A typology of Newcomblike problems

Johannes Treutlein and Caspar Oesterheld

Abstract

This paper introduces a typology of Newcomblike problems—decision problems in which evidential decision theory and causal decision theory come apart. Such problems involve an evidential dependence between an action and an outcome that does not stem from the causal effects of the action. We distinguish two types of Newcomblike problems on the basis of how this dependence arises: (i) *reference class*, and (ii) *common cause*. In the former, the dependence arises from an inference about the actions of similar agents or models of agents, as in Newcomb's problem. In the latter, it is the result of a common cause between action and outcome, as in the Smoking Lesion. We argue that several plausible, recently proposed variants of causal decision theory take the distinction into account in their action recommendations. Moreover, the distinction sheds new light on the debate about the tickle defense. We argue that if the tickle defense works, then it applies only to common cause problems, not to reference class problems.

1 Introduction

The two most prominent theories of rational choice are evidential decision theory (EDT) (Jeffrey, 1965; Horgan, 1981; Eells, 1982; Price, 1986; Ahmed, 2014b), and causal decision theory (CDT) (Nozick, 1969; Gibbard and Harper, 1978; Lewis, 1981; Skyrms, 1982; Joyce, 1999; Pearl, 2009). Decision problems in which the recommendations come apart are often called Newcomblike problems (see, e.g., Sobel, 1990),[1] after a decision problem which was first published by Nozick (1969). These problems provide fertile soil for philosophical dispute. One reason for the persistent discussion seems to be that in addition to widely diverging intuitions between individuals, many scholars feel there are intuitively convincing cases against both EDT and CDT (MacAskill, 2016, p. 425; see Egan, 2007, p. 113; Briggs, 2010, p. 1). In hope of helping to resolve this dilemma, we would like to propose an illuminating new consideration. Newcomblike problems involve a dependence between an action and an outcome that does not stem from the causal effects of the action. We argue that two types of Newcomblike problems can be distinguished, based on how this dependence arises:

(i) from an inference about the decisions of similar agents or models of agents (e.g., a psychologically similar opponent in the prisoner's dilemma, or the Predictor's model of

[1] These problems are sometimes also called Newcomb situations, or Newcomb problems (Ledwig, 2000, p. 66). Some problems where EDT and CDT agree are also sometimes considered Newcomblike. For instance, the toxin puzzle (Kavka, 1983, pp. 33–4), Parfit's Hitchhiker (Parfit, 1986, p. 7; Barnes, 1997, pp. 2–3), or a version of Newcomb's problem with transparent boxes (Gibbard and Harper, 1978, pp. 153–4). These problems, involving the issue of *binding* (cf. Weirich, 1985; Arntzenius, Elga, and Hawthorne, 2004; Meacham, 2010), where both EDT and CDT give the same recommendations, are outside of the scope of this paper.

an agent in Newcomb's problem), based on observing the agent's action. We say that these agents or models are in the agent's *reference class*.

(ii) from a common cause between action and outcome (e.g., a genetic disposition, as in the Smoking Lesion). Action and outcome are hence dependent, but conditionally independent given this common cause.

Here, problems involving (i) are referred to as *reference class* problems, while those involving (ii) are referred to as *common cause* problems.[2] Some dependences may also arise from both (i) and (ii) simultaneously—i.e., there can be a common cause between both an agent's action and the action of a member of their reference class. In this case, the associated problem belongs to the reference class type.[3] As becomes apparent through this brief characterization, this typology is qualitative rather than quantitative. Rather than classifying problems based merely on their formal properties, we ask how an agent would realistically obtain the credence function that the decision problems supposes.

In Section 2, we briefly revisit EDT and CDT and introduce relevant notations.[4] In Section 3, we examine the two problem types more closely and identify their properties. In Sections 4 and 5, we argue that our typology is a fruitful contribution to the philosophical debate. First, we examine several recently proposed variants of CDT that take our distinction into account—following EDT in reference class, and CDT in common cause problems—thus avoiding some of the alleged pitfalls associated with EDT and CDT. Secondly, we discuss our distinction's implications for the debate surrounding the tickle defense. The tickle defense has been advanced as an argument that some or all Newcomblike problems are incoherent (see, e.g., Eells, 1982, chap. 7). We argue that if the tickle defense works, it can do so only in common cause problems, not in reference class problems. Hence, if no other argument against the tickle defense succeeds, then only reference class problems remain as Newcomblike problems, and EDT arrives at the same recommendations as the abovementioned variants of CDT. This makes EDT more plausible. In Section 6, we examine prior discussions of differences between Newcomblike problems. For instance, Sobel (1994, pp. 33–35) and Hurley (1991) have outlined distinctions similar to the one introduced in this paper. Our typology adds to this previous work in that it provides a novel account of these differences and outlines potential implications.

2 Evidential and causal decision theory

Newcomb's problem illustrates the differences between EDT and CDT:

> *Newcomb's problem:* There are two boxes; one is opaque, and one translucent. The translucent box contains \$1,000. Alice must choose between (a_1) taking only the opaque box ("one-boxing"), or (a_2) taking both boxes ("two-boxing").

[2] We partly adapt the terminology of Almond (2010, sec. 4).

[3] We elaborate on the reasoning behind this in Section 3.1. Here, we assume that both (i) and (ii) give rise to the same dependence, such that knowledge of the common cause and its causal effects screens of (i), and knowledge of the action of the reference class member screens off (ii). For the sake of simplicity, we don't look at problems in which the dependence between action and outcome arises via some more complicated mechanism.

[4] For a more thorough introduction to the decision theory of Newcomblike problems, see, e.g., Ahmed (2014b, chap. 2), or Weirich (2016).

> To determine the contents of the opaque box, the Predictor—a powerful but benevolent being—predicted Alice's choice in advance. If (s_1) she predicted Alice would one-box, the Predictor placed \$1M in the opaque box. If (s_2) she predicted Alice would two-box, the Predictor left the box opaque box empty. Should Alice one-box or two-box?

The situation can be summarized using the following *desirability matrix* (Figure 1):

	s_1	s_2
a_1	1M	0
a_2	1M+1K	1K

Figure 1: Here, columns denote possible world states, and rows denote available actions. Both states and actions are propositions—i.e., subsets of a set of possible worlds Ω—and the set of states S and the set of actions A both form a partition of Ω (e.g., Lewis, 1981). Each conjunction $s \wedge a$ of a state and an action (defined as the intersection of the corresponding sets) is a possible outcome. The table specifies the "news value", or desirability $V(s \wedge a)$ of each outcome (under the simplifying assumption that money is linear in value) (see, e.g., Jeffrey, 2004, chap. 6; Ahmed, 2014b, chap. 2.1; cf. Lewis, 1981, p. 5).

According to EDT, Alice ought to one-box. EDT recommends choosing the most auspicious action; that is, the action with the highest "news value" (e.g., Ahmed, 2014b, chap. 2.4). Given any partition S of Ω, the news value or desirability $V(a)$ of an action a can be calculated as

$$V(a) = \sum_{s \in S} P(s|a)\ V(s \wedge a), \tag{2.1}$$

where $P(s|a) = \frac{P(s \wedge a)}{P(a)} \in [0, 1]$ denotes the probability of state s conditional on taking the action a (e.g., Jeffrey, 2004, chap. 6).

