we take our agents to operate according to a decision procedure that prescribes a choice when given a decision problem, this amounts to the predictor being causally influenced by the decision procedure that the agent employs. However, an agent's choice of decision procedure can causally influence neither the presence of a lesion, nor the presence of termites. We desire for our ideal decision theory to correctly capture these relations.

1.5 Functional Decision Theory (FDT)

Functional decision theory (FDT)[42] has recently emerged as a candidate for a decision theory that can correctly capture the decision-relevance of predictors and simulations. The basic intuition behind FDT is that there is some *decision-relevant*

[42] Proposed by Yudkowsky and Soares (2017) and Soares and Levinstein (2017), as a progression from Soares and Fallenstein (2015) and Yudkowsky (2010).

respect in which predictor-like things depend upon an agent's future action (such as in NP), and lesion-like things do not (such as in SL). Advocates of FDT denote their claimed decision-relevant notion of dependence as *subjunctive dependence*.[43]

In this section, I shall first suggest a precise definition of subjunctive dependence, and show how taking subjunctive dependence to be the decision-relevant notion of dependence may potentially yield the rational responses to TPD and XOR-B, insofar as they accord to the intuitions of one-boxers such as myself. However, I will conclude the section by arguing that, in spite of such successes, FDT cannot be the ideal normative theory of instrumental rationality. This is because FDT takes the *wrong kind* of counterfactual as decision-relevant when evaluating the outcome of a different decision procedure.

1.5.1 Subjunctive Dependence

Functional decision theorists imagine that the agent thinks of her decision process as *an implementation of a fixed mathematical decision function*. The function itself is a collection of rules and methods for taking the agent's desires and credences and selecting an action. In this sense, the agent's mathematical decision function captures precisely the aforementioned notion of the *kind* of decision-maker the agent is.

The functional decision theorist evaluates the consequences of her actions by evaluating the different hypothetical scenarios in which her decision function takes on a different *logical output*. Since the decision process is an implementation of a fixed function, this amounts to the FDT agent posing to herself for each action under consideration the question: "What if *this very decision process* produced a different conclusion?"

What does subjunctively depend on the agent's actions? The important cases

[43] The term originates from Drescher (2006). However, while Drescher defines subjunctive dependence by making modifications to EDT-style evidential dependence, Yudkowsky and Soares take the concept of subjunctive dependence as already given. Hence, despite being motivated by similar intuitions in Newcomblike cases, it is unclear whether or not the similarity between Drescher-style subjunctive dependence and FDT-style subjunctive dependence is merely verbal.

for this dependence involve predictors and simulations. Taking the latter as an example: if two physical systems are computing the same function, Yudkowsky and Soares say that the behaviours of these systems 'subjunctively depend' upon that function. The function's existence is posited as a common determinant of the behaviour of the two systems, analogously to Reichenbach's[44] inference of a common physical cause between two simultaneously correlated events.

Importantly, however, in the FDT agent's world-model, the function does not exist as a physical component of the universe, and the dependence relation between the function output and the system behaviour is *not* causal, in the sense of CDT. Instead, subjunctive dependencies are taken as a *strict superset* of causal dependencies. Moreover, evidential dependencies are taken as a *strict superset* of subjunctive dependencies.

Yudkowsky and Soares do not give a strict definition of subjunctive dependence that captures every instance of such a relation holding between two proposition. However, their explanation of the concept suggests a definition in which proposition B is said to *subjunctively depend* on proposition A just in case one of the following conditions holds:

(i) A and B describe physical events, and event B causally depends on event A.

(ii) A describes an equation governing the logical output of an agent's decision function on a given input, and B states a logical consequence[45] of A.

(iii) A describes an equation governing the logical output of an agent's decision function on a given input, and B describes the physical behaviour (the act) of the agent.

Condition (i) captures the idea that all causal dependencies are subjunctive dependencies. Condition (ii) describes the logical consequences of the logical output of the agent's very decision function having a different value. Condition

[44] Reichenbach (1991).

[45] The exact notion of *logical consequence* employed could arguably be taken as a free parameter for FDT. Typically, however, the term is used to refer to proofs in *classical logic*.

(iii) is intended to capture the link between the agent's representations of physical objects, and the agent's representation of logico-mathematical objects.

1.5.2 FDT in Action

Yudkowsky and Soares use Pearl's graphical apparatus in order to calculate the agent's credences about subjunctive dependencies. Whereas the nodes and edges in a CDT graph represented physical variables and the causal dependencies between them, in an FDT graph the nodes *also* represent logical function outputs and the subjunctive dependencies between them. As in CDT, we suppose FDT decision graphs contain a node OUTCOME, and an FDT agent seeks to maximise $\mathcal{U}(\textsc{Outcome})$. Unlike CDT graphs, FDT graphs will contain a node of the form $\textsc{fdt}(\underline{Cr}, \underline{G})$, a variable that represents the output of FDT when run with a credence Cr over graph G.[46] This node, by condition (iii), will always have an edge directed towards the agent's ACT node. Finally, we evaluate the expected utility of different actions by intervening on the $\textsc{fdt}(\underline{Cr}, \underline{G})$ node, as opposed to the ACT node. Therefore, expected utility for FDT is defined[47]

$$\mathcal{EU}_{\text{FDT}}(a) := \sum_{o \in \mathcal{O}} Cr(o \mid \mathtt{do}(\textsc{fdt}(\underline{Cr}, \underline{G}) = a)) \cdot \mathcal{U}(o). \tag{1.5.1}$$

Using this formalisation, we can now construct the graphs for TPD and XOR-B in order to determine FDT's prescription in each of the cases.

Twin Prisoners' Dilemma

The agent and her psychological twin in TPD are twins, in that they are two physical systems computing the same function. By Yudkowsky and Soares' claim, this means that the physical behaviour of the agent and the physical behaviour of the

[46] Here, we use underlines to represent dequoting, i.e. if $x := 3$ then $Z\underline{x}$ denotes the variable name $Z3$.

[47] I have simplified the formalisation of FDT given by Yudkowsky and Soares (2017), which also assumes that the FDT algorithm takes an *observation history* as another input. This complication is not necessary for my assessment of FDT.

psychological twin both subjunctively depend on the output of their shared decision function, $\text{FDT}(\underline{Cr}, \underline{G})$. Intervening on this node, it is clear that $\text{FDT}(\underline{Cr}, \underline{G}) = \textit{cooperate}$ determines that both the agent and her twin will cooperate, and $\text{FDT}(\underline{Cr}, \underline{G}) = \textit{defect}$ determines that both the agent and her twin will defect. Mutual cooperation, having higher utility for the agent than mutual defection, is therefore the output of the FDT algorithm: in other words, FDT agents rightly mutually cooperate in TPD. By Lewis' equivalence argument,[48] FDT agents will rightly one-box in NP.

XOR Blackmail

The important feature in this case is that TERMITES is *not* a descendant of $\text{FDT}(\underline{Cr}, \underline{G})$. Therefore, we may hold the value of TERMITES fixed while we consider the possibilities $\text{FDT}(\underline{Cr}, \underline{G}) = \textit{pay}$ and $\text{FDT}(\underline{Cr}, \underline{G}) = \textit{refuse}$. Refusing, being the better option regardless of the presence of termites, becomes the (rightly) advocated option in XOR-B.

1.5.3 The Problem of Counterlogicals

Ultimately, subjunctive dependence cannot be used as the decision-relevant notion of dependence for ideal decision theory. It does appear to capture the decision-relevance of predictors and simulations while avoiding the pitfalls of EDT. However, a closer inspection shows that the premise upon which subjunctive dependence is based can lead to counterintuitive consequences that make FDT an inadequate choice of decision procedure for an ideally rational agent.

To wit: the idea that the rational decision-maker evaluates different hypothetical scenarios in which the *same* decision function takes on different logical outputs is problematic. *Qua* well-defined mathematical function, it is logically necessary that the decision procedure has no more than one output for any given input. Therefore, in any decision problem in which the agent evaluates the ex-

[48] Lewis (1979).

pected utility of more than one potential output, all but one such evaluation will begin from a logical contradiction (setting aside ties for maximal expected utility). Under classical logic, by the principle *ex falso quodlibet* (EFQ), this means that any proposition follows as a logical consequence (in that it can be proven) of the supposition that the decision function has that particular output. Using condition (ii) of my proposed definition of subjunctive dependence, this means that we could use EFQ to recommend *any* action by 'showing' it has an arbitrarily high utility!

To elaborate further: consider a decision problem in which there are two distinct available actions, i.e., $\mathcal{A} = \{a, b\}$. Suppose further that actions a and b subjunctively determine outcomes o_a and o_b respectively, with $\mathcal{U}(o_a) > \mathcal{U}(o_b)$. In this case, FDT must prescribe action a. However, in its evaluation of the expected utility of action b, the FDT agent uses the do-calculus to make the node-value assignment $\text{FDT}(\underline{Cr}, \underline{G}) = b$. Since the true value of $\text{FDT}(\underline{Cr}, \underline{G})$ is a, this is a logical falsehood. By condition (ii) and EFQ, we may suppose this value assignment determines that $\text{OUTCOME} = o'_b$, where $\mathcal{U}(o'_b) > \mathcal{U}(o_a)$. Therefore, FDT should in fact prescribe action b, contradicting the original assumption.

We call a counterfactual whose antecedent is a logical falsehood a *counterlogical*.[49] Worries about reasoning counterfactually from contradiction are particularly salient when we represent FDT graphically, as Yudkowsky and Soares propose. If any proposition can be proved from a contradiction, then is there any limit on what the nodes of the graph can be? Are there any facts that we can hold fixed while evaluating the expected utility of a particular function (*de re*) producing a different output? If we cannot know what we are to hold fixed, then we have not given an adequate specification of the decision-relevant notion of dependence.

The problem of utilising counterlogicals in FDT is acknowledge by its advocates.[50] Yudkowsky and Soares dismiss the issue as being "technical rather than

[49] See Cohen (1990), Bjerring (2013), Bernstein (2016) for cases for the non-triviality of counterlogicals.

[50] For instance, see Yudkowsky and Soares (2017, pp. 7-8) and Soares and Fallenstein (2015, p. 12) for discussions. One proposed solution is to allow for an agent to reason probabilistically

philosophical":[51] they claim that humans are generally capable of reasoning in the face of uncertainty about logical claims, and in particular they are capable of deducing the 'true' subjunctive consequences of a decision function having a different logical output.

For the sake of argument, we can concede this point. After all, causal dependence is an archetypal example of a relation whose instances we can intuitively identify, yet which has eluded a precise metaphysical analysis.[52] Nevertheless, we do not take it to be a fatal flaw of CDT that we lack such an analysis.

