Approval-directed agency and the decision theory of Newcomb-like problems

Caspar Oesterheld

January 13, 2018

Abstract

Decision theorists disagree about how instrumentally rational agents, i.e., agents trying to achieve some goal, should behave in so-called Newcomb-like problems, with the main contenders being causal and evidential decision theory. Since the main goal of artificial intelligence research is to create machines that make instrumentally rational decisions, the disagreement pertains to this field. In addition to the more philosophical question of what the right decision theory is, the goal of AI poses the question of how to implement any given decision theory in an AI. For example, how would one go about building an AI whose behavior matches evidential decision theory's recommendations? Conversely, we can ask which decision theories (if any) describe the behavior of any existing AI design. In this paper, we study what decision theory an approval-directed agent, i.e., an agent whose goal it is to maximize the score it receives from an overseer, implements. If we assume that the overseer rewards the agent based on the expected value of some vNM utility function, then such an approval-directed agent is guided by two decision theories: the one used by the agent to decide which action to choose in order to maximize the reward and the one used by the overseer to compute the expected utility of a chosen action. We show which of these two decision theories describes the agent's behavior in which situations.

Keywords: reinforcement learning, causal decision theory, evidential decision theory, Newcomb's problem, AI safety, philosophical foundations of AI

Contents

1 Introduction

In decision theory, there is a large debate about how an instrumentally rational agent, i.e., an agent trying to achieve some goal or maximize some utility function, should decide in Newcomb's problem (introduced by Nozick, 1969) and variations thereof (a list is given by Ledwig, 2000, pp. 80-87). Consequently, different normative theories of instrumental rationality have been developed. The best known ones are evidential (sometimes also called Bayesian) decision theory (EDT) (Ahmed, 2014; Almond, 2010; Price, 1986; Horgan, 1981) and causal decision theory (CDT) (Gibbard and Harper, 1981; Joyce, 1999; Lewis, 1981; Skyrms, 1982; Weirich, 2016), but many have attempted to remediate what they view as failures of the two theories by proposing further alternatives (Spohn, 2003; Spohn, 2012; Poellinger, 2013; Arntzenius, 2008; Gustafsson, 2011; Wedgwood, 2013; Dohrn, 2015; Price, 2012; Yudkowsky, 2010; Soares and Levinstein, 2017).

Because the main goal of artificial intelligence is to build machines that make instrumentally rational decisions (Russell and Norvig, 2010, sect. 1.1.4, 2.2; Legg and Hutter, 2007; Doyle, 1992), this normative disagreement has some bearing on how to build these machines (cf. Soares and Fallenstein, 2014a, sect. 2.2; Soares and Fallenstein, 2014b, sect. 1; Bostrom, 2014b, ch. 13, sect. "Decision theory"). The differences between these decision theories are probably inconsequential in most situations (Ahmed, 2014, sect. 0.5, ch. 4; Briggs, 2017),[1] such that it may not matter in most current applications of AI. That said, the differences may have implications in some situations (Ahmed, 2014, ch. 4-6; Soares, 2014a; Oesterheld, 2017b; Bostrom, 2014a). In fact, AI may expose them more often. For example, Newcomb's problem and the prisoner's dilemma with a replica (Kuhn, 2017, sect. 7) are easy to implement for agents with copyable source code (cf. Yudkowsky, 2010, pp. 85ff. Soares and Fallenstein, 2014b, sect. 2; Soares, 2014b; Cavalcanti, 2010, sect. 5). Indeed, the existence of many copies is the norm for (successful) software, including AI-based software. While copies of present-day software systems may only interact with each other in rigid, explicitly pre-programmed ways, future AI-based systems will make decisions in a more autonomous, flexible and goal-driven way. Overall, the decision theory of Newcomb-like scenarios may not rank among the most pressing issues in the near-term deployment of AI, but it is a central foundational issue which may become practically important in the longer term.

The problem for AI research posed by the disagreement among decision theorists can be divided into two questions:

1. What decision theory do we want an AI to follow?
2. How could we implement such a decision theory in an AI? Or: How do decision theories and AI frameworks or architectures map onto each other?