In Newcomb's problem, Alice's action provides strong evidence about the prediction of her action. Conditional on action a_1, state s_1 is much more likely than s_2. Suppose the predictor's reliability is 90%, i.e., $P(s_1|a_1) = P(s_2|a_2) = 0.9$. Alice can then calculate both actions' news values as

$$V(a_1) = 0.9 \cdot 1,000,000 + 0.1 \cdot 0 = 900,000 \tag{2.2}$$
$$V(a_2) = 0.1 \cdot 1,001,000 + 0.9 \cdot 1,000 = 101,000. \tag{2.3}$$

This is unsurprising: one-boxers are much more likely than two-boxers to be rich, so one-boxing produces the better news.

CDT, on the other hand, recommends two-boxing, as this is the most efficacious action. CDT rests on the idea that what matters are only the causal effects of actions. It recommends maximizing the action's U-score (e.g., Lewis, 1981, sec. 9):

$$U(a) = \sum_{s \in S} P(a \rightarrow s)\ V(s \wedge a),^{5} \tag{2.4}$$

where $P(a \rightarrow s)$ are unconditioned probabilities of "causal counterfactuals" (see Gibbard and Harper, 1978, p. 125). These probabilities differ from conditional probabilities in that they are only sensitive to the causal effects of the antecedent on the probability of the consequent. In (2.4), it is assumed that the partition S is chosen such that each conjunction of a state s and an action a specifies all the agent cares about—i.e., for all propositions x such that $P(s \wedge a \wedge x) > 0$, it is $V(s \wedge a \wedge x) = V(s \wedge a)$. If this is the case, then the probabilities of causal counterfactuals express all the necessary information regarding the causal relationships in a given decision problem (Lewis, 1981, secs. 6, 9; see also Gibbard and Harper, 1978, p. 130).

As Alice's actions do not causally affect the states in Newcomb's problem (specified by Figure 1), it follows that $P(a_1 \rightarrow s_1) = P(a_2 \rightarrow s_1) = P(s_1)$ and $P(a_1 \rightarrow s_2) = P(a_2 \rightarrow s_2) = P(s_2)$. Moreover, given any action, the states specify everything Alice cares about. Suppose both s_1 and s_2 are equally likely *a priori*, i.e., $P(s_1) = P(s_2) = 0.5$.[6] Then:

$$U(a_1) = 0.5 \cdot 1,000,000 + 0.5 \cdot 0 = 500,000 \tag{2.5}$$

$$U(a_2) = 0.5 \cdot 1,001,000 + 0.5 \cdot 1,000 = 501,000. \tag{2.6}$$

Since Alice cannot causally influence the content of the boxes, two-boxing always causally leads to the better outcome.

In this paper, we often make qualitative assumptions about the causal structure of a decision problem, based on that problem's description. These assumptions are depicted using causal graphs (see Spirtes, Glymour, and Scheines, 2000, chap. 3.2.4), in which vertices or nodes represent partitions of Ω, while directed edges represent dependences between nodes that are based on causal relationships. So, for instance, a causal arrow from a node representing the set of states to a node representing the set of actions means that the world state causally affects the choice of action.

3 Two types of Newcomblike problems

As outlined in the introduction, our typology attempts to partition Newcomblike problems into two types: (i) *reference class* and (ii) *common cause*, based on how the dependence that sets EDT and CDT at odds arises.[7]

[5] There exist several other specifications of CDT (ibid.). The differences between them are not relevant for the purpose of this paper.

[6] Since there is a causally dominant action in this problem (see, e.g., Joyce, 1999, p. 151), CDT's verdict does not change based on different prior credences in states.

[7] Whether or not the typology is exhaustive depends on the question of whether any Newcomblike problems exist in which there are a state and an action such that $P(a \rightarrow s) \neq P(s|a)$, but there is neither a common cause that explains this dependence, nor is the dependence a reference class dependence (i.e., between similar agents or models thereof). If the common cause principle (see Section 3.1) applies to all cases that are not of the reference class type, then our typology is exhaustive. There may be counterexamples, however, depending on the correct theory of causality and the definition of the common cause principle. For instance, it is possible that quantum mechanical cases (see Ahmed, 2014a), or those involving anthropic probabilities or time travel, may lie outside our typology.

Before further elaboration on this classification, it is helpful to consider another Newcomblike problem, the Smoking Lesion. This problem is one of many medical Newcomb problems, a class based on medical conditions—e.g., genetic dispositions to certain diseases—which are said to influence human behavior. Most medical problems we are aware of are common cause problems (see Section 3.2).

> *The Smoking Lesion*:[8] "Susan is debating whether or not to smoke (a_2 and a_1, respectively). She believes that smoking is strongly correlated with lung cancer, but only because there is a common cause—a condition that tends to cause both smoking and cancer. Once we fix the presence (s_2) or absence (s_1) of this condition, there is no additional correlation between smoking and cancer. Susan prefers smoking without cancer ($a_2 \wedge s_1$) to not smoking without cancer ($a_1 \wedge s_1$), and she prefers smoking with cancer ($a_2 \wedge s_2$) to not smoking with cancer ($a_1 \wedge s_2$). Should Susan smoke?" (Egan, 2007, p. 94, with trivial variations)

Here, EDT advises Susan to abstain from smoking, because smokers are more likely to develop cancer than non-smokers. CDT recommends smoking, because in the Smoking Lesion, smoking cannot causally influence whether Susan does or does not have cancer, and causally leads to the better outcome.

Both the agent's desirability matrix and their subjective (conditional) credences in the Smoking Lesion are similar to those in Newcomb's problem. Although these problems are both Newcomblike, we maintain there is an important difference between them. To determine what constitutes this difference, we will turn our attention to reference class problems and try to develop a theory for the dependences involved.

3.1 Reference class problems

(a) (b)

Figure 2: Causal graphs of generic (a) reference class and (b) common cause problems. Both problems involve a dependence between action (A) and outcome (O) that does not stem from the causal effects of the action. In reference class problems, the dependence arises from an inference about the action of a member of the agent's reference class (R). Since this dependence is not necessarily rooted in a causal relationship, we indicate it using a dotted, double-arrowed arc. In common cause problems, the dependence stems from a common cause (C) between action and outcome.

[8] To our knowledge, this problem was first introduced as "Fisher's problem" by Jeffrey (1965, p. 15).

The dependence in reference class problems stems from a similarity between an agent's own decision-making process and those of other agents, or models of agents (Figure 2). For example, the Predictor in Newcomb's problem might base her prediction on a model of the agent's reasoning process: in this case, it is possible for the agent to infer the action that the model predicted by observing their own action. Another example would be a prisoner's dilemma with a psychologically similar opponent, which is also often regarded as a Newcomblike problem (Brams, 1975; Gibbard and Harper, 1978, sec. 12; Lewis, 1979; Horgan, 1981, sec. 10; Sobel, 1994, chap. 3; Ahmed, 2014b, chap. 4.6). More generally, in any game with players that use similar decision procedures and are in a symmetric decision situation, the players' decisions are dependent: $P(Y = a|X = a) > P(Y = a)$, where $X = a$ and $Y = a$ denote the proposition that player X respectively Y chooses action a. We refer to decision problems fulfilling these criteria as *reference class* problems, since they involve making inferences about the actions of members of a reference class of similar agents, or models thereof. A list of reference class problems includes:

- Newcomb's problem (Nozick, 1969)
- Meta-Newcomb problem (Bostrom, 2001)
- Death in Damascus (Gibbard and Harper, 1978, pp. 157–9)
- Prisoner's dilemma (Gibbard and Harper, 1978, sec. 12; Lewis, 1979; cf. Rapoport, 1966, pp. 139–41)
- Dancing chief (Dummett, 1964, pp. 348–9)
- Superpower arms race (Brams, 2014, chap. 7.3)
- Voting in a mass election (Grafstein, 1991, 2002; Ahmed, 2014b, pp. 117–8)
- Nomination race (Brams, 1976, chap. 8.3)
- Platonia dilemma (Hofstadter, 1996, pp. 737–55)

In some reference class problems, a common cause exists between an agent's own action and that of a member of their reference class. In Newcomb's problem, for instance, the Predictor might base her model of Alice on a past state of Alice's brain. This past state then determines both Alice's decision when faced with the problem and the outcome of the prediction (Figure 3).[9] Newcomb's problem is therefore also categorized as a "common cause problem" by other authors (Eells, 1982, pp. 210–1; Hurley, 1991; Burgess, 2004, sec. 3, 2012, p. 323). Conversely, we assign all of these problems to the reference class type, regardless of the existence of an additional common cause; as long as they simultaneously involve a reference class inference, they share the relevant properties with all other reference class problems. That is, as we will argue in Sections 4 and 5, several alternative versions of CDT "one-box" in these problems, and the tickle defense does not apply in these cases.