However, the appearance of a problem as peculiar as the problem of counterlogicals suggests that FDT may be mistaken in its approach to reasoning counterfactually about different decision procedures. Rather than posing counterfactuals concerning the logical output of the agent's decision function, the agent could consider the consequences of her whole decision procedure changing. On a *de re* reading, the claim that the agent's decision function could prescribe a different action is evidently false, by the above argument. However, on a *de dicto* reading, the claim is more plausible. The *de dicto* reading has the further advantage that its corresponding counterfactuals reason from antecedents that, while false, are not logically impossible.

The *de dicto* reading accords better with the intuitions of one-boxers in TPD-like cases. An advocate for cooperation in TPD will make a claim of the form: 'Holding constant the knowledge that the agent and her twin employ the same decision function,[53] then whatever the agent does, her twin necessarily does the same thing.' On the *de re* reading, the agent's decision function is treated as a given part of the environment over which she has no control. On the *de dicto* reading, the agent *does* decide which decision function she employs. The assumption

about logical statements such as the output of her decision function. See Garrabrant et al. (2017) for a technical overview of this kind of doxastic reasoning. The above problem is a variant of the '5 and 10 Problem' discussed by Benson-Tilsen (2014).

[51] Yudkowsky and Soares (2017, p. 7).

[52] For a sample of the debate around causality, refer to Sosa and Tooley (1993) and Collins et al. (2004).

[53] This assumption about TPD is crucial to the claim.

in TPD on the latter reading is that, no matter what decision function the agent is employing, her twin employs the same one. I take this latter view to be a more plausible reading of TPD (and any situation involving predictors and simulations), and it is one I shall elaborate on in the next chapter.

Chapter 2

Policy-Based Causal Decision Theory

2.1 Policy Selection

The previous chapter investigated functional decision theory (FDT),[54] whose central claim was that *subjunctive dependence* is the decision-relevant influence relation between acts and outcomes. This relation included causal links, in addition to *logical links*. These links are an attempt to capture the line of counterfactual reasoning in TPD-like cases that the output of a deterministic algorithm varies with the output of any copies of that algorithm, *even if the actions of such copies are mutually causally independent*.

Such an approach was found unsatisfactory: the exact nature and description of situations whose nodes describe *function outputs* as well as physical acts and outcomes is ambiguous when we take the possibility of the decision function having a different output as *de re*.

However, I claim that we still ought to find a decision theory that views behavioural correlation between an agent and her simulators and predictors as arising from a non-evidential decision-relevant dependence. In this chapter, I argue that such a theory is possible within a framework where the model of our decision-theoretic agent contains only a description of the physical world and the causal relations between physical events in this world. Such an account will

[54] Yudkowsky and Soares (2017).

make a distinction between the physical *acts* of an agent and the *policy* of the agent that determines those acts.[55]

2.1.1 Three Problems for Policy Selection

Consider NP with a perfectly reliable psychologist. A regular causal decision theorist argues that, since *the physical act of taking one or two boxes* has no causal influence *ex hypothesi* on the prediction of the psychologist, we may hold the prediction fixed while investigating the outcome of having performed one of these physical acts. Regardless of the value at which we fix the prediction, the outcome of having taken two boxes is always $1,000 greater than the outcome of having taken one box. Hence we should take two boxes.

Previously, I asserted that we ought not to hold the prediction fixed while varying the act. Yet if we are taking a graphical approach to deciding on the act, and the act is indeed causally independent of the prediction, are we not justified in doing so? We must conclude that our decision is about more than just which physical act to take, and so the node in a decision graph representing our physical act is not, in fact, the node on which we should intervene.

Indeed, there is a sense in which we do intervene on a node distinct from the physical act. For when we make our decision about which act to take in NP, say, there is a sense in which we make a decision about which act to take in all such problems. For example, the CDT agent knows how she will behave in a Newcomblike situation by virtue of what CDT prescribes. Her decision about what to do in Newcomblike situations applies in all cases, not only the case she happens to find herself in. However, according to Joyce,[56] a CDT agent merely *wishes* she had the options of the kind of agent that one-boxes, rather than treating those options as actually available. This is mistaken.

Instead, I claim, choosing to one-box in NP for a decision-theoretic agent also

[55] Refer to Gauthier (1986, 1988, 1994), McClennen (1990), Meacham (2010), and Spohn (2012) for earlier accounts in this vein.

[56] Joyce (1999, p. 153).

means choosing to be (indeed, choosing to have been) *the kind of agent that one-boxes*. And it is the latter choice that results in a Newcomb predictor making its prediction that we one-box.

How ought we describe such a node? Recall that, in the previous chapter, each decision theory is described as a procedure that maximises an expected utility. The expected utility is itself a function of a decision problem $\mathcal{D}$, and the agent's credence and utility functions. Fixing the latter two functions, the decision-theoretic procedure is now a function from a *decision problem* to the *action* to be taken in such a problem. Call such a problem-to-action mapping a *policy*.[57] Intervening on a node that tells you what *kind* of agent you are, I claim, is the same as iterating over possible policies to find one that is optimal with respect to the right notion of expected value.[58]

Call a decision theory that intervenes on policies rather than acts a *policy-based decision theory*.[59] I will subsequently formalise a policy-based decision theory that takes causal dependence as the decision-relevant notion of dependence. First, however, I shall defend policy-based decision theory from three points of criticism that one may make against such an account.

The Domain of Control

Firstly, one might claim that it is not within the power of a decision-theoretic agent to choose which policy to employ. Instead, the agent is *given* by a policy, who then makes a choice about which acts to take in each decision problem in accordance with the policy. An FDT advocate, holding the agent's decision function fixed while varying its output, may support such a claim.

[57] The term is borrowed from the field of reinforcement learning by way of Soares and Fallenstein (2015). In reinforcement learning, a policy is a map from *observations* to actions. In this setting, the content of observations is encoded in the specification of the decision problem itself.

[58] FDT does incorporate this idea of policy selection by intervening on the output of a decision algorithm, which in turn determines the physical action as a separate node. However, I wish in this chapter to divorce the idea of policy selection from the problematic notion of subjunctive dependence between algorithm outputs and predictors, and work with a physical model.

[59] The approach has also been referred to by Greene (forthcoming) as treating the *decision theory* as the independent variable rather than the *act*.

However, I would argue that if we have the choice of which act to perform within each decision problem, we are in effect making a choice of policy, for a policy is simply an aggregate of decision problem-action pairs (*qua* function from decision problems to actions). Typically, the case for either one-boxing or two-boxing in NP is made as a result of principles which may be applied in a determinate fashion for other decision problems; consequently, the choice we make can be seen as a choice between sets of policies that share a particular problem-action pair. If we regard the agent as having the capacity to freely choose between all available actions in a given problem, we must likewise attribute to the agent the capacity to make the corresponding choice between these underdetermined policies. A supporter of CDT may grant this, while still regarding the dependencies stemming from the policy change as decision-irrelevant.

Changing the Problem

A CDT agent, if forewarned that she is to enter a Newcomblike scenario and given the opportunity to change her decision procedure, would self-modify to become the type of agent that one-boxes (like an EDT agent, say). Yet that fact is widely considered *irrelevant* to the question of what to do when one *already* finds oneself in a Newcomblike scenario. Consequently, one may argue that a decision theory which reasons about what it would be like to have a different decision policy in order to arrive at its prescriptions may be changing the decision problem.[60]

I reject this criticism. We can still specify a policy-based decision theory as a procedure for identifying an expected utility-maximising *act*, rather than *policy*. What changes in a policy-based account is the way in which one evaluates the expected utility of a given act.

[60] An analogous criticism is made by Arntzenius (2008) and Joyce (2012) about the invocation of mixed acts in the decision problem *Psychopath Button*, proposed by Egan (2007).

Rational Precommitment and Rational Irrationality

Thirdly, we may worry that a policy-based decision theory commits us to certain principles governing rationality that are widely regarded as implausible. I agree that policy-based decision theories do carry such commitments: however, I will show that they are more benign than they first appear.

The first concerns the idea of *rational precommitment*. Having chosen the optimal policy, the recommendations of that policy for which act to perform if and when one observes that one is in a given decision problem will be fixed both before and after that observation has been made (if it has been made at all). Hence, an agent following a policy will always act upon receiving such an observation as she would have precommitted to acting on that observation, *before* receiving said observation. Yet the principle

> 'if it is rational to precommit to something, then it is rational to predictably behave as though one has precommitted'

does not appear to be universally valid.

Consider the following counterexample:[61]

Greaves' Gym (GG) *The date is January 1st, 2018. I am offered a gym membership for this year costing \$600 that will give me the equivalent of \$550 in utility. Last year, an offer was made that if I filled in a preregistration form with the gym expressing an intention to get a 2018 membership, I would be offered a \$100 discount on signing up for that membership at the beginning of the year. I am uncertain as to whether I preregistered last year: in fact, I have only a credence p in having done so. Should I now sign up to the gym?*

In this case, the act of preregistering represents our precommitment. This act is evidently part of the rational course of action: by preregistering and signing up for the membership, an agent in GG stands to gain \$50 in utility, whereas not

[61] The set-up of the problem here in which one has uncertainty over one's prior actions is due to Hilary Greaves (personal communication). The original scenario and treatment of the problem is my own.

preregistering and signing, or not signing at all, result in -\$50 and \$0 in utility, respectively.

However, while it is rational in GG to precommit by preregistering, it is *not* rational to predictably behave as though one has preregistered. Indeed, the smaller p is, the less rational such behaviour becomes, since it becomes more likely that one stands to make a net *loss* in utility. CDT is capable of capturing this: preregistration occurs in the causal past to signing up to the gym, and so it can be held fixed. The reader may verify that

$$\mathcal{EU}_{\text{CDT}}(sign) = (p \cdot \$50) + ((1-p) \cdot -\$50) = \$(100p - 50), \tag{2.1.1}$$

so the CDT agent signs up to the gym just in case $p \geq 0.5$. In particular, CDT rightly treats p as a relevant factor when deciding whether to sign up to the gym.

On an initial reading, it would appear that a policy-based decision theory may neglect this nuance and behave irrationally. The rational agent is the one who preregisters and subsequently signs up for the membership: if a policy-based decision theory requires that one acts in accordance with the best possible policy, then the theory will endorse signing up regardless of the value of p. In other words, the principle (and policy-based decision theory) fail to account for the possibility that one has acted irrationally in the past.

A policy-based version of CDT can be rescued from this consequence, and can be shown to endorse the same rationale as CDT in GG. To show this requires a clear formalisation of a policy-based account, and so I defer the full treatment of the case until after such a formalisation has been given.