Although it certainly requires further discussion, there already is a large literature

[1]In fact, Eells (1981) has argued that EDT and CDT always behave in the same way, but I disagree with this assessment based on the reasons given by Ahmed (2014, sect. 4.3-4.6) and Price (1986).

related to the first question.[2] In this paper, we would thus like to draw attention to the second question.

Specifically, we would like to investigate how approval-directed agents behave in Newcomb-like problems, where by an approval-directed agent we mean an agent choosing actions to maximize the score it receives from an overseer (cf. Christiano, 2014). Approval-agency is intended as a model of reinforcement learning (RL) agents (see Sutton and Barto, 1998; Russell and Norvig, 2010, ch 17, 21, for introductions to RL), for whom the reward function fills the role of the overseer. Since reinforcement learning is such a general and commonly studied problem in artificial intelligence (Hutter, 2005, e.g. ch. 4.1.3; Russell and Norvig, 2010, p. 831; Sutton and Barto, 1998, ch. 1), it is an especially attractive target for modeling.[3] However, because decision theories are usually defined only for single decisions, we will only discuss single decisions whereas reinforcement learning is usually concerned with sequential interactions of agent and environment. However, this decision can also be a policy choice to model sequential decision problems.[4] Moreover, we will not discuss the learning process and simply assume that the agent has already formed some model of the world. A model of approval-directed agency that allows us to describe Newcomb-like situations is described and discussed in sect. 2.

If we assume that, after an action has been taken, the overseer rewards the agent based on the expected value of some von Neumann-Morgenstern utility function, the agent is implicitly driven by two decision theories: The overseer can use the regular conditional expectation or the causal expectation to estimate the value of its utility function; and the agent itself can follow CDT or EDT when maximizing the score it receives from the overseer (sect. 3).

We then show how the overall decision theory depends on these two potentially conflicting decision theories. If the overseer bases its expected value calculations

[2] Of course, the existing literature asks about the right decision theory proper. The answer to that question might differ from the answer to the AI-specific question (Kumar, 2017). After all, even if we have identified the right decision theory for ourselves, we may want to implement a different decision theory in an AI. One reason could be that the main contenders are not self-recommending – it has been pointed out that EDT and CDT both recommend to self-modify into slightly different decision theories (Meacham, 2010; Soares and Fallenstein, 2014b, sect. 3; Yudkowsky, 2010, sect. 2; Greene, forthcoming). The same arguments imply that even if one is convinced of CDT or EDT one would not want the AI to use CDT and EDT. That said, one could also leave the self-modification to the AI.

[3] Reinforcement learning and approval-directed agency are also common outside of artificial intelligence. For example, Achen and Bartels (2016, ch. 4) review evidence which shows that the electorate often votes retrospectively to punish or reward incumbents.

[4] This is consistent with what reinforcement learning algorithms usually do – they choose policies rather than individual actions. This is because the utility of a single action usually cannot be evaluated without knowing how the agent will deal with situations that might arise as a result of taking that action.

When individual actions *can* be evaluated in isolation, the *ex ante* policy choice sometimes differs from the choice of individual actions (see the absent-minded driver, introduced by Piccione and Rubinstein, 1997; cf. Aumann, Hart, and Perry, 1997; the Newcomb-like scenarios discussed by, e.g., Hintze, 2014; Soares and Levinstein, 2017, sect. 2; and the problems in anthropics discussed by Armstrong, 2011). While it is rarely discussed in the debate between evidential and causal decision theorists, a few authors regard this discrepancy as crucial and have argued that a proper decision theory should be be about optimal policy choices (e.g. Hintze, 2014; Soares and Fallenstein, 2014b, sect. 2.1; Soares and Levinstein, 2017, sect. 2). However, this issue is beyond the scope of the present paper.

Further issues in sequential Newcomb-like problems are discussed by Everitt, Leike, and Hutter (2015).

on looking only at the world, then the agent's decision theory is decisive. If the overseer bases its estimates only on the agent's action, then the overseer's decision (or perhaps rather action evaluation) theory is decisive.