Importantly, reference class problems without any common causes also exist. As Burgess (2012, p. 323) points out, "there is no such common cause to be recognized in relation to the prisoners' dilemma"—we could imagine a prisoner's dilemma in which there is no causal interaction, and therefore no common cause, between an agent and their opponent. Still, if both agents use similar decision-making processes, then one agent's decision will be evidence

[9]To the degree that Alice's decision procedure manifests itself in the (past) physical structures of Alice's brain, the common cause in Newcomb's problem might again be considered a member of her reference class.

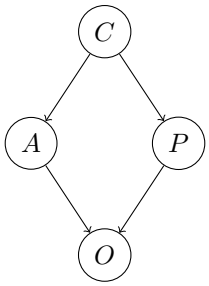

Figure 3: A possible causal graph of Newcomb's problem. Alice's past brain state (C) determines both her action (A) and the prediction (P).

about the other agent's decision. Similar considerations apply to the *class action* approach to decision making in an infinite universe, in which inferences are made from one agent's actions about the actions of identical agents in causally disconnected parts of the universe (Bostrom, 2011, pp. 38–40). Another good example of a reference class problem without any common causes is Oesterheld's *Multiverse-wide Cooperation via Correlated Decision Making* (2017), which involves correlations between similar, but not necessarily identical, causally disconnected agents.

Cases in which dependences exist between agent's actions absent of any causal relations violate Reichenbach's (1956) "principle of the common cause" (see Yudkowsky, 2010, sec. 10; cf. Spohn, 2012, p. 101).[10] In Reichenbach's (1956) original formulation[11], this principle states that if any two events a and b are dependent, then either one is a cause of the other, or there is a third event c which is a common cause of a and b and satisfies the following conditions:

(3.1) $0 < P(c) < 1$

(3.2) $0 < P(c) < 1$

(3.3) $P(a \wedge b|\neg c) = P(a|\neg c)P(b|\neg c)$

(3.4) $P(a|c) > P(a|\neg c)$

(3.5) $P(b|c) > P(b|\neg c)$.

For example, the actions of two agents in a prisoner's dilemma can be interpreted as events in Reichenbach's sense. If a dependence exists between these events, but there is neither causal interaction between the two agents nor a third event that causes both actions, then this dependence appears to violate the common cause principle.

We believe that the reason dependences in reference class problems can violate the common cause principle is that (i) an agent's action and those of members of their reference class are *logically* related, and (ii) if two events are logically related, and hence not "distinct" (see

[10] Depending on how the common cause principle is interpreted, this could be disputed, e.g., by claiming that some property of the initial state of the universe might count as a common cause. If this is so, then there is indeed a common cause to be found in these problems.

[11] This formulation is flawed in that it does not account for situations with more than one common cause between two events (Arntzenius, 2010, sec. 1.1). In Section 4, we introduce a related principle without this restriction, the causal Markov condition.

Lewis, 1986, sec. 4), then the common cause principle can fail (Hitchcock, 2016, sec. 2.3, para. ii). It is easy to see why (ii) is the case. For example, suppose e, f, and g are three completely independent events, i.e., $P(e \wedge f \wedge g) = P(e) \cdot P(f) \cdot P(g)$. We define two events a and b such that $a = e \wedge f$ and $b = e \wedge g$. Then a and b are dependent, although there may be no common cause between them.[12] The remainder of this section focuses on (i), following arguments by Yudkowsky (2010, sec. 11).

For humans and all other agents, the physical processes of deliberation between different choices could be abstractly described by a *decision algorithm*[13] performing a *decision computation*. The algorithm implemented by a human brain appears to be quite complex, but a decision algorithm could also take a simpler form. For example, a procedure that calculates the values that EDT assigns to each action and then selects the action with the highest news value could be considered a decision algorithm. When deliberating, the properties of the decision algorithm in combination with a particular decision situation determine the agent's actions and simultaneously, those of all other agents in symmetrical decision situations using an identical or functionally equivalent decision algorithm. This means that if an agent's action is predicted based on a model of their decision algorithm—and if that prediction is likely to be accurate—the outcome of the prediction is caused by an algorithm that is in logical relation to the algorithm the agent executes when facing the decision themself. Assuming perfect equivalence of both algorithms, given the decision to *one-box*, it is logically impossible for the algorithm examined by the Predictor to output *two-box*. Or, as Horgan (1981, p. 343) explains, if one is "completely certain that the being has predicted correctly, then the state of [the opaque box] depends logically upon [one's] current beliefs together with [one's action]."

To illustrate, consider two pocket calculators. Both calculators always give the same output if queried with the same algebraic expression, even if there is no event that could count as a common cause for both calculators' outputs. This is because both pocket calculators' outputs depend on the same logical fact. After observing physical evidence regarding the output of one of the calculators, one updates one's credence regarding the logical fact, and consequently, one's credence regarding the output of the other calculator. The credences regarding the outputs of these two calculators are thus logically dependent, even if there is no causal connection between the two.

We believe that the dependences in reference class problems work analogously. Dependences without a common cause exist because the events representing actions of members of the same reference class are not "distinct"—if two agents implement a similar decision procedure, they arrive at the same decision given the same decision situation. Theoretically, it might even be possible to specify exactly how decisions are being computed in a human brain, and subsequently to introduce a formal representation of this algorithm. In Newcomb's problem, once observing their own action, a person could update their credence regarding the logical fact about this algorithm, and consequently, their estimation of the actions of all members of their reference class, including the Predictor's model.

To summarize, the dependences in reference class problems can be interpreted as logical dependences between different instances of the same or similar decision algorithms. In that

[12] According to Lewis (1986), for instance, neither a nor b stand in a *causal* relation to e. Given Lewis' theory of causality, there may be no common cause between a and b.

[13] Here, the term algorithm is used in its broadest sense. We do not contend that the human brain necessarily implements a Turing-computable algorithm.

sense, these problems differ relevantly from common cause problems. Before discussing the potential implications of this distinction, we briefly turn our attention to common cause problems.

3.2 Common cause problems

The dependence in common cause problems arises not from an inference about a similar agent or model thereof, but from a common cause between actions and outcomes. Observing their action provides an agent with evidence about the causes of their action. Therefore, if there is a common cause between action and outcome, the agent can make an inference from their action to the outcome via the common cause that is not based on the action's causal effect.