The second problematic principle concerns *rational irrationality*. In general, unlike CDT, adopting a policy-based decision theory precludes the possibility that a predictor can reward an agent for behaving in accordance with a policy that she takes to be irrational. This is because, on a policy-based theory, as soon as a large enough such reward is present, the policy-based theorist (unlike the

CDT agent) takes behaving in accordance with that policy to be rational.[62]

One may worry, then, that there are examples in which irrational behaviour clearly *is* rewarded. If so, then the rational course of action before entering such a decision problem would be to make oneself temporarily irrational. Yet a policy-based decision theorist cannot, as argued above, ever wish to be the kind of agent whose actions in a given decision problem would differ from her own: she bases her decisions on the best kind of agent she could be for a particular decision problem.

CDT advocates would claim that NP serves as an example: if a CDT agent could bind herself to one-boxing *before* the psychologist makes his prediction, then she would, even if she takes two-boxing to be the rational course of action in NP. Given that I hold one-boxing to be rational, NP cannot serve as our counterexample here. Instead, we may consider the following example as one in which policy-based decision theory may conflict with the rational course of action:[63]

Schelling's Answer to Armed Robbery (SAAR) *A man breaks into my house. He hears me calling the police. But, since the nearest town is far away, the police cannot arrive in less than fifteen minutes. The man orders me to open the safe in which I hoard my gold. He threatens that, unless he gets the gold in the next five minutes, he will start shooting my children, one by one.*

It would not be rational to give this man the gold. The man knows that, if he simply takes the gold, either I or my children could tell the police and make the number of the car in which he drives away. So there is a great risk that, if he gets the gold, he will kill me and my children before he drives away.

It would also be irrational to ignore the man's threat. There is a great risk that he will kill one of my children, to make me believe his threat that, unless he gets the gold, he will kill my other children.

However, I have a special drug, conveniently at hand. This drug causes one to be,

[62] Meacham (2010, p. 56) makes a similar point regarding EDT on NP, as a response to the claim that in NP the psychologist 'rewards irrationality'.

[63] Abridged from Parfit (1984, pp. 12-13).

> *for a brief period, very irrational. Before the man can stop me, I reach for the bottle and drink. Within a few seconds, it becomes apparent that I am crazy. Reeling about the room, I say to the man: 'Go ahead. I love my children. So please kill them.' The man tries to get the gold by torturing me. I cry out: 'This is agony. So please go on.'*
>
> *Given the state that I am in, the man is now powerless. He can do nothing that will induce me to open the safe. Threats and torture cannot force concessions from someone who is so irrational. The man can only flee, hoping to escape the police. And, since I am in this state, he is less likely to believe that I would record the number of his car. He therefore has less reason to kill me.*
>
> *While I am in this state, I shall act in irrational ways. There is a risk that, before the police arrive, I may harm myself or my children. But, since I have no gun, this risk is small. And making myself irrational is the best way to reduce the great risk that this man will kill us all.*

As Parfit claims, on any plausible theory about rationality, it would be rational for an agent in SAAR to cause herself to become for a period irrational by taking the drug. Given his description of the alternative options available to the agent,[64] I would agree with Parfit's claim that taking the drug is the rational option.

However, it is unclear to me why one's behaviour after having the taken the drug is irrational. After all, the series of actions described by Parfit *do* result in the outcome with the highest utility for the agent. In which case, a policy-based decision theory would seemingly treat such actions as part of the policy that yields the highest expected utility, making the actions *rational* by the agent's lights. How might we understand what mistake is being made in the claim that the 'irrational' actions *are* genuinely irrational in this circumstance?

The key to understanding the claim that the actions are irrational is in Parfit's assumption that the agent cannot make herself merely *appear* irrational.[65] To elab-

[64] I assume that one's choices in SAAR are exhausted by the three options *give the gold*, *refuse to give the gold*, and *take the drug*.

[65] Parfit (1984, Fn. 4a).

orate, we know that the home invader wishes to not get caught by the police, and will therefore kill the agent's children unless he can be *sure* that the agent will not record his number plate when he flees. The problem for the agent is that, once the home invader *has* fled, she will no longer face any repercussions from him for recording his number plate, and a world in which she does so is preferable to a world in which she does not (since the former makes it more likely that the man is brought to justice). Therefore, the drug is necessary in order to bind the agent not only to the actions she will take in the home invader's presence, but also to her choice to allow the home invader to avoid capture.

Yet even this final action will not be irrational by the policy-based theorist's lights: being a component of the ideal course of action, the policy-based theorist will dutifully neglect to record the number plate. This is true even if, by some miracle, the agent survives the ordeal unscathed without having had to take the drug - *ceteris paribus*, taking the drug and not recording the number plate is part of the ideal course of action, and so the policy-based theorist executes the second part of that plan.[66] Consequently, not recording the number plate is best understood as *irrational* only insofar as it is irrational *for a CDT agent*, who holds the past fixed. For a policy-based theorist, it is perfectly rational to not record the number plate.

Regardless, the policy-based theorist still chooses to take the drug - and rationally so! *Ex hypothesi*, taking the drug is the unique signal to the home invader that he can certainly avoid capture. We do not endow the home invader with the same predictive capacity of the psychologist in NP, and so taking the drug is the only way that the agent can signal her policy. All of the actions taken while under the influence of the drug are recommended by a policy-based decision theory, and are therefore rational. Hence SAAR is not a counterexample to policy selection.

[66] The virtue of a policy-based theorist being able to rationally execute previously-made assurances in this way reoccurs in Chapter 3.

2.1.2 Decision-Relevant Dependence for Policy Selection

Finally, it is worth asking whether, when iterating over different policies, we should consider the consequences of those policy choices by considering their causal effects, or by conditioning probabilistically on the event that the policy in question has been chosen. I argue that we should take the causal rather than the evidential approach. While both advocate one-boxing in NP, only the causal approach advocates smoking in SL. While being the kind of person who one-boxes in NP determines that that box will contain $1,000,000, being the kind of person who smokes in SL does not determine whether one gets cancer - only the presence of the lesion, which does not depend on one's policy, does. However, the correlation still exists, and so an evidential approach to policy selection will advocate refraining from smoking, hence advocating Lewis' *irrational policy of managing the news*.[67]

2.2 Formalising PCDT

We are now in a position to give a formal account of **policy-based causal decision theory** (PCDT). PCDT is a theory that accords with the intuitions of one-boxers by factoring in the behaviour of predictors and simulations, yet at the same time treats *causal* dependence as the decision-relevant notion of dependence between events.

Informally, then, our PCDT agent finds the best action by evaluating, for each possible policy, the *causal* expected utility of choosing that policy. Once she has found the optimum policy, she simply evaluates that policy for the decision problem at hand to find the appropriate course of action.

Formally, PCDT may be specified analogously to CDT. Take $\mathcal{D} = (\Omega, \mathcal{O}, \mathcal{S}, \mathcal{A})$ to be the decision problem for which we seek the correct action $a \in \mathcal{A}$. As in CDT, the agent comes equipped with a graph G representing the *causal structure* of $\mathcal{D}$:

[67] Lewis (1981b).

the nodes of the graph represent variables (with subsets of Ω as their ranges), with a node $\textsc{Outcome}$ taking values in $\mathcal{O}$ and a node $\textsc{Act}$ taking values in $\mathcal{A}$. A PCDT graph must contain a node $\textsc{Policy}$, representing the agent's decision problem-to-action mapping. As one might expect, any PCDT graph should contain a directed edge from $\textsc{Policy}$ to $\textsc{Act}$.

We specify the range of $\textsc{Policy}$ node by supposing that, for each act $a \in \mathcal{A}$ causally depending on the policy, $\textsc{Policy}$ may take as a value the proposition $\pi_{\mathcal{D},a}$ meaning 'be the kind of agent that in decision problem $\mathcal{D}$ takes action a.' If $\Pi \subset \Omega$ is the range of the $\textsc{Policy}$ variable, let us suppose for now that

$$\Pi = \{\pi_{\mathcal{D},a} \mid a \in \mathcal{A}\}.$$

In other words, all policies that agree with their prescribed action in the decision problem the agent finds herself in are considered equivalent. Finally, in order to assure the relationship between taking an action and being the kind of agent that takes that action, we stipulate for each $a \in \mathcal{A}$ that

$$Cr(\textsc{Act} = a \mid \textsc{Policy} = \pi_{\mathcal{D},a}) = 1. \tag{2.2.1}$$

The PCDT agent seeks to find the action $a \in \mathcal{A}$ that maximises

$$\mathcal{EU}_{\text{PCDT}}(a) := \sum_{o \in \mathcal{O}} Cr(o \mid \texttt{do}(\textsc{Policy} = \pi_{\mathcal{D},a})) \cdot \mathcal{U}(o). \tag{2.2.2}$$

As with regular CDT, we can employ Pearl's framework[68] to evaluate $Cr(o \mid \texttt{do}(\textsc{Policy} = \pi_{\mathcal{D},a}))$, which is interpreted as the agent's degree of belief that her being the *kind* of agent that takes action a will effect outcome o.

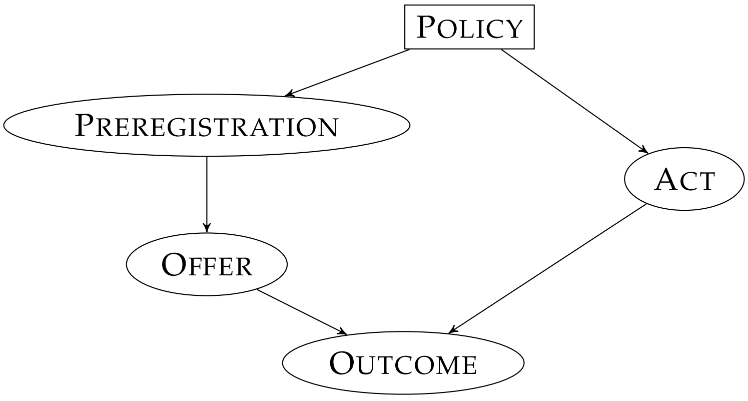

Figure 2.1: A PCDT graph for Greaves' Gym. This may be modified to a CDT graph by shifting the node of intervention from POLICY to ACT.