2 Approval-directed agency

We first describe a model of approval-directed agency. To be able to apply both CDT and EDT, we will use causal models in Pearl's (2009) sense. Consequently, we use Pearl's *do*-calculus-based version of CDT (Pearl, 2009, ch. 4). We will, throughout this paper, assume that the agent has already formed a (potentially implicit) model of the world[5] – e.g., based on past interactions with the environment. Also, we will only consider single decisions rather than sequential problems of iterative interaction between agent and environment.

A causal model of such a one-shot Newcomb problem from the perspective of the approval-directed agent is given in figure 1. In this model, the agent decides to take some action A, which may causally affect some part of the environment history, i.e., the history of states, H. We will call that part of the history the agent's causal future H_f. Furthermore, the agent may be causally influenced by some other part of the environment history, which we will call the agent's causal past H_p. H may contain information other than H_f and H_p, which we will assume to be independent of A.[6] The *overseer*, physically realized by, e.g., some module physically attached to the agent or a human supervisor, observes the agent's action and partially, via some percept O, the state of the world[7]. The overseer then calculates the reward R. To set proper incentives to the agent, we will assume the overseer to know not only the action and observation, but also everything that the agent knows (cf. Christiano, 2016). The overseer may also have access to some additional piece of information

[5]There is a broad philosophical literature on whether causal relationships exist and whether they can be inferred in cases where the agent is part of the environment. See, e.g., the edited volume by Price and Corry (2007).

[6] For simplicity, we will ignore dependences not resulting from causation (Arntzenius, 2010). For example, if you play against a copy, there is a logical dependence between your and your copy's decision. Even if you know a set of nodes in the causal graph that *d*-separates your and your copy's decision (e.g., if you know your common source code), the dependence persists. We exclude these dependences because such situations cannot be modeled by standard causal graphs.

However, we could adapt causal graphs to accomodate for these kinds of dependences. First, we could modify our definition of causality in such a way that dependence does imply causation, as has been proposed by Spohn (2003, 2012), Yudkowsky (2010) and others. For instance, we could model the dependence between the outputs of two instances of an algorithm by introducing a logical node as a common cause of the two. This logical node would then represent the output of the abstract algorithm that the two copies implement. While changes to the concept of causation may affect CDT's implied behavior, the results from this paper can be directly transfered to such modifications.

Alternatively, we could extend causal graphs to also include non-causal dependences (cf. Poellinger, 2013). Such extension necessitates a new CDT formalism, so the proofs from this paper do not directly transfer to this case. That said, I expect our results to generalize given that both EDT and CDT would probably treat non-causal dependences just like they treat causal arrows directed toward the action.

[7] Christiano (2014) does not define approval-directed agency formally, but judging from a comment he made at `https://medium.com/paulfchristiano/i-agree-that-the-key-feature-of-approval-directed-agents-is-that-the-causal-picture-is-736b4474910e`, he considers it crucial to his conception that the overseer only looks at the agent's action and does not observe the action's consequences (cf. the distinction introduced in sect. 3).

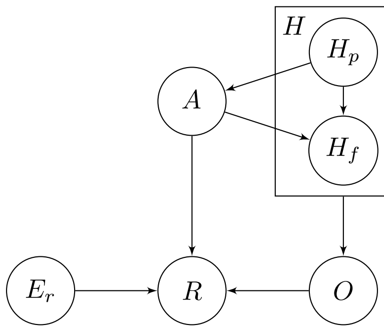


Figure 1: A causal model of an approval-directed agent in a Newcomb-like decision problem. A denotes the agent's action, H the environment history, O the observation on which the overseer bases the reward, R is that reward, and E_r is information about the way the reward is computed that is only available to the overseer. The box is used to indicate that H includes the two random variables H_p and H_f. All of H may have a causal influence on O.

E_r about the way the reward is to be calculated.[8] Lastly, we assume that the sets of possible values of A, O and E_r are finite.

In principle, the overseer could reward the agent in all kinds of ways. E.g., it could reward the agent "deontologically" (Alexander and Moore, 2016) for taking a particular action independently of the consequences of taking that action. In this paper, we will assume that the reward estimates the value of some von Neumann-Morgenstern utility function U that only depends on states of the world. I use the capital U to indicate that the utility function, too, is a random variable (in the Bayesian sense). For simplicity's sake, we will, again, assume that the set of possible values of U is finite.