Most common cause problems in the literature involve medical conditions, but a common cause may influence agents' actions in other ways as well: for example, via subliminal advertising, as in Sobel's (1986) Popcorn problem. The following problems may be classified as common cause:[14]

- Smoking lesion (Fisher's problem) (Jeffrey, 1965, p. 15; Stalnaker, 1980)
- Expand or contract (Muth, 1961; Frydman, O'Driscoll, and Schotter, 1982; Broome, 1989)
- Eggs-benedict-for-breakfast-problem (Skyrms, 1980, pp. 128–9)
- Loafing-weak-heart-problem (Lewis, 1981, pp. 9–11)
- Uncle Albert's problem (Skyrms, 1982, p. 700)
- Popcorn problem (Sobel, 1986, pp. 411–3)
- Pregnant women's problem (Price, 1986, p. 196)
- Nasty medicine (Mellor, 1987, pp. 170–1)
- Coco's problem (Price, 1991, p. 7)
- CGTA (chewing-gum throat-abscess) problem (Yudkowsky, 2010, p. 2)
- Newcomb's soda (ibid., pp. 11–12)
- Toxoplasmosis problem (Altair, 2013, p. 4)
- Weiner's robot problem (Weiner, 2004)
- 2-possible-fathers-1-son-case (Nozick, 1969, p. 125)
- Betting on the past (Ahmed, 2014b, chap. 5.1)
- Betting on the laws (ibid., chap. 5.2)
- Calvinist's problem (Resnik, 1987, p. 111; Ahmed, 2014b, chap. 0.6)

[14] Most of this list is taken from Ledwig (2000, pp. 80–7). The last seven examples might also be considered reference class problems, depending on how exactly the dependences in these problems are explained. We discuss the reasons for this in Sections 5 and 6.

- Solomon's problem (Gibbard and Harper, 1978, pp. 135–7)
- Jones' problem (ibid., pp. 137–8)

There is some debate about which kinds of propositions (representing events) can count as common causes; however, this topic is beyond the scope of this paper, and we take an agnostic stance. In the context of this typology, a common cause must fulfill conditions (3.1)-(3.5), and the dependence between the common cause and the two dependent propositions must be a causal one. For instance, it is left to one's conceptualization of causality to determine which propositions have a causal effect on each other.

Most philosophers feel that EDT's recommendations in common cause problems are wrong. For this reason, common cause problems are often proposed as intuitively convincing counterexamples to EDT (Nozick, 1969, p. 135; Skyrms, 1982, p. 700; Hurley, 1991, p. 174; Egan, 2007, p. 96; Pearl, 2009, sec. 4.1.1; Ahmed, 2014b, p. 91). In at least some reference class problems, on the other hand, a sizeable minority of scholars endorse EDT's recommendations (see Bourget and Chalmers, 2014; cf. Shafir and Tversky, 1992). Nozick (1969, p. 136), who defends CDT's recommendations in all Newcomblike problems, concedes that in Newcomb's problem, there is a stronger feeling of being able to exert a non-causal influence than in medical Newcomb problems. Some reference class problems, such as "Death in Damascus", are even treated as counterexamples to CDT (see, e.g., Ahmed, 2014a).

These intuitive judgements are supported by our analysis of the two types of problems. Since their relevant properties differ, it is reasonable to treat them separately, both in the discussion of Newcomblike problems and in practical decision-making. For instance, it is not clear that a failure of EDT's recommendations in common cause problems necessarily also refutes those recommendations in reference class problems, and vice versa for CDT. Instead, one might choose to act according to EDT in reference class problems, and according to CDT in common cause problems. Sections 4 and 5 present two ways of justifying such an approach.

4 Alternative decision theories

Several more recently proposed variations of CDT take the distinction outlined in this paper into account, including Spohn's (2003, 2012) variation of CDT, Poellinger's (2013) variation of CDT, and Yudkowsky and Soares' (2017) "functional decision theory". Here, we examine how and why these theories arrive at this distinction.

Each of these three theories follows a causal approach to decision making, and they are all based on graphical causal models. Such models are powerful tools in empirical investigation and statistical modeling, and they provide a practical formalization of CDT. One important assumption behind such models is the common cause principle, in this context represented by the causal Markov condition (see Spirtes, Glymour, and Scheines, 2000, chap. 3.4.1; see also Pearl, 2009, chap. 1.3).[15] Given a joint probability distribution P over a set of random variables $V = \{X_1, \ldots, X_n\}$ and a causal graph $G := (V, E)$, the causal Markov condition is fulfilled if any variable X_i is conditionally independent of all variables other than its descendants, given the set of its parents, PA_i. That is, the value of each variable depends

[15] There are differences between Reichenbach's common cause principle and other, related principles, such as the causal Markov condition. But these differences don't matter to us here.

probabilistically on only the value of its parents; hence, there are no dependences between variables except those expressed by the edges of the causal graph, and the causal graph with the probability distribution is a causal Bayesian network. Although a full discussion of the mathematical details is not possible here, it is important to note that such a network can be used to model outside interventions on variables. An intervention differs from ordinary conditioning in that only the causal effects of a variable on other variables are taken into account when setting the variable to a specific value. CDT can be formalized by modelling actions as interventions on a variable in the causal graph.

In Section 3.1, we argued that some reference class problems violate the common cause principle. This leads to difficulties in modeling these problems via a causal Bayes network. For instance, in the prisoner's dilemma, the naive causal graph used to represent this problem would contain no causal edges between the two agents (see Figure 2, in which the non-causal dependence between both agents is indicated by a dotted, double arrowed arc). However, since there is a dependence between these agents' actions, this causal graph violates the causal Markov condition (see Yudkowsky, 2010, sec. 10).[16]

In order to model all reference class problems within this framework, it is necessary to augment the causal graphs. Assuming the interpretation of reference class dependences as dependences between logical facts about decision algorithms introduced in Section 3.1, this might be accomplished as follows. First, a way to model the agent's lack of logical omniscience—that is, to represent their uncertainty about logical facts—is needed. For instance, one might choose to include both possible and impossible worlds in the set of worlds Ω (see, e.g., Nolan, 2013; Bjerring, 2014). Secondly, causal graphs could be augmented with nodes representing logical facts about decision algorithms (Yudkowsky, 2010, sec. 11; Yudkowsky and Soares, 2017, sec. 5). These logical facts could then be introduced as common causes for the actions of agents of the same reference class.

Given such an augmented graph, decisions might be modelled not as interventions on the node representing the agent's action, but as interventions on the node representing the logical fact about the agent's decision algorithm. This approach is taken by Yudkowsky and Soares' functional decision theory. Since an intervention is still performed, common cause dependences—here, the common cause is not a descendant of the decision algorithm node—are still disregarded. However, since the actions of members of the agent's reference class are descendants of the decision algorithm node, reference class dependences are taken into account.

One reason for performing interventions on the decision node rather than the action node is that the latter would lead to logically impossible counterfactuals. If an agent models actions as interventions on the action node, this leaves the actions of all other agents from the same reference class untouched. Hence, the resulting probability distribution violates that agent's knowledge about the logical relations between the actions of agents of the same reference class. According to Yudkowksy (2010, p. 89), the correct counterfactuals or closest worlds in which the agent arrives at a particular decision would be the worlds in which not only the given agent, but all members of their reference class, assume this decision.

In a similar approach to Yudkowsky and Soares, Poellinger (2013) introduces epistemic

[16]Pearl augments his causal graphs with bidirected edges (dotted, double arrowed arcs) "to denote the existence of unobserved common causes" (Pearl, 2009, p. 12). A similar notation could be used to denote dependences that do not arise from a causal relationship.

contours, "non-directed, non-causal but rather informational link[s]" (ibid., p. 13). These links augment a simple causal model into a "causal knowledge pattern" (ibid., pp. 13–14). Poellinger explicitly mentions that the links should "work like synonyms, mathematical inter-definitions, or logical relations" (ibid., p. 13), which is consistent with what we have written about reference class dependences in Section 3.1. For Poellinger, these epistemic contours connect the node representing the agent's act[17] with the act's prediction or, for example, with the act of a psychologically similar opponent in the prisoner's dilemma. Since epistemic contours are not severed by interventions, his version of CDT also "one-boxes" in reference class problems. And since epistemic contours represent only reference class—not common cause—inferences, Poellinger's CDT behaves like standard CDT in common cause problems.