Revisiting Greaves' Gym

We are now in a position to discuss the response of PCDT to GG, and the relation such a response has to idea of *rational precommitment*. Fig. 2.1 depicts the causal graph for GG. The same graph may be used for the CDT agent, though with the intervention on ACT rather than POLICY. Let $sign$ and $refuse$ be the available acts of signing and refusing to sign, respectively. A PCDT agent must evaluate the outcome of the policies $\pi_{\mathcal{D},sign}$ and $\pi_{\mathcal{D},refuse}$ in order to make her decision. What of her uncertainty over whether she has preregistered? PCDT can accommodate this by using the following conditional credences:

$$\begin{aligned}
Cr(\text{PREREGISTER} = \textit{yes} \mid \text{POLICY} = \pi_{\mathcal{D},sign}) &= p \\
Cr(\text{PREREGISTER} = \textit{no} \mid \text{POLICY} = \pi_{\mathcal{D},sign}) &= 1 - p \\
Cr(\text{PREREGISTER} = \textit{yes} \mid \text{POLICY} = \pi_{\mathcal{D},refuse}) &= p \\
Cr(\text{PREREGISTER} = \textit{no} \mid \text{POLICY} = \pi_{\mathcal{D},refuse}) &= 1 - p.
\end{aligned}$$

With the credences assigned thus, both possible values of PREREGISTER must be considered, with the utility of the overall outcome weighted accordingly. This leads to precisely the same evaluation of expected utility as in the CDT case described earlier.

[68] Pearl (2009).

The crucial step here is in PCDT treating two policies as distinct only insofar as they disagree on the act presently being decided upon. Hence, $\pi_{\mathcal{D},sign}$ denotes the proposition that one signs up for the gym membership, and says nothing about whether one has already preregistered.

One might charge that the reliance on such a step is *ad hoc*. It is equally plausible, a critic may add, that a PCDT agent should be able to further fine-grain the available policies so as to treat preregistering and signing ($y\&s$), and *not* preregistering but still signing ($n\&s$), as two distinct possibilities, $\pi_{\mathcal{D},y\&s}$ and $\pi_{\mathcal{D},n\&s}$ respectively. Yet if one does so (and continues to assume a credence 1 in acting according to your policy), we again arrive at $y\&s$, and hence signing, as the rational action.

I claim in response that the step is only as contrived as the original decision problem. If we are stipulating a genuine uncertainty over the previous act of preregistering, we are effectively taking that act as being independent of the policy that we exercise control over. Taking the contrapositive, if our choice of policy is sufficiently fine-grained so as to distinguish preregistering and not preregistering, then we cannot assume as in the decision problem that we are in fact uncertain as to how we have acted. As a result, it is the particular structure of GG that imposes a more coarse-grained partition of the available policies, rather than some external, *ad hoc* principle. PCDT is thus better described as adhering to the principle of rational precommitment only in the cases in which the agent is not genuinely uncertain about her past actions.

2.3 Comparison with Related Accounts

2.3.1 Causal Decision Theory

The key difference between CDT and PCDT is what we take our *choice* to be about, and so using Pearl's graphical apparatus, the *node* upon which we intervene. In both cases, however, we are interested only in the *causal* effects of such an inter-

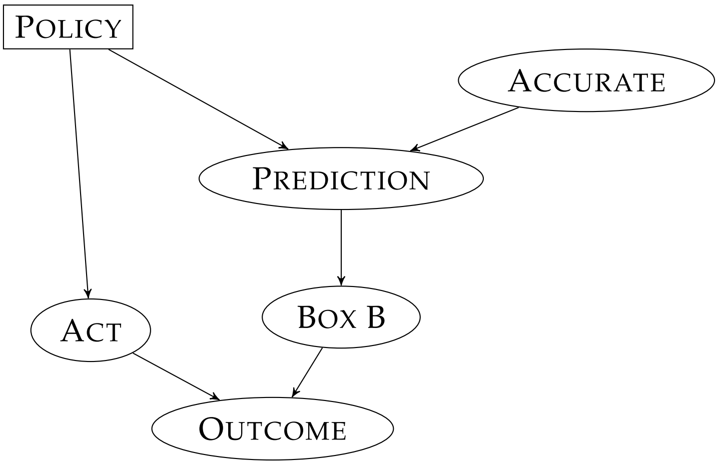

Figure 2.2: NP, as modelled by a PCDT agent. The sole difference is the node at which the agent intervenes.

vention.

Figs. 1.1 and 2.2 show this distinction at work in NP. Both CDT and PCDT employ the same graph, with the same variables.[69] The difference stems from the node at which one intervenes to calculate the relevant causal probabilities, represented by the node having a rectangular (as opposed to round) border.

In the case of CDT, our choice for ACT has no causal effect on the value of BOX B, whose value stems from that of PREDICTION, which in turn takes its value solely from the POLICY node. This allows us to hold the value of BOX B fixed, while looking at the effect of ACT on OUTCOME. Since setting ACT $= twobox$ always yields a greater pay-off than ACT $= onebox$ regardless of the value of BOX B, we conclude that CDT prescribes two-boxing.

PCDT, however, intervenes on the POLICY node, which takes the values $\pi_{\mathcal{D},onebox}$ and $\pi_{\mathcal{D},twobox}$. If the psychologist is *perfectly* reliable, setting the value of POLICY completely determines the values of all of the other nodes, from which we easily see that setting POLICY $= \pi_{\mathcal{D},onebox}$ (which in turn tells us what our action should be) is the most lucrative option.

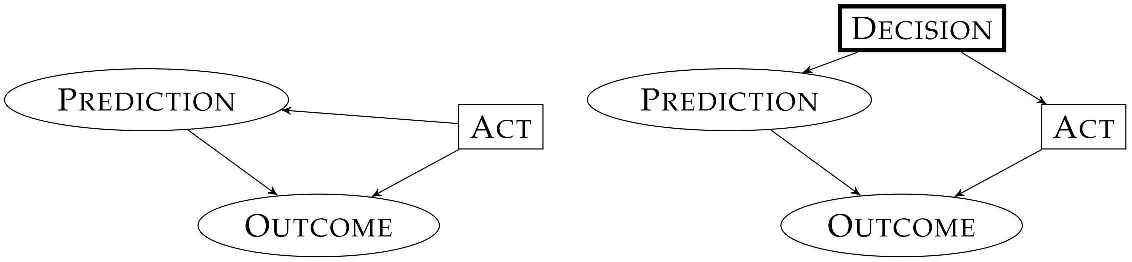

Figure 2.3: Spohn's reduced (left) and reflexive (right) causal diagrams for NP.

2.3.2 Spohn's CDT

Spohn[70] defends a different account that attempts to defend the compatibility of CDT with the act of one-boxing in NP. Fig. 2.3 shows his two graphical representations of the problem.

According to Spohn, the right-hand diagram gives the full picture for NP: a 'reflexive decision graph' in which intention to perform a realisation of ACT, modelled by DECISION, is the common cause of both the act and the prediction. The graph is *reflexive* insofar as it reflects on the ordinary decision situations deliberating on and causing the actions, in addition to the relevant action and non-action (elliptical) nodes.[71]

The left-hand diagram represents a 'reduced' version of this graph, which the agent may use to determine the appropriate course of action in the decision problem. Spohn shows how we may systematically 'reduce' a reflexive decision graph to a regular Pearl-style decision graph, deleting any bold-rectangular nodes that represent the decision situation itself and drawing new arrows that capture the influence that stems from the agent's deliberation on how to act leading to the decision problem. In particular, the new arrows make any non-rectangular child of the DECISION node directly causally dependent on all the parents and all the rectangular children of the DECISION node.[72]

This leads to a graph containing arrows that may seemingly be interpreted as cases of backwards causation, as shown by the arrow drawn between ACT and

[69] The design of the graph is from Yudkowsky and Soares (2017).
[70] Spohn (2012).
[71] Spohn (2012, p. 100).
[72] Spohn (2012, pp. 118-119).

PREDICTION in the left-hand diagram. Yet Spohn insists that there is no such thing as backwards causation. Instead, the new arrow is instrumentally useful in showing the connection between the action and prediction, and is justified by the implicit presence of a bold-rectangular node as the true common cause between ACT and PREDICTION. Querying the left-hand diagram, we arrive at the conclusion that one-boxing is the rational option.

The motivation behind Spohn's account and my own remains the same - the common cause between one's act and any prediction of one's act is the decision rule that leads to the act, something over which one has control. Indeed, this shared motivation leads me to believe that the decision theory I have independently arrived at turns out to be equivalent to that of Spohn's.[73]

However, I believe my exposition to be clearer than that of Spohn's. In PCDT, the values of the POLICY node are transparent propositions of the form $\pi_{\mathcal{D},a}$, which state that the agent's policy is such that she takes action a in decision problem $\mathcal{D}$. In Spohn's account, this meaning is far more opaque: the values of the DECISION node are 'decision situations', essentially causal Bayesian networks for use in a Pearl-style decision theory. Moreover, the (earliest) DECISION node takes only one value, δ_0, with probability 1, representing the reduced version of the present decision problem with the DECISION node deleted. The shape of any other value of a DECISION node[74] is left as the subject of further theoretical work.

Moreover, Spohn merely sketches what the relevant decision rule is that prescribes for each decision situation δ that is a realisation of a DECISION node a rational action.[75] Yet, as Spohn must make use of such a decision rule in order to determine the relevant conditional probabilities of any acts that directly causally depend on that decision situation, it makes his account far more unwieldy. By

[73] This may turn out to be *false* if Spohn disagrees on 'veil of ignorance'-type cases discussed in Chapter 3. However, it is unclear from Spohn (2012) how his version of CDT would respond to such cases.

[74] Spohn allows for more than one bold-rectangular DECISION node in a given graph, in order to accommodate cases in which one's credence and utilities change as a result of previous actions. As I shall mention in Chapter 4, I leave this as an open area of development for PCDT. One might ask in particular whether multiple POLICY nodes are the best model for credence and utility shifts.

[75] Spohn (2012, pp. 120-121).

contrast, the prescribed action of a given policy in PCDT can be clearly read from the notation.

2.3.3 Meacham's Cohesive Expected Utility Theory

Meacham[76] claims that it is a regrettable feature of any decision theory that an agent employing it would act differently as she would have bound herself to act if she had had the prior opportunity to do so.[77] In other words, Meacham takes the consequence of PCDT that SAAR seeks to refute (which he calls *cohesiveness*) as a *feature*, rather than a bug. To show that this regrettable feature need not always be the case for a decision theory, he proposes a kind of policy-based decision theory.

In Meacham's terminology, a *decision problem* is an ordered triple consisting of an agent's current credences, utilities, and the set of available acts. A *comprehensive strategy* is a function which maps every decision problem to one of its available acts - in other words, it is a policy that is specified not only for the decision problem at hand, but for *every* decision problem. The agent then seeks a policy π to maximise her *cohesive expected utility*, defined as

$$\mathcal{EU}_{\text{CoDT}}(\pi) := \sum_i Cr(ICr_i) \sum_{o \in \mathcal{O}} ICr_i(o : \pi) \cdot \mathcal{U}(o). \tag{2.3.1}$$

Here, the $ICr_i(\cdot)$ represent the agent's possible 'initial credence' functions (with i ranging over such functions), and the credence $ICr_i(o : \pi)$ uses ':' as an intentionally ambiguous symbol between different readings of decision-relevant dependence. As previously argued, however, a rational response to NP, XOR-B, and SL demands that this dependence relation is taken as a causal one. Furthermore, we must now assume that the set of propositions $\mathcal{O}$ is sufficiently fine-grained so as to respect the difference in outcomes in *all* decision problems, rather than the decision problem in question, since we are choosing a policy rather than

[76] Meacham (2010).