Spohn also augments his causal graphs by adding additional nodes. These nodes stand for the agent's decision situation: "an action is caused precisely by the decision situation in which it is best, where a decision situation is a subjective state of the agent consisting of desires and beliefs and represented by a truncated Bayesian net with a distinguished action variable and a utility function" (Spohn, 2005, p. 5). The graphs that include these nodes are called "reflexive decision graphs." Consistent with our argument, Spohn justifies the need for the insertion of these nodes in terms of Reichenbach's common cause principle. In Newcomb's problem, the decision situation serves as a common cause for prediction and action.

(a) (b)

Figure 4: (a) Spohn's reflective decision graph containing the decision situation (D) as a causal parent to action (A) and prediction (P), and (b) the graph that is reduced by the decision situation with an edge going from action to prediction.

Spohn justifies "one-boxing" in reference class problems in the following terms. Before deciding, the reflexive decision graph is reduced by the "decision situation" node. That is, the node is removed from the graph, and a new edge is inserted between nodes previously adjacent to the now-removed node (Figure 4). Introducing these new edges ensures that all dependences between nodes are still represented by the reduced graph. Normally, a common cause between two nodes would be reduced to a directed edge in accordance with the temporal order of the events these nodes represent—i.e., the edge is directed from the earlier to the later event. According to Spohn, an exception should be made for the decision node; it should be reduced such that, where the decision node was a common cause between

[17] Here, "act" means the passive, observable variable, while the action is the exogenous intervention on the act variable (cf. ibid., sec. 4.1.1).

action node A and some other node X, the new edge should always go from A to X. In Newcomb's problem, for example, this means that after the reduction, there will be a causal edge (backwards in time) from action to prediction. Since the prediction is now a causal child of the action node, intervening in the graph reduced by the decision node yields one-boxing.

This approach appears to result in the same conclusions arrived at by Yudkowsky and Soares, and by Poellinger. However, Spohn's path to these conclusions appears more arbitrary than the former two approaches; his reason for reducing the decision node such that edges will leave the action node instead of the node occurring earlier in time is that in the second case, the reduction would depend on the temporal structure of the decision problem. If, for instance, the prediction in Newcomb's problem was made prior to the action, this would result in two-boxing, while if in some way the agent had a chance to first act and only then the prediction would be carried out (without looking at the act), the theory would recommend one-boxing. In Spohn's words, "The temporal location of the [decision node's children] is irrelevant to the causal structure of the situation and should not make any difference" (Spohn, 2012, p. 119). Nonetheless, this does not explain why these edges should always leave the action node, instead of always entering it, for instance.

In conclusion, the concept of reference class problems serves as an effective explanation for the dependences seen in the prisoner's dilemma and Newcomb's problem, and can facilitate modelling these problems with augmented causal graphs. The fact that the three discussed theories all "one-box" in reference class problems appears to be motivated primarily by the intuition that even causalists *ought* to take reference class dependences into account. The causal graph framework and its necessary augmentation provide a convenient opportunity for doing so without having to entirely succumb to evidentialism.

5 The tickle defense

The tickle defense, an argument against the coherence of Newcomblike problems, has been advanced in many forms. A common (and, we believe, plausible) version of the argument is as follows (see Ahmed, 2014b, chap. 4.3.2; see also Weirich, 2016, sec 2.5):

(5.1) A rational agent's action can be caused by modifying (i) the agent's beliefs, (ii) the agent's desires, and (iii) the agent's decision algorithm.

(5.2) An agent employing EDT has sufficient knowledge of (i) and (ii) to make their actions evidentially irrelevant to (i) and (ii).

(5.3) Therefore, Newcomblike problems in which a common cause influences only (i) or (ii) (or both) are incoherent.

Roughly, the argument for (5.2) is that an agent employing EDT must, at minimum, have sufficient knowledge of their beliefs and desires to be able to assert which actions have highest news value. Therefore, a cause that influences the agent's beliefs or desires such that in combination, one action has higher news value than another, is immediately known to the agent by judging the action's news value. The agent's action can thus never tell the agent anything new about this cause. (5.3) is a consequence of the causal Markov condition, which implies that knowledge of an action's immediate causes makes the action evidentially irrelevant to its non-descendants—the knowledge "screens off" all dependences between

action and non-descendant nodes (Pearl, 2009, chap. 1.1.5). Hence, in a Newcomblike problem in which a common cause influences only (i) or (ii), dependences that do not stem from the causal effects of the action are screened off. EDT thus recommends the same action as CDT—the action with the superior causal effects. The problem is incoherent, because it stipulates beliefs and desires that an agent will never hold in a realistic version of the problem; in reality, the most efficacious option has highest news value.

While the tickle defense has been employed against all Newcomblike problems (Eells, 1982, chap. 7, 1984), some scholars argue that it applies to problems such as the Smoking Lesion, but not to Newcomb's problem or to the Prisoner's dilemma (Price, 1986, p. 209; Ahmed, 2014b, p. 119). In this section, we support this argument on the basis of our characterization of the two types of Newcomblike problems. A full discussion of the arguments for and against the different versions of the tickle defense is beyond the scope of this paper;[18] instead, we assume the version outlined above and show that it applies to all common cause problems. Moreover, we argue that no version of the tickle defense that is based on the causal Markov condition—i.e., that works via an agent screening off a common cause dependence, based on knowledge of their action's causes—can apply to reference class problems.

In reference class problems without a common cause, the reference class dependence violates the causal Markov condition. Thus, there is no way to screen off the dependence (except by knowing the variable of interest itself), and no version of the tickle defense that is based on the causal Markov condition can apply. Hence, only Newcomblike problems involving a common cause are of any concern; in this case, our argument is as follows:

(5.4) In reference classes problems, if there is a common cause, then either the common cause or its causal effect is unknown to the agent. Thus, the reference class dependence cannot be screened off by knowledge of the common cause.

(5.5) Common cause problems work only via (i) or (ii) (or both).

(5.6) Therefore, the tickle defense as outlined in (5.1)-(5.3) applies to all common cause problems, while no version of the tickle defense that is based on the causal Markov condition can apply to reference class problems.

Assuming that (5.1)-(5.3) is sound and that all versions of the tickle defense that are not based on the causal Markov condition are unsound, it follows that common cause problems are not Newcomblike problems, while reference class problems are. Hence, EDT gives the same recommendations as Yudkowsky and Soares' theory and others.

The truth of (5.4) may be illustrated by Newcomb's problem, in which a past brain state could serve as a common cause for both the agent's action and the outcome of the prediction of that action. In the spirit of the tickle defense, it could be claimed that if the agent knows that one-boxing has higher news value, they already know all relevant facts about their past brain state. But apparently, this line of reasoning fails. As long as the decision rule implemented by the past brain state remains unknown to the agent, the effect the news value of an action has on the decision implied by the past brain state remains unknown as well. Hence, since the decision algorithm is what makes the relevant news, the tickle defense does not apply.