[77] This is as a response to Arntzenius et al. (2004), who is taken to support the principle: *If a theory of decision making has a counterintuitive result that only arises for agents who cannot bind themselves, this result is not a mark against the theory of decision making in question.*

an act.

What are the agent's initial credences, and why ought we use them as part of the decision procedure of an ideally rational agent? Meacham develops his account with respect to the first question, yet has little to say with regard to the second. He proposes that one takes ICr to be the agent's *ur-priors*: the credences that an agent ought to have if she had no evidence whatsoever.[78]

It is more difficult to motivate an answer to the second question. Indeed, as Greene[79] has noted, when reasoning about cases such as NP and TPD, we are reasoning causally from a different *temporal* perspective, rather than a different *evidential* perspective. Therefore, even if an agent's ur-priors could be specified in a non-arbitrary way, by disposing of them in our reasoning we fail to capture the truly relevant factors in a given decision problem. PCDT, using the agent's current credences to reason causally from a different choice of policy by the agent, does not suffer from this difficulty.

2.3.4 Graphical Updateless Decision Theory

Finally, I compare PCDT to **updateless decision theory** (UDT).[80] According to UDT, we choose an optimal policy by calculating the expected value of adopting that policy - this effectively gives an expected utility formula that we seek to maximise that is formally identical to (2.2.2). However, as in the case of FDT, the arrows in a Graphical UDT graph represent "logical relations", in addition to causal relations.[81] These logical relations tell us the consequences of a given policy, considered as the 'logical output of the agent's decision algorithm'.[82] Hence Graphical UDT is a policy-based, or *updateless*, version of FDT.

While we should consider our decision to be about *policy* rather than *actions*,

[78] Meacham (2010, p. 70, Fn. 34). As Meacham goes on to note, objective Bayesians will hold that all agents have the same ur-prior, while subjective Bayesians will hold that different agents can have different ur-prior functions.

[79] Greene (2013, pp. 37-41), Greene (forthcoming, pp. 15-17).

[80] As mooted in Soares and Fallenstein (2015) in tandem with Pearl's graphical apparatus. The original UDT proposal is due to Dai (2009).

[81] Soares and Levinstein (2017, p. 2) consider FDT as a generalisation of UDT.

[82] Soares and Fallenstein (2015, p. 9).

our story about the consequences of our policy choice should be purely causal, with the policy node itself standing for a physical representation of the policy as opposed to the abstract, logical representation favoured by UDT.[83]

Consequently, despite their similarities, the prescriptions of PCDT and Graphical UDT may in fact diverge for a certain class of decision problems. Indeed, PCDT is the decision theory that one obtains if one rejects the claim of Soares and Fallenstein that the logical/subjunctive relations between events is what matters in decision-making. This claim is motivated in response to the following decision problem:[84]

Retro Blackmail Problem (RBP) *There is a wealthy artificially intelligent agent, and an honest AI researcher with access to the agent's original source code. The researcher may deploy a virus that will cause $150 each in damages to both the agent and the researcher, and which may only be de-activated if the agent pays the researcher $100. The researcher is risk-averse and only deploys the virus upon becoming confident that the agent will pay up.*

The agent knows the situation and has an opportunity to make a certain self-modification after the researcher acquires her original source code but before the researcher decides whether or not to deploy the virus. (The researcher knows this, and has to factor this into their prediction.) The effect of this self-modification is that the agent will always refuse to pay if she is being blackmailed in the manner above. Should the agent self-modify?

In this case, such a self-modification is advisable. According to Soares and Fallenstein, the researcher will consequently deduce from the agent's original source code that she is the type of agent which would self-modify so that she always refused to pay up, and so would not deploy the virus in the knowledge that she

[83] It is telling that Soares and Fallenstein do not pursue the the graphical approach to UDT very far, arguing that there is no principled, formalised way to construct the right 'logical' graph, as there is for Pearl's CDT. Nevertheless, the graphical approach re-emerges with the same issues highlighted for FDT in Yudkowsky and Soares (2017). I discuss some of the issues raised by Soares and Fallenstein about graphical representations of decision algorithms later in this chapter.

[84] Soares and Fallenstein (2015, p. 6).

would lose $150 if she were to do so.

Soares and Fallenstein claim further that a policy-based CDT agent would *not* choose to self-modify. Their argument is as follows: since the behaviour of the copy of the agent that the researcher reasons about is causally independent to that of the agent herself, performing a self-modification would not have any impact over whether the researcher chooses to blackmail her. Indeed, if the agent did self-modify, then in the case where the researcher does choose to issue the blackmail (considered possible due to the presumed causal independence between the agent's and the researcher's choice), then the agent would lose $150, rather than potentially getting away with losing over $100.

While this account seems to hold of the regular CDT agent, I reject the claim that a PCDT agent would not self-modify. On the contrary, PCDT successfully captures the result that it is rational to self-modify. A PCDT agent, evaluating how to respond in the case of a blackmail, will reason that choosing to pay up effectively guarantees that such a blackmail will occur. Hence, she will advocate not paying up in the case she receives the blackmail. This will be reflected in the *original source code* of the PCDT agent, provided that that original source code determines that the agent will act according to PCDT.

The original source code is what we represent by the POLICY node in PCDT. This source code has a causal link to the policy of her copy, since the copy originates from the original source code. Hence, the original source code acts as a common cause of the behaviour of the copy (and hence, the behaviour of the researcher), and of the action that one takes after the copy has been made.

Soares and Fallenstein picture an agent reasoning causally, even if deciding on the policy she has always had, to have no causal influence over her original source code, even if her eventual choice in the decision problem is to be reflected in the original source code. This seems incoherent. Soares and Fallenstein assume the decision the agent makes will be reflected in her original source code. Yet if this is the case, then the original source code is what the POLICY node represents, as it is meant to tell you about the *kind* of agent you are with respect to how you choose to act in the given decision problem.

I hold RBP to not be relevantly different to NP: the agent's choice of action is reflected in her policy, which has a causal influence over the psychologist. In both cases, the actionable policy node should represent the original source code, which in RBP has a causal influence over the researcher by way of the copy made of the source code.

To elaborate, if we take the original source code of the agent to be the representative of the POLICY node, we may use our graph for NP to construct a graph for RBP, making the substitutions

$$\text{PREDICTION} \leftrightarrow \text{COPY}$$
$$\text{BOX B} \leftrightarrow \text{BLACKMAILER}.$$

It is worth noting that this equivalence demonstrates that CDT does *not* prescribe self-modification. To elaborate: the CDT agent in RBP is causally independent to her copy, and so she is also causally independent to the researcher's choice of whether or not to send the blackmail. This means that we can hold the researcher's decision *fixed*, assigning her probability of sending the blackmail as p, while varying the choice of whether to self-modify.

If the agent does self-modify so as to bind herself to refusing the blackmail, her expected utility in doing so is $-\$150p$. If she does not self-modify, then if she receives the blackmail she will act according to CDT. Since paying up is causally efficacious for de-activating the virus, and not paying will mean the virus damages the agent, CDT prescribes paying if the blackmail is issued and the virus is deployed. This means that not self-modifying has an expected utility of $-\$100p$. Hence, a CDT agent will not self-modify in such a circumstance.

2.4 What a Policy Node Represents

Though Soares and Fallenstein do not discuss the issue in great detail, I take the crux of our disagreement in RBP to be the temporal location of the POLICY node upon which the agent has the ability to intervene. As such, if we are to construct a graph for a given decision problem that (exclusively) describes the causal relation between the agent's policy, her physical action, and other simulations and predictors, it should be made precise exactly what the policy node *represents*, in addition to the temporal location at which it is located.

As such, here is my proposal: **The POLICY node stands for the earliest representative of the physical system that faithfully implements the decision procedure that the agent employs**.

The principal motivation of the proposal is as follows. We wish for the POLICY node to stand for a configuration of particles that actually *represents and implements* the decision procedure that the agent employs.[85] Without this stipulation, then in a deterministic universe, we might imagine that the POLICY node would simply represent the configuration of particles at the initial state of the universe![86] This is because, under the assumption of determinism, one could infer precisely what the agent's policy would be from an observation of the initial state of the universe.

This would be an unfortunate construal of the term *policy*.[87] In the initial motivating examples for a policy-based decision theory, I claimed that there was a sense in which the agent had control over the *kind* of agent she was, insofar as this is defined with respect to how she chooses to act in a given decision problem. The scope of this control does not, however, extend to facts about the universe

[85] For a discussion of what it means for a physical system to represent an abstract computation, see Piccinini (2015). Chalmers (1994) provides an alternative account of implementation in which a physical system implements a computation if the causal structure of the system mirrors the formal structure of the computation. Both accounts are non-trivial, in that no physical system is said to implement *every* abstract computation.

[86] Thanks to Hilary Greaves for pointing this out.

[87] This issue does not seem to occur if we do not assume determinism. If this is the case, then the agent's decision procedure is not guaranteed to be determined completely until her inception in the physical universe.

that have a causal influence over the nature of the policy that obtain before the policy has been designed. Concretely: suppose in RBP that the agent's original source code was written by another researcher. Whereas we may imagine that the agent has control over her original source code (according to which she is currently operating), we would not extend that control to be over the researcher who originally wrote that source code.

My proposal precludes this construal: even if knowledge of the initial state of the universe is sufficient to determine what the policy is, that initial state does not *implement* the policy under any satisfactory account of implementation, nor does it *represent* a future implementation under any satisfactory account of representation.

Using the proposal, we see that if the agent is an artificially intelligent system, the node will represent the agent's original source code. The medium of this source code is irrelevant - it can be the (physical implementation of) a `.py` file or pseudocode written on a napkin - provided that there is a causal nexus to the physical system implementing the code. This physical system will itself, if the agent is a digital computer, be some collection of transistors. If the agent is a human, then the node will consist of some collection of neurons.[88]

This proposal compels us to be much more precise in specifying the physical story of any decision problem. For example, consider the earlier case of TPD, in which you are an agent playing against a distinct physical agent employing the same decision algorithm. FDT may identify such a twin as a 'logical copy' of the agent, whose actions will necessarily co-vary with the actions of the agent herself. In contrast, I claim that such a story is underspecified: we need to know just what is it about the physical origin of these two agents that makes one a 'logical copy'

[88] You might worry that, in the Smoking Lesion (SL) case, the lesion cannot be considered distinct to the policy, due to its being an influential factor in whether you smoke. This would be problematic for PCDT: in subsuming the lesion into POLICY, we would pretend to assume control over the presence of the lesion and thus over the presence of cancer, which would lead to the irrational decision to refrain from smoking. Though it intuitively seems to be the case that we *should* treat the lesion as merely an influential factor in our policy rather than an essential component of it, the question of where in general to draw the line appears worryingly arbitrary.

of another. Consider these two potential scenarios:

Scenario 1 A professor writes a precise decision algorithm on a whiteboard. Two graduate students see this writing, and each one faithfully implements the same algorithm on two distinct physical systems. One of these systems represents the agent, and one represents the 'logical copy' of the agent. Both employ the same decision procedure.