However, even if an agent knew the decision algorithm implemented by the common cause—

[18] For an exposition of the debate, see, e.g., Ahmed (2014b, chap. 4.3).

and hence, their own decision algorithm—then this still would not suffice to screen off the dependence. Given (empirical) knowledge of the agent's decision algorithm, there is still (logical) uncertainty about the decision implied by this rule. This is because knowing both the decision algorithm and which decision it implies would create a paradox—it would mean the agent knowing their own decision before having decided. As long as the actual action is not carried out, a rational agent must retain the illusion of being free to change their mind. So long as that is the case, the agent does not know which decision is implied by their decision algorithm, nor which prediction their past brain state yields (cf. Eells, 1982, p. 150).

This finding can be interpreted in two ways. Either the common cause in the causal graph denotes whatever the past brain state implies, or the common cause is the past brain state. In the former case, the common cause (i.e., a logical fact about the agent's decision algorithm) is unknown, but its causal effect on its children is known. In the latter case, the common cause (i.e., the past brain state) is known, but its effect is uncertain; thus, even after conditioning on the common cause, the action still yields additional information about the prediction. If the causal effect of the common cause on its children is unknown, then the causal Markov condition is violated, and the dependence induced by this common cause cannot be screened off (see Cartwright, 1999, sec. 2.2). The latter line of reasoning is consistent with a remark by Ahmed (2014b, p. 97). As he points out, the structure of such a problem is similar to a counterexample to the causal Markov condition by Cartwright (1999, p. 7).

We believe this argument carries over to other reference class problems as well. Reference class problems are defined as problems involving an inference about the members of an agent's reference class. This means that the dependence has to be between the agent's action and the action of something that's at least faintly agent-like. No matter how indirect, a decision procedure must eventually be involved and executed, and the information available to this procedure and the posed decision problem must be sufficiently similar to the information and problem the agent faces. The agent might know all the facts regarding how their decision (and that of their reference class member) is caused; however, this still leaves open the actual decision resulting from these factors. Moreover, it is only the decision itself—as the result of a computation involving beliefs, desires, and a decision algorithm—that can screen off the reference class member's decision. The agent's logical uncertainty about their decision algorithm's output manifests itself in uncertainty about the reference class member's decision algorithm's output. There can be no tickle defense.

This argument also applies to reference class problems in which a common cause influences only (i) or (ii) (or both). For instance, there could be a common cause for the beliefs and desires of two identical agents in a prisoner's dilemma. Given knowledge of the agents' decision algorithms and their implications, knowledge of such a common cause would screen off the dependence between both agents. Moreover, since the common cause influences only (i) and (ii), it is known to both agents. But the effect of the common cause is mediated by the agents' decision algorithms and their implications, at least the latter of which are unknown to the agents even after learning about the common cause. Thus, the common cause's effect is uncertain, and (5.4) applies.

Anticipating a possible objection, we must address the possibility of so-called "screening by inclination" (Ahmed, 2014b, p. 105). That is, for example, if the prediction of an agent's action in Newcomb's problem happens by finding out about their initial inclination, the agent is able to notice their inclination during deliberation and hence know the outcome of the prediction prior to their final decision. We maintain that such a Newcomb problem is

not a reference class problem. A reference class problem requires a similar agent or model of agent, such that one agent's decision tells them something about the other agent's or model's decision. The initial inclination is not the final decision, but could be understood of as one cause of the decision. A prediction by inclination therefore has a similar cause as the agent's decision, but the inclination doesn't serve as a member of the agent's reference class. This is clear from the different information available to the algorithm arriving at the inclination, and the algorithm executed afterwards. The former can't change the decision based on knowing the inclination, while the latter can. Thus, the algorithms are in different information states. Since the agent's decision algorithm (i.e., whatever guides the computation that leads to the actual decision) and the algorithm that generates the inclination are not in the same reference class, there is no reference class problem, and our argument here does not apply.

Turning to (5.5), our argument for (5.4) does not apply to common cause problems that work via an influence on beliefs and desires. It is possible to know which beliefs and desires are associated with which disease, for instance, prior to making a decision. An agent can infer knowledge about the disease from their beliefs or desires. There is no logical uncertainty, since the common cause's effect does not depend on the output of the agent's decision algorithm.

That leaves as a potentially coherent common cause problem only one in which a common cause influences the agent's decision algorithm. For instance, there might be a genetic condition that causes evidentialism or causalism, and is also associated with a disease. Then, what is left to demonstrate is that if there exists a coherent Newcomblike problem that works via an influence on the agent's decision algorithm, then such a problem belongs to the reference class type.

The reason for this is similar to the reason for (5.4). Knowing the correlation between decision algorithm and common cause is not enough. It is necessary to know what the decision algorithm implies. Otherwise, no Newcomblike problem emerges. As an example, consider Nozick's (1969, pp. 127–8) 2-possible-fathers-one-son-case:

> *2-possible-fathers-one-son-case*: There exists a gene which determines whether a person has a tendency to follow EDT or CDT. Bob does not know whether (s_1) a man who carries the gene for EDT, or (s_2) a man who carries the gene for CDT passed his genes on to him. Now Bob is faced with a choice similar in structure to the one displayed in Figure 1. "The choice matrix might have arisen as follows. A deadly disease is going around, and there are two effective vaccines against it. (If both are given, the person dies.) For each person, the side effects of vaccine a_1 are worse than that of vaccine a_2, and each vaccine has worse side effects on persons in s_2 than either does on persons in s_1." (ibid., 128, with trivial variations). Thus, Bob has inherited an additional condition from his father, which makes him either more or less susceptible to the side effects of the vaccine. Suppose it is known that both $P(s_1|a_1)$ and $P(s_2|a_2)$ are high, while $P(s_2|a_1)$ and $P(s_1|a_2)$ are low, i.e., proponents of EDT are likely to take action a_1, while proponents of CDT are likely to take action a_2.

Given either state, Bob would cause the better outcome by taking a_2, but taking a_2 would indicate that Bob is more likely to be in the worse state. So CDT recommends a_2, while EDT recommends a_1.

This appears to be a common cause problem in which the common cause influences (iii).

However, we argue that in addition to a common cause, there is a crucial dependence based on the agent's reference class, without which there would be no Newcomblike problem. Hence, the 2-fathers problem is a reference class problem.

The problem depends on the assumption that people with the gene for evidentialism would likely choose a_1. For example, if evidentialists were likely to choose a_2, then Bob's action would provide him with no evidence about his tendency to follow either decision algorithm. To be able to state a dependence between action and cause based on causing the agent's decision algorithm, it must be known which decision algorithm causes which action. In Nozick's case, for example, this knowledge could be gathered by performing a study among people with both genes. From observing his action, Bob would then gain information about his reference class—namely, whether agents in his reference class that participated in the study chose a_1 or a_2. The fact that he can make this inference is decisive for establishing a dependence between action and gene.

The reference class nature of the problem becomes clearer when Bob's situation prior to learning about the result of the study is considered. Before Bob learns about the dependence between genes and actions, taking action a_1 makes it more likely that, whichever genetic disposition he has, the agents in his reference class did so, too. It does not change the probability of having either genetic disposition. Hence, absent a dependence between states and actions, even EDT would advise Bob to commit himself to action a_2, if given the choice before learning about the result of the study (cf. Meacham, 2010, p. 11; Ahmed and Price, 2012, pp. 22–23).[19]

While it might be possible to cause an agent to adhere to a certain decision algorithm, it is not possible to influence which algorithm leads to which outcome: the output of the algorithm given a decision situation is logically determined by the algorithm. In order to determine which decision algorithm leads to which action, then, the decision algorithm itself must be studied. A Newcomblike problem emerges because the agent can make an inference from their decision to the output of their decision algorithm. We therefore assign such a problem to the reference class type. It follows that all common cause problems work via (i) or (ii).