Scenario 2 A professor writes the instruction 'Design an agent employing your favourite decision theory' on a whiteboard. Two graduate students see this writing, and each one, without having communicated to the other, implements the same algorithm on two distinct physical systems. One of these systems represents the agent, and one represents the 'logical copy' of the agent. Both employ the same decision procedure.

The graphical representation of the decision problem in PCDT identifies the POLICY node with the earliest configuration of particles completely determining the policy that the agent employs. Assume that our graduate students cannot make any mistakes, and will faithfully follow any instructions left on the whiteboard. In the first scenario, then, it is the fully-specified decision algorithm written on the whiteboard that completely determines the agent's policy, and so the whiteboard writing is what the POLICY node represents. The infallibility assumption will also entail that there is no possible world in which the graduate students implement *distinct* algorithms.

In the second scenario, however, the POLICY node cannot stand for the professor's instruction, since it is not a faithful implementation of the decision procedure that the agent will employ. Instead, the POLICY node will stand for the design of the agent as devised by just *one* of the graduate students. Hence, under the assumption that our graduate students are infallible, it is not necessarily true (either logically or metaphysically) that the same decision algorithm will be employed by the two agents. In particular, there exists a possible world in which

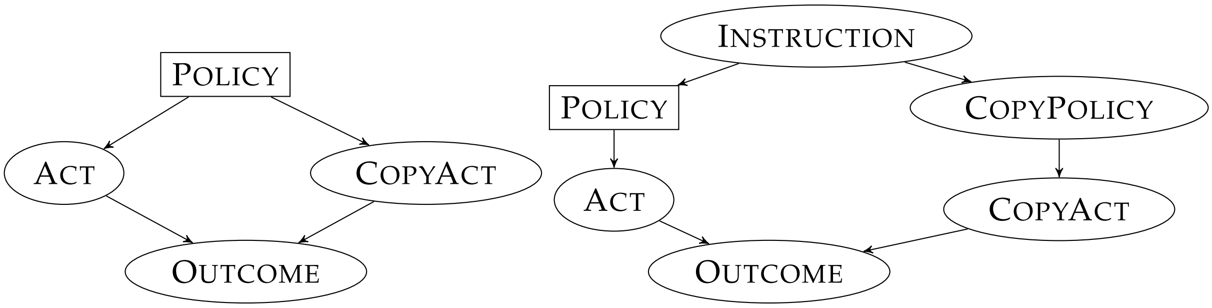

Figure 2.4: A comparison of the causal graphs depicting Scenario 1 (left) and Scenario 2 (right) for TPD.

the two graduate students prefer different decision theories, and so implement different decision procedures.

This will result in different decision graphs, and hence different recommendations, for the two disambiguations of TPD. The graphical representation of the two disambiguations can be seen in Fig. 2.4.

To summarise, policy-based causal decision theory (PCDT) is a procedure that captures the idea that, under certain circumstances, an agent has an influence over predictors and simulations that she ought to take into account in her decision-making process. In contrast to FDT and Graphical UDT, it does so while taking into consideration only the *causal* effects that one's choice of policy may have, making it more readily specifiable. Moreover, I have made clear the key modelling choice that one must make in constructing the appropriate decision graph: namely, the choice about what physical system the POLICY node over which the agent has power represents.

Chapter 3

The Veil of Ignorance

One of the challenges of decision theory is in disambiguating an ordinary-language decision problem: depending on how one formalises a particular problem, the same decision theory may prescribe different outcomes. Though unavoidable to some extent, we should be wary of decision theories that demand of a seemingly clear decision problem too many hidden structural assumptions in order for a clear outcome to be specified, and we should be even more wary if the outcome is overly dependent on seemingly unmotivated assumptions.

In this chapter, I consider a decision problem in which PCDT appears to possess this disadvantage over CDT and EDT. In response, I motivate a principle that may be applied to disambiguate such decision problems for PCDT, and show this principle to be advantageous in other cases.

3.1 The Curious Benefactor

Consider the following problem:[89]

The Curious Benefactor (CB) *A wealthy psychologist decides to play a game with you. He flips a fair coin. If it comes up tails he will ask you to pay him $5. If it comes up heads, he will give you $1,000,000, but only if he predicts that you*

[89] Problem originally due to Nesov (2009), under the name *Counterfactual Mugging*. This version is adapted from Hintze (2014, p. 3).

> *would have given him the \$5 if the coin had come up tails. He then flips the coin and it comes up tails. He explains the situation to you, and asks you for the \$5. Should you give him the \$5?*

In CB, the prescriptions of both CDT and EDT coincide. An agent following either decision procedure must condition her credences on the evidence she has observed, viz., the coin having come up tails.[90] *Given that the coin has come up tails*, both CDT and EDT agree that giving the \$5 yields a lower expected utility than not giving the \$5. Hence both CDT and EDT prescribe not giving the \$5. Conditioning on the coin's having come up tails effectively holds that fact fixed, rendering the hypothetical scenario in which the coin comes up heads irrelevant.

Does PCDT recommend the same course of action? It is not immediately clear. Recall that PCDT, rather than evaluating interventions of the form $\textsc{Act} = a$, instead evaluates interventions of the form $\textsc{Policy} = \pi_{\mathcal{D},a}$. In CB, the agent makes her choice about the $\textsc{Policy}$ node before coin is flipped. We therefore have a choice between two modelling assumptions when propagating the value of $\textsc{Policy} = \pi_{\mathcal{D},a}$ forwards: we either (i) respect our observation of the actual coin outcome and fix the value of $\textsc{Coin}$ as tails; or we (ii) forget our observation of the coin outcome, and model $\textsc{Coin}$ as a stochastic variable taking values either h or t each with credence $1/2$.

Our choice between (i) and (ii) is crucial. If we choose (i), then PCDT advocates not giving the \$5, by a similar line of reasoning to that of CDT. However, if we choose (ii), PCDT advocates giving the money. This is because we must now take into account what would happen if the coin came up heads instead of tails:

$$\begin{aligned}\mathcal{EU}_{\text{PCDT}}(\textit{give}) &= \left(\frac{1}{2} \times -\$5\right) + \left(\frac{1}{2} \times \$1,000,000\right) = \$499,997.50,\\ \mathcal{EU}_{\text{PCDT}}(\textit{refuse}) &= \left(\frac{1}{2} \times \$0\right) + \left(\frac{1}{2} \times \$0\right) = \$0.\end{aligned}$$

[90] Both CDT and EDT agents condition on all evidence they have collected up until the present. This is a feature of the theories I have thus far omitted to mention, though the expected utility formulas for both theories may be straightforwardly adjusted to reflect this requirement. For an account of how the CDT and PCDT agents update their causal credences on receiving evidence, refer to the Appendix.

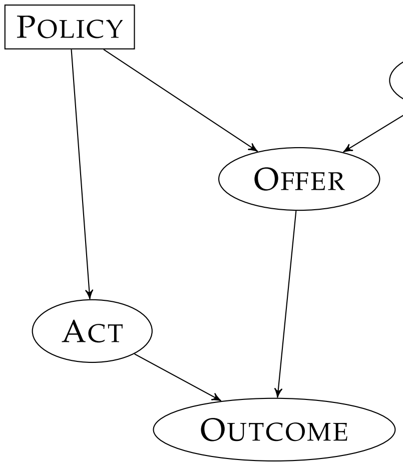

Figure 3.1: CB, as modelled by a PCDT agent. Modelling choices (i) and (ii) correspond with the range of COIN being $\{t\}$ and $\{h, t\}$, respectively.

Responding that the PCDT prescription depends on the formalisation of the problem won't cut it. As we have seen, this nuance in modelling CB is not present in CDT or EDT, which condition on all observations up until ACT. In order to defend PCDT, one must therefore propose some well-motivated principle whose purpose it to disambiguate *all* such cases.

3.1.1 Generalising the Problem in CB

Call the time that lies strictly after the temporal location of the POLICY node and strictly before the temporal location of the ACT node the *pre-act post-policy* (PAPP) time.[91] Suppose that, at the time of the decision problem, the agent has observed that a set X of variables located in the PAPP time has value x. The question for PCDT is as follows: should a PCDT agent condition her credences on the fact $X = x$ when evaluating $\mathcal{EU}_{\text{PCDT}}$? In other words: where should one draw the *veil of ignorance* over observations in the PAPP time?

I claim that the PCDT agent should *not* condition her credences in $X = x$. In particular, if the value of X is defined according to some joint probability distribution over the variables in X, the agent should treat the value of X as if it is drawn randomly from that distribution, even if the agent knows the actual out-

[91] We may assume without loss of generality that, in an agent's causal graph, no two variables are temporally co-located.

come for X. Therefore, in CB the agent should take choice (ii) and give the \$5. Better luck next time! I shall now defend this principle from two criticisms.

Susceptibility to Blackmail

One might worry that, like with XOR Blackmail and Retro Blackmail, an agent who is predisposed to paying in CB can be duped by a psychologist into giving all of her money away. For example, the psychologist might flip his coin until it comes up tails, and then propose the problem to an unwitting PCDT agent to extract \$5, at no monetary cost to the psychologist. Given that a scam of this form is perfectly avoidable, it would be irrational if PCDT were to reliably lose money through such a scheme.

However, if the psychologist were to behave in the way described in the previous paragraph, it would create a different decision problem. Applying PCDT to *that* decision problem, we see that in *that* case PCDT would not recommend giving the \$5, although it does in the original CB case.

Only the Actual World Matters

The biggest issue for PCDT is that it is *prima facie* irrational to ignore one's evidence about the state that one in fact finds oneself in. Why should a possible world in which one *knows* one is not located have any relevance when choosing an action? To quote Gandalf:[92]

> *All we have to decide is what to do with the time that is given us.*

Of course, there is some sense in which PCDT takes into account the actual observations. In particular, having chosen a policy, the PCDT agent uses that policy to decide which act to take depending on what *actual* observation is made.