6 Related work

Distinctions between different kinds of Newcomblike problems have often been introduced implicitly through novel decision theories (Fisher, n.d. Spohn, 2003; Yudkowsky, 2010; Spohn, 2012; Poellinger, 2013; Yudkowsky and Soares, 2017). Apart from these, Sobel and Hurley have published writing relating to our topic, which we discuss below.

Sobel (1994, pp. 33–35) distinguishes four different cases of Newcomblike problems: (i)

[19] Analysed in this way, the problem is similar to Arntzenius' "Yankees vs. Red Sox" (Arntzenius, 2008, pp. 286–7), or the "XOR Blackmail" (Soares and Fallenstein, 2015, p. 2; Soares and Levinstein, 2017, sec. 2). In all of these problems, both an agent's choice and the world state is predicted beforehand (in Nozick's case, via a study among agents in the agent's reference class, i.e., with the same genetic condition), and then it is revealed to the agent that either (i) they will choose a_1 and they are in state s_1, or (ii) they will choose a_2 and they are in state s_2. Moreover, the agent's preference ranking is $(a_2 \wedge s_1) \succ (a_1 \wedge s_1) \succ (a_2 \wedge s_2) \succ (a_1 \wedge s_2)$. In Yankees vs. Red Sox, the mapping between preferences and states changes based on whether it is predicted that the agent wins or loses, but otherwise it is the same principle.

Predictor cases, (ii) *Causal cases*, (iii) *Similar decision processes*, and (iv) *Signs of character*. As in the typology presented here, this distinction is based on how "evidential bearings of actions on certain features are made plausible" (ibid., p. 33). Sobel's (ii) include what we have classified as common cause problems, while (i) and (iii) translate into the reference class type. In our framework, (iv) might belong to either the common cause or the reference class type, depending on the exact specification of the problem. This is because "personality" or "character" are vague terms, which could refer to several things simultaneously: for instance, the agent's goals; specific patterns in how the agent deviates from ideal rationality (cf. Tversky and Kahneman, 1974); or, of course, the agent's decision processes. Depending on which of these causally influences the outcome, the problem could belong to either the common cause or the reference class type.

Hurley (1991) introduces a distinction between common cause problems and problems involving "collective action" that is also similar to the one presented here. In the latter type of problem, all agents have similar preferences regarding outcomes—that is, they all prefer a collective cooperative action to CDT's Nash equilibrium. In Hurley's words, there is no "collective act-outcome independence"—if all agents agree to cooperate, then this produces a better outcome for everyone than one that would result from causally rational action. According to Hurley, this definition does not include Newcomb's problem, since predictor and predictee do not necessarily agree in their preferences towards outcomes—it is not the case that the predictor also favors a one-boxing predictee (cf. Sobel, 1994, chap. 4). In that, his distinction is different from ours—our reference class type, which otherwise matches the "collective action" criterion, does include Newcomb's problem. This is because, as outlined above, it is the predictor's *model* of the predictee which has to believe it is in the same decision situation as the predictee, and which hence also prefers one-boxing. Thus, in our view, a "collective causal power" (Hurley, 1991, p. 176) between model and actual agent, resulting in a mutually beneficent outcome, can be established. (If the model had relevantly different preferences from the actual agent, then the model's actions would be less predictive of the agent's actions). Contrary to Hurley (ibid., p. 174) and Sobel (1994, chap. 4), we therefore also place Newcomb's problem in the same category as the prisoner's dilemma. Nevertheless, we agree with Hurley and Sobel that common cause problems (such as those listed in Section 3.2) are not like prisoner's dilemmas: while those problems all depend on a common cause between action and outcome, the relevant dependences in prisoner's dilemmas are based on inferences about the actions of other agents, and hence indicate the reference class type.

Acknowledgements

We are grateful to Arif Ahmed, Daniel Kokotajlo, Max Daniel, Simon Knutsson, Kenny Easwaran, David Lanius, and David Althaus for helpful comments on earlier drafts of the paper. We also thank R. Healy for proofreading and Alfredo Parra for typesetting.

References

Ahmed, Arif (2014a). "Dicing with Death". In: *Analysis* 74.4, pp. 587–592.

— (2014b). *Evidence, Decision and Causality*. Cambridge: Cambridge University Press.
Ahmed, Arif and Huw Price (2012). "Arntzenius on 'Why ain'cha rich?'" In: *Erkenntnis* 77.1, pp. 15–30.
Almond, Paul (2010). *On Causation and Correlation Part 1: Evidential decision theory is correct*. URL: https://casparoesterheld.files.wordpress.com/2016/12/almond_edt_1.pdf (visited on 03/23/2018).
Altair, Alex (2013). *A Comparison of Decision Algorithms on Newcomblike Problems*. Tech. rep. Machine Intelligence Research Institute. URL: https://intelligence.org/files/Comparison.pdf.
Arntzenius, Frank (2008). "No Regrets, or: Edith Piaf Revamps Decision Theory". In: *Erkenntnis* 68.2, pp. 277–297.
— (2010). "Reichenbach's Common Cause Principle". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N Zalta. Fall 2010. Metaphysics Research Lab, Stanford University.
Arntzenius, Frank, Adam Elga, and John Hawthorne (2004). "Bayesianism, Infinite Decisions, and Binding". In: *Mind* 113.450, pp. 251–283.
Barnes, R Eric (1997). "Rationality, Dispositions, and the Newcomb Paradox". In: *Philos. Stud.* 88.1, pp. 1–28.
Bjerring, Jens Christian (2014). "On counterpossibles". In: *Philos. Stud.* 168.2, pp. 327–353.
Bostrom, Nick (2001). "The Meta-Newcomb Problem". In: *Analysis* 61.272, pp. 309–310.
— (2011). "Infinite Ethics". In: *Analysis and Metaphysics* 10, pp. 9–59.
Bourget, David and David J Chalmers (2014). "What Do Philosophers Believe?" In: *Philos. Stud.* 170.3, pp. 465–500.
Brams, Steven J (1975). "Newcomb's Problem and Prisoners' Dilemma". In: *J. Conflict Resolut.* 19.4, pp. 596–612.
— (1976). *Paradoxes in Politics: An Introduction to the Nonobvious in Political Science*. New York: Free Press.
— (2014). *Rational Politics: Decisions, Games, and Strategy*. Boston: Academic Press.
Briggs, Rachael (2010). "Decision-Theoretic Paradoxes as Voting Paradoxes". In: *Philos. Rev.* 119.1, pp. 1–30.
Broome, John (1989). "An Economic Newcomb Problem". In: *Analysis* 49.4, pp. 220–222.
Burgess, Simon (2004). "The Newcomb Problem: An Unqualified Resolution". In: *Synthese* 138.2, pp. 261–287.
— (2012). "Newcomb's Problem and its Conditional Evidence: A Common Cause of Confusion". In: *Synthese* 184.3, pp. 319–339.
Cartwright, Nancy (1999). "Causal Diversity and the Markov Condition". In: *Synthese* 121.1-2, pp. 3–27.
Dummett, Michael (1964). "Bringing About the Past". In: *Philos. Rev.* 73.3, pp. 338–359.
Eells, Ellery (1982). *Rational Decision and Causality*. Cambridge: Cambridge University Press.
— (1984). "Metatickles and the Dynamics of Deliberation". In: *Theory Decis.* 17.1, pp. 71–95.
Egan, Andy (2007). "Some Counterexamples to Causal Decision Theory". In: *Philos. Rev.* 116.1, pp. 93–114.
Fisher, Justin C (n.d.). "Disposition-Based Decision Theory". URL: http://www.justin-fisher.com/papers/DBDT.pdf (visited on 03/23/2018).
Frydman, Roman, Gerald P O'Driscoll, and Andrew Schotter (1982). "Rational Expectations of Government Policy: An Application of Newcomb's Problem". In: *South. Econ. J.* 49.2, pp. 311–319.