Yet this is not enough to defend PCDT from this criticism. For while PCDT may *act* on its actual observations, the claim is that a PCDT agent is acting irrationally by considering what would be the case were she to have made observa-

[92] Tolkien (1954, p. 60). Wedgwood (2013) proposes a decision theory based upon this principle.

tions she *knows* she has not made. This is the kind of behaviour alluded to by the original name of CB, *Counterfactual Mugging*: the idea that one should make a sacrifice that will only pay out in a world that is not the actual one.

I believe that, in cases such as CB, one should simply bite the bullet and take as decision-relevant a state of the world one knows one does not find oneself in. Regrettably, I see no compelling argument that endorses such behaviour in the specific case of CB. Instead, in order to make my case that the bullet is worth biting in CB, I shall consider a different case in which ignoring PAPP observations is rationally motivated.

3.2 Making Reliable Assurances

In this section, I give an example in which the ability to ignore one's observations by using unconditional credences is in fact a virtue. The case is as follows:[93]

Parfit's Hitchhiker (PH) *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.*

Neither the CDT nor the EDT agent will be able to convincingly make this promise. Having reached her home, the agent will condition on her observation and straightforwardly deduce that paying the reward will be worse for her than not paying it. Being aware of this, however, the driver will choose not to rescue the agent.

This does not hold for the PCDT agent. She will not condition on her observation that she is home, since this variable occurs in the PAPP time. This means that she will evaluate paying the money to be the better outcome, since she believes it is worse to die than it is to pay the great reward (and her refusal to condition on

[93] Parfit (1984, p. 7). For a version of this scenario as a decision-theoretic problem, see, e.g., Hintze (2014, p. 3).

her observation means she still entertains the possibility of dying in the desert, despite having been rescued). Knowing this, the driver rescues the agent.

This opportunity arises for the PCDT agent only because of her ability to evaluate policy outcomes unconditionally, even when there is evidence available upon which one may condition. Therefore, the PCDT treatment of the veil of ignorance allows for PCDT agents to make credible assurances to predictors, an ability which CDT and EDT agents lack.[94]

A CDT or EDT advocate may question why the ability to make credible assurances is taken as a desideratum of instrumental rationality. In response, I say that it is clear that making the assurance and then keeping it is the path of decisions that makes the agent's life go as well as it possibly can for her. Moreover, by acting according to the prescriptions of PCDT, the agent can make these choices without having to utilise some additional ability to make binding commitments that later force one to make a certain choice.

The agent has control over the kind of decision procedure she employs: such is the underlying motivation behind PCDT. CDT and EDT do not grant the agent this degree of control, thus denying the agent her opportunity to escape the desert in PH.[95] If making credible assurances results in a better outcome for the agent overall, the agent should have the capability to make these assurances. PCDT grants the agent this opportunity, and it does so crucially through its refusal to condition on observations made in the PAPP time. Consequently, we should treat this refusal as a feature of PCDT, rather than a bug.

[94] Refer to Gauthier (1994) for a discussion of assurances and threats in the context of decision theory. Bratman (1987) develops an account in which the *intention* to perform an act affects the rationality of performing the act, which would be applicable in PH if we take the driver to read intentions. PCDT does not require this addition to the agent's ontology.

[95] Meacham (2010) shows that Arntzenius et al. (2004)'s response to such a case, that the inability to bind oneself should not be taken as a mark against CDT, leads to undesirable consequences for decision theory as a whole.

Chapter 4

Conclusion and Future Work

I have developed policy-based causal decision theory (PCDT) as a decision theory that (i) respects the decision-relevance of predictors and simulations; (ii) correctly identifies *causal dependence* between physical entities as being decision-relevant; and (iii) is capable of executing ideal sequences of decisions. The distinctive features of PCDT are more likely to show up for an AI than for any other kind of agent, motivated as it is through examples in which access to the AI's source code provides an easy means of prediction or simulation. Yet, in principle, PCDT captures the ideal normative theory of instrumental rationality for *any* kind of agent. Moreover, it is sufficiently well-specified so as to allow for clear predictions of behaviour.

Throughout this thesis, I have assumed that the utility and credence functions of the agent remain *constant* throughout the time period considered. However, there exists a rich literature of decision problems[96] in which this assumption is relaxed, and even a decision theory - **deliberational decision theory**[97] - intended to account for the time-instability of credence functions. It remains to be seen how PCDT interacts with such scenarios, or even whether any modification is necessary to the theory in order to allow for the possibility of such deliberation.[98]

[96] Egan (2007) introduces *Psychopath Button* and *Murder Lesion* as cases intended as counterexamples to CDT. Joyce (2009, 2012) shows how CDT may be developed to account for these cases.

[97] Arntzenius (2008) develops this account. See Skyrms (1990) for an earlier development of deliberation in decision theory.

[98] Omohundro (2008, pp. 5-6) claims that, in most circumstances, an AI will in fact try to pre-

Furthermore, as alluded to in the Introduction, it is unclear whether as humans we ought to program PCDT into an AI. The goals that determine what decision theory an AI *should* employ from our perspective - the avoidance of human catastrophe for one - do not necessarily align with the goals an AI might otherwise have. Both questions are of interest to those who seek long-term human flourishing.

serve its utility function.

Appendix A

A Review of Causal Bayesian Networks

A.1 Bayesian Networks

Causal Bayesian networks have enjoyed widespread use in the field of AI since the 1980s, when they were developed as a tool for statistical modelling by Pearl, amongst others.[99] However, the tool is less well-known amongst the philosophical community, and so I briefly develop the account here insofar as it is necessary in order to specify Causal and Policy-Based Causal Decision Theory.

Definition A.1.1 (Nielsen and Jensen (2009), Definition 2.3)**.** A **Bayesian network** consists of the following:

- A set of **variables** and a set of **directed edges** between variables.
- Each variable has a finite set of mutually exclusive states.
- The variables together with the directed edges form a **directed acyclic graph** (DAG); a directed graph is **acyclic** if there is no directed path $A_1 \to ... \to A_n$ where $A_1 = A_n$.

[99] Pearl (2009) remains the most comprehensive account of modelling causation with networks. See Spirtes et al. (2000) for an alternative account.

- To each variable A with parents $B_1, ..., B_n$, a conditional probability table $P(A \mid B_1, ..., B_n)$ is attached. If A has no parents, then the table reduces to the unconditional probability table $P(A)$, referred to as the **prior probability** for A.

Rather than requiring the whole table of probabilities, the *chain rule* for Bayesian networks yields a more compact means of calculating the required probabilities.

Theorem A.1.2 (Nielsen and Jensen (2009), Theorem 2.1)**.** *Let BN be a Bayesian network over $\mathcal{A} = \{A_1, ..., A_n\}$. Then BN specifies a unique joint probability distribution $P(\mathcal{A})$ given by the product of all conditional probability tables specified in BN:*

$$P(\mathcal{A}) = \prod_{i=1}^{n} P(A_i \mid pa(A_i)), \tag{A.1.1}$$

where $pa(A_i)$ are the parents of A_i in BN, and $P(\mathcal{A})$ reflects the properties of BN.

A.1.1 Adjusting for Evidence

Bayesian networks can also capture observations that a given variable has a certain value, or that it in fact has a certain set of values. If A is a variable with n states with $P(A) = (x_1, ..., x_n)$ and e is the information that A can be only in state i or j, then $P(A, e) = (0, ..., 0, x_i, 0, ..., 0, x_j, 0, ..., 0)$.

Definition A.1.3 (Nielsen and Jensen (2009), Definition 2.4)**.** Let A be a variable with n states. A **finding** on A is an n-dimensional table of zeros and ones.

Theorem A.1.4 (Nielsen and Jensen (2009), Theorem 2.2)**.** *Let BN be a Bayesian network over the universe $\mathcal{A}$, let $\boldsymbol{e}_1, ..., \boldsymbol{e}_m$ be findings, and let e be the statement representing these findings. Then*

$$P(\mathcal{A}, e) = \prod_{A \in \mathcal{A}} P(A \mid pa(A)) \cdot \prod_{i=1}^{m} \boldsymbol{e}_i, \tag{A.1.2}$$

and for $A \in \mathcal{A}$ we have

$$P(A \mid e) = \frac{\sum_{\mathcal{A} \setminus \{A\}} P(\mathcal{A}, e)}{P(e)}. \tag{A.1.3}$$

This result holds also for *likelihood evidence* that expresses a relative distribution over certain states. For instance, if A has possible states a_1 and a_2 and we receive the evidence that a_1 is three times as likely to be the state of A than a_2, we represent such evidence with the vector $(0.75, 0.25)$.

A.1.2 d-separation

Following the rules of *d-separation*, we can decide for any pair of variables in a Bayesian network whether they are independent given the evidence entered into the network. These rules, which cover all of the ways in which evidence may be transmitted through a variable, are formulated in the following definition.

Definition A.1.5 (Nielsen and Jensen (2009), Definition 2.1)**.** Two distinct variables A and B in a Bayesian network are **d-separated** if for all (undirected) paths between A and B, there is an intermediate variable V (distinct from A and B) such that either

- the connection is serial (i.e., $A \rightarrow ... \rightarrow V \rightarrow ... \rightarrow B$) or diverging (i.e., $A \leftarrow ... \leftarrow V \rightarrow ... \rightarrow B$) and the value of V is known; or
- the connection is converging (i.e., $A \rightarrow ... \rightarrow V \leftarrow ... \leftarrow B$), and neither V nor any of V's descendants have received evidence.

The link between d-separation and conditional independence is given in the following theorem.

Theorem A.1.6 (Pearl (2009), Theorem 1.2.4)**.** *If sets X and Y are d-separated by Z in a DAG G, then X is independent of Y conditional on Z in every Bayesian network with graph G. Conversely, if X and Y are not d-separated by Z in a DAG G, then X and Y are dependent conditional on Z in at least one Bayesian network with graph G.*

A.2 Learning Bayesian Networks

We say that the **skeleton** of a Bayesian network N is the undirected graph obtained by removing directions from all edges in N. Assume we have a set of variables, and access to queries of the form $I(A, B, \mathcal{X})$, which denote that A is d-separated from B given $\mathcal{X}$ (or equivalently, that A is independent from B given $\mathcal{X}$). We wish to determine the causal structure behind these variables. We can learn the skeleton of N by making the link $A - B$ part of the skeleton just in case $\neg I(A, B, \mathcal{X})$ for all $\mathcal{X}$ not containing A or B.

Suppose now we wish to recover the directions of these edges. We can do so by following only four rules:[100]

Introduction of v-structures If you have three nodes, A, B, C, such that $A - C$ and $B - C$, but not $A - B$, then introduce the v-structure $A \to C \leftarrow B$ if there exists an $\mathcal{X}$ (possibly empty) such that $I(A, B, \mathcal{X})$ and $C \notin \mathcal{X}$.