Gibbard, Allan and William L Harper (1978). "Counterfactuals and Two Kinds of Expected Utility". In: *Foundations and Applications of Decision Theory*. Ed. by Clifford Alan Hooker, James J Leach, and Edward Francis McClennen. The University of Western Ontario Series in Philosophy of Science. Springer Netherlands, pp. 125–162.

Grafstein, Robert (1991). "An Evidential Decision Theory of Turnout". In: *Am. J. Pol. Sci.* 35.4, pp. 989–1010.

— (2002). "What Rational Political Actors Can Expect". In: *J. Theor. Polit.* 14.2, pp. 139–165.

Hitchcock, Christopher (2016). "Probabilistic Causation". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N Zalta. Winter 2016. Metaphysics Research Lab, Stanford University.

Hofstadter, Douglas (1996). *Metamagical Themas: Questing For The Essence Of Mind And Pattern*. New York: Basic Books.

Horgan, Terence (1981). "Counterfactuals and Newcomb's Problem". In: *J. Philos.* 78.6, pp. 331–356.

Hurley, Susan L (1991). "Newcomb's Problem, Prisoners' Dilemma, and collective action". In: *Synthese* 86.2, pp. 173–196.

Jeffrey, Richard (1965). *The Logic of Decision*. Chicago: University of Chicago Press.

— (2004). *Subjective Probability: The Real Thing*. Cambridge: Cambridge University Press.

Joyce, James M (1999). *The Foundations of Causal Decision Theory*. Cambridge: Cambridge University Press.

Kavka, Gregory S (1983). "The toxin puzzle". In: *Analysis* 43.1, pp. 33–36.

Ledwig, Marion (2000). "Newcomb's problem". PhD thesis. URL: https://kops.uni-konstanz.de/bitstream/handle/123456789/3451/ledwig.pdf (visited on 03/23/2018).

Lewis, David (1979). "Prisoners' Dilemma is a Newcomb Problem". In: *Philos. Public Aff.* 8.3, pp. 235–240.

— (1981). "Causal Decision Theory". In: *Australas. J. Philos.* 59.1, pp. 5–30.

— (1986). "Events". In: *Philosophical Papers Vol. II*. Ed. by David Lewis. Oxford: Oxford University Press, pp. 241–269.

MacAskill, William (2016). "Smokers, Psychos, and Decision-Theoretic Uncertainty". In: *The Journal of Philosophy* 113.9, pp. 425–445.

Meacham, Christopher J G (2010). "Binding and its consequences". In: *Philos. Stud.* 149.1, pp. 49–71.

Mellor, D H (1987). "Fixed Past, Unfixed Future". In: *Michael Dummett*. Ed. by Barry M Taylor. Nijhoff International Philosophy Series. Springer Netherlands, pp. 166–186.

Muth, John F (1961). "Rational Expectations and the Theory of Price Movements". In: *Econometrica* 29.3, pp. 315–335.

Nolan, Daniel P (2013). "Impossible Worlds". In: *Philosophy Compass* 8.4, pp. 360–372.

Nozick, Robert (1969). "Newcomb's problem and two principles of choice". In: *Essays in honor of Carl G. Hempel*. Springer, pp. 114–146.

Oesterheld, Caspar (2017). *Multiverse-wide Cooperation via Correlated Decision Making*. Tech. rep. Foundational Research Institute. URL: https://foundational-research.org/multiverse-wide-cooperation-via-correlated-decision-making/.

Parfit, Derek (1986). *Reasons and Persons*. Oxford: Oxford University Press.

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

Poellinger, Roland (2013). *Unboxing the Concepts in Newcomb's Paradox: Causation, Prediction, Decision in Causal Knowledge Patterns*. URL: http://philsci-archive.pitt.edu/9887/7/newcomb_in_ckps.pdf (visited on 03/23/2018).
Price, Huw (1986). "Against Causal Decision Theory". In: *Synthese* 67.2, pp. 195–212.
— (1991). "Agency and probabilistic causality". In: *Br. J. Philos. Sci.* 42.2, pp. 157–176.
Rapoport, Anatol (1966). *Two-Person Game Theory*. Ann Arbor: University of Michigan Press.
Reichenbach, Hans (1956). *The Direction of Time*. Berkeley: University of California Press.
Resnik, Michael D (1987). *Choices: An Introduction to Decision Theory*. Minneapolis: University of Minnesota Press.
Shafir, Eldar and Amos Tversky (1992). "Thinking Through Uncertainty: Nonconsequential Reasoning and Choice". In: *Cogn. Psychol.* 24.4, pp. 449–474.
Skyrms, Brian (1980). *Causal Necessity*. New Haven: Yale University Press.
— (1982). "Causal Decision Theory". In: *J. Philos.* 79.11, pp. 695–711.
Soares, Nate and Benja Fallenstein (2015). "Toward Idealized Decision Theory". In: arXiv: 1507.01986 [cs.AI].
Soares, Nate and Ben Levinstein (2017). "Cheating Death in Damascus". In: *Formal Epistemology Workshop (FEW)*. (University of Washington, Seattle, USA). URL: http://intelligence.org/files/DeathInDamascus.pdf (visited on 03/23/2018).
Sobel, Jordan Howard (1986). "Notes on decision theory: Old wine in new bottles". In: *Australas. J. Philos.* 64.4, pp. 407–437.
— (1990). "Newcomblike Problems". In: *Midwest Studies In Philosophy* 15.1, pp. 224–255.
— (1994). *Taking Chances: Essays on Rational Choice*. Cambridge: Cambridge University Press.
Spirtes, Peter, Clark N Glymour, and Richard Scheines (2000). *Causation, Prediction, and Search*. Cambridge: MIT Press.
Spohn, Wolfgang (2003). "Dependency Equilibria and the Causal Structure of Decision and Game Situation". In: *Homo Oeconomicus* 20, pp. 195–255.
— (2005). "The 5 Questions". In: *Formal Philosophy*. Ed. by Vincent F Hendricks and John Symons. Copenhagen: Automatic Press.
— (2012). "Reversing 30 Years of Discussion: Why Causal Decision Theorists Should One-Box". In: *Synthese* 187.1, pp. 95–122.
Stalnaker, Robert C (1980). "Letter to David Lewis". In: *IFS*. Ed. by William L Harper, Robert Stalnaker, and Glenn Pearce. The University of Western Ontario Series in Philosophy of Science. Springer Netherlands, pp. 151–152.
Tversky, Amos and Daniel Kahneman (1974). "Judgment under Uncertainty: Heuristics and Biases". In: *Science* 185.4157, pp. 1124–1131.
Weiner, Matt (2004). *A Limiting Case for Causal/Evidential Decision Theory*. Opiniatrety: Half- to Quarter-Baked Thoughts. URL: http://mattweiner.net/blog/archives/000406.html (visited on 03/23/2018).
Weirich, Paul (1985). "Decision Instability". In: *Australas. J. Philos.* 63.4, pp. 465–472.
— (2016). "Causal Decision Theory". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N Zalta. Winter 2016. Metaphysics Research Lab, Stanford University.
Yudkowsky, Eliezer (2010). *Timeless Decision Theory*. Tech. rep. Machine Intelligence Research Institute. URL: https://intelligence.org/files/TDT.pdf.
Yudkowsky, Eliezer and Nate Soares (2017). "Functional Decision Theory: A New Theory of Instrumental Rationality". In: arXiv: 1710.05060 [cs.AI].