Avoid new v-structures When the first rule has been exhausted, and you have $A \to C - B$ (and no link between A and B), then direct $C \to B$.

Avoid cycles If $A \to B$ introduces a directed cycle in the graph, then do $A \leftarrow B$.

Choose randomly If none of the above rules can be applied anywhere in the graph, choose an undirected link and give it an arbitrary direction.

A.3 Causality

Evidently, the last of these rules might appear too arbitrary for a Bayesian network to truly represent the causal structure of a decision problem. Indeed, the *causal* graphs used in CDT and PCDT will obey further constraints:

Definition A.3.1 (Pearl (2009), Definition 1.3.1)**.** Let $P(\nu)$ be a probability distribution on a set V of variables, and let $P_x(\nu) = P(\nu \mid \texttt{do}(X = x))$ denote the

[100] Nielsen and Jensen (2009, Section 7.1).

distribution resulting from the intervention $\mathtt{do}(X = x)$ that sets a subset X of variables to constants x. Denote by $\boldsymbol{P}_*$ the set of all interventional distributions $P_x(\nu)$ such that $X \subseteq V$. A directed acyclic graph G is said to be a **causal Bayesian network** compatible with $\boldsymbol{P}_*$ just in case the following three conditions hold for every $P_x \in \boldsymbol{P}_*$:

1. P_x and G form a Bayesian network;

2. $P_x(v_i) = 1$ for all $V_i \in X$ whenever v_i is consistent with $X = x$;

3. If pa_i is a value of $pa(V_i)$, then $P_x(v_i \mid pa_i) = P(v_i \mid pa_i)$ for all $V_i \notin X$ whenever pa_i is consistent with $X = x$.

We assume that any decision problem can be accurately represented thus. Once we do, we are able to calculate the effect on an intervention on a set of variables, as the following theorem shows.

Theorem A.3.2 (Pearl (2009), Theorem 3.2.2). *Let Y be any set of variables disjoint of $\{X_i\} \cup pa(X_i)$. The effect of the intervention $\mathtt{do}(X_i = x_i')$ on Y is given by*

$$P(y \mid \mathtt{do}(X_i = x_i')) = \sum_{pa_i} P(y \mid x_i', pa_i) \cdot P(pa_i). \tag{A.3.1}$$

Equation (A.3.1) calls for conditioning $P(y \mid x_i')$ on the parents of X_i and then averaging the result, weighted by the prior probability of $pa(X_i) = pa_i$. The operation defined by this conditioning and averaging is known as 'adjusting for $pa(X_i)$.'

Bibliography

A. Ahmed. *Evidence, Decision and Causality*. Cambridge University Press, 2014.

F. Arntzenius. No regrets, or: Edith Piaf revamps decision theory. *Erkenntnis*, 68 (2):277–297, 2008.

F. Arntzenius, A. Elga, and J. Hawthorne. Bayesianism, infinite decisions, and binding. *Mind*, 113(450):251–283, 2004.

T. Benson-Tilsen. UDT with Known Search Order. Technical Report 2014–4, Machine Intelligence Research Institute, 2014. URL `https://intelligence.org/files/UDTSearchOrder.pdf`.

S. Bernstein. Omission impossible. *Philosophical Studies*, 173(10):2575–2589, 2016.

J. C. Bjerring. On counterpossibles. *Philosophical Studies*, (2):1–27, 2013.

M. Bratman. *Intention, Plans, and Practical Reason*. Center for the Study of Language and Information, 1987.

R. Briggs. Normative theories of rational choice: Expected utility. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2017 edition, 2017.

D. J. Chalmers. On implementing a computation. *Minds and Machines*, 4(4):391–402, 1994.

D. Cohen. On what cannot be. In J. Dunn and A. Gupta, editors, *Truth or Consequences*, pages 123–132. Kluwer Academic Publishers, 1990.

J. Collins, N. Hall, and L. Paul. *Causation and Counterfactuals*. MIT Press, 2004.

W. Dai. *Towards a New Decision Theory*. Less Wrong (blog), 2009. http://lesswrong.com/lw/15m/towards_a_new_decision_theory/ [Accessed: 7 March 2018].

G. L. Drescher. *Good and Real: Demystifying Paradoxes from Physics to Ethics*. MIT Press, 2006.

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

E. Eells. Metatickles and the dynamics of deliberation. *Theory and Decision*, 17(1): 71–95, 1984.

E. Eells. *Probabilistic Causality*. Cambridge University Press, 1991.

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

S. Garrabrant, T. Benson-Tilsen, A. Critch, N. Soares, and J. Taylor. A formal approach to the problem of logical non-omniscience. In *Proceedings Sixteenth Conference on Theoretical Aspects of Rationality and Knowledge, TARK 2017, Liverpool, UK, 24-26 July 2017.*, pages 221–235, 2017. doi: 10.4204/EPTCS.251.16. URL https://doi.org/10.4204/EPTCS.251.16.

D. Gauthier. *Morals by Agreement*. Oxford University Press on Demand, 1986.

D. Gauthier. In the Neighbourhood of the Newcomb-Predictor (Reflections on Rationality). In *Proceedings of the Aristotelian Society*, volume 89, pages 179–194. JSTOR, 1988.

D. Gauthier. Assure and threaten. *Ethics*, 104(4):690–721, 1994.

A. Gibbard and W. Harper. Counterfactuals and Two Kinds of Expected Utility. In A. Hooker, J. J. Leach, and E. F. McClennen, editors, *Foundations and Applications of Decision Theory*, pages 125–162. D. Reidel, 1978.

P. Greene. *Rationality and Success*. PhD thesis, Graduate School, New Brunswick, Rutgers, The State University of New Jersey, 2013.

P. Greene. Success-first decision theories. In A. Ahmed, editor, *Newcomb's Problem*. Cambridge University Press, forthcoming.

S. Hargreaves Heap, M. Hollis, B. Lyons, S. Robert, and A. Weale. *The Theory of Choice: A Critical Guide*. Oxford: Basil Blackwell, 1992.

D. Hintze. Problem class dominance in predictive dilemmas. Honors Thesis, Arizona State University, 2014.

R. C. Jeffrey. *The Logic of Decision*. University of Chicago Press, 1990.

J. M. Joyce. *The Foundations of Causal Decision Theory*. Cambridge University Press, 1999.

J. M. Joyce. Ratifiability, stability and the role of act probabilities in decision theory. In *Progic 2009: 4th Workshop on Combining Probability and Logic*. University of Groningen, The Netherlands, 2009.

J. M. Joyce. Causal reasoning and backtracking. *Philosophical Studies*, 147(1):139, 2010.

J. M. Joyce. Regret and instability in causal decision theory. *Synthese*, 187(1): 123–145, 2012.

N. Kolodny and J. Brunero. Instrumental rationality. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edition, 2016.

D. Lewis. Prisoners' dilemma is a Newcomb problem. *Philosophy & Public Affairs*, pages 235–240, 1979.

D. Lewis. 'Why ain'cha rich?'. *Noûs*, pages 377–380, 1981a.

D. Lewis. Causal decision theory. *Australasian Journal of Philosophy*, 59(1):5–30, 1981b.

E. F. McClennen. *Rationality and Dynamic Choice: Foundational Explorations*. Cambridge University Press, 1990.

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

A. Morton. Epistemic virtues, metavirtues, and computational complexity. *Noûs*, 38(3):481–502, 2004.

V. Nesov. *Counterfactual Mugging*. Less Wrong (blog), 2009. `http://lesswrong.com/lw/3l/counterfactual_mugging/` [Accessed: 7 March 2018].

T. D. Nielsen and F. V. Jensen. *Bayesian Networks and Decision Graphs*. Springer Science & Business Media, 2009.

R. Nozick. Newcomb's problem and two principles of choice. In *Essays in Honor of Carl G. Hempel*, pages 114–146. Springer, 1969.

S. M. Omohundro. The Basic AI Drives. In *AGI*, volume 171, pages 483–492, 2008.

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

J. Pearl. Causation, action, and counterfactuals. In *Proceedings of the 6th Conference on Theoretical Aspects of Rationality and Knowledge*, pages 51–73. Morgan Kaufmann Publishers Inc., 1996.

J. Pearl. *Causality*. Cambridge University Press, 2009.

J. Pearl. Physical and metaphysical imaging. Technical Report R-359, University of California, Los Angeles, 2010. URL `http://web.cs.ucla.edu/~kaoru/r359-ss.pdf`.

G. Piccinini. *Physical Computation: A Mechanistic Account*. Oxford University Press, 2015.

H. Reichenbach. *The Direction of Time*, volume 65. University of California Press, 1991.

L. J. Savage. *The Foundations of Statistics*. Courier Corporation, 1972.

B. Skyrms. *Causal Necessity: A Pragmatic Investigation of the Necessity of Laws*. Yale University Press, 1980.

B. Skyrms. *The Dynamics of Rational Deliberation*. Harvard University Press, 1990.

N. Soares and B. Fallenstein. Toward idealized decision theory. *CoRR*, abs/1507.01986, 2015. URL `http://arxiv.org/abs/1507.01986`.

N. Soares and B. A. Levinstein. Cheating death in damascus, 2017. URL `http://intelligence.org/files/DeathInDamascus.pdf`.

E. Sosa and M. Tooley. *Causation*. Oxford University Press, 1993.

P. Spirtes, C. Glymour, S. N., and Richard. *Causation, Prediction, and Search*. MIT Press: Cambridge, 2000.

W. Spohn. Reversing 30 years of discussion: why causal decision theorists should one-box. *Synthese*, 187(1):95–122, 2012.

R. C. Stalnaker. Letter to David Lewis. In *Ifs*, pages 151–152. Springer, 1980.

P. Suppes. *A Probabilistic Theory of Causality*. Amsterdam: North-Holland Pub. Co., 1970.

J. R. R. Tolkien. *The Fellowship of the Ring: Being the First Part of the Lord of the Rings*. London: Allen and Unwin, 1954.

R. Wedgwood. Gandalf's solution to the Newcomb problem. *Synthese*, pages 1–33, 2013.

P. Weirich. *Realistic Decision Theory: Rules for Nonideal Agents in Nonideal Circumstances*. OUP USA, 2004.

P. Weirich. Causal decision theory. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edition, 2016.

E. Yudkowsky. Timeless decision theory. The Singularity Institute, San Francisco, 2010. URL `http://intelligence.org/files/TDT.pdf`.

E. Yudkowsky and N. Soares. Functional decision theory: A new theory of instrumental rationality. *CoRR*, abs/1710.05060, 2017. URL `http://arxiv.org/abs/1710.05060`.