We will view U as representing the system designer's preferences over world states.[9] While other ways of assigning the reward are possible, this is certainly an attractive way of getting an approval-directed agent to achieve goals that we want it to achieve. After all, in real-world applications, we will usually care about the outcomes of the agent's decisions, such as whether a car has reached its destination in time or whether a human has been hurt.

The standard way of estimating $U(H)$ (or any quantity for that matter) is the familiar conditional expectation. Thus, the overseer may compute the reward as

$$r = \mathbb{E}\left[U(H) \mid e_r, a, o\right], \tag{1}$$

where r, a, e_r, and o are values of R, A, E_r, and O, respectively.[10] But since this reward will be used to incentivize an agent, a causal decision theorist may want the

[8] One reason for the overseer to have access to such additional information is that some of the human supervisor's values may not be expressible in a way that the approval-directed agent's algorithm can utilize (cf. Muehlhauser and Helm, 2012, sect. 3, 4, 5.3).

[9] Some have tried to modify the reward relative to the designer's preferences to make the reinforcement learning problem easier to solve (Sorg, 2011), although Sutton and Barto (1998, sect. 3.2) explicitly discourage such tricks in their reinforcement learning textbook.

[10] At first sight this may be confusing to some readers, because in reinforcement learning utility sometimes refers to expected cumulative reward (Russell and Norvig, 2010, ch. 17, 21), although

reward to be computed according to

$$r = \mathbb{E}\left[U(H) \mid e_r, do(a), o\right], \tag{2}$$

although, of course, this term does not correctly estimate the utility after the agent has already decided to take action a.[11] Here, $do(a)$ refers to Pearl's do-calculus, where conditioning on $do(a)$ roughly means intervening from outside the causal model to set A to a. For an introduction to the do-calculus, see Pearl (2009).

An approval-directed agent is characterized by maximizing the reward it receives from the overseer.[12] However, decision theory offers us (at least) two different expected values, the regular expected value of EDT

$$\mathbb{E}\left[R \mid a\right], \tag{3}$$

and CDT's causal expected value

$$\mathbb{E}\left[R \mid do(a)\right]. \tag{4}$$

We leave the interesting question of which (if any) decision theory describes the behavior of current RL algorithms to future research[13] and in the following assume that the agent is known to implement either CDT or EDT.

3 The conflict of the decision theories of agent and overseer

When viewed together with the overseer, our agent may now be seen as containing two decision theories, one for computing the reward and one in the algorithm that tries to find the action to maximize that reward. These decision theories may not always be the same.

Given this potential discrepancy, the question is which of the two decision theories prevails, i.e., for which configurations of the two decision theories the overall agent acts like a CDT agent and for which it acts like an EDT agent w.r.t. u.

As it turns out, the answer to this question depends on the decision problem in question. In particular, it depends on whether the overseer updates its estimate of

others use the term *value function* instead (Sutton and Barto, 1998, sect. 3.7). Here, u does not refer to utility in that sense but in the decision-theoretical sense of representing intrinsic values. So, in the present case, we have two "layers" of goals: first, the agent maximizes the reward. Second, the agent as incentivized by the overseer's way of calculating rewards maximizes u.

One cause of confusion is that in model applications of RL, the reward function possesses full knowledge of the world state and thus does not require the use of the expectation operator.

[11] If the disagreement in Newcomb's problem is to be about CDT versus EDT rather than the predictive abilities of "the being", Omega or the psychologist, then after requesting both boxes a proponent of two-boxing knows that she will probably receive only $1,000. Causal and evidential decision theorists agree that regular conditional expectation is the correct way of updating one's beliefs about the state of the world after an action has been taken (cf. the distinction between "acts" and "actions" in Pearl, 2009, sect. 4.1.1). However, since the overseer's reward function is the optimization criterion for the agent's decisions, causal decision theorists may want it to reward the agent not based on regular expected utility but on causal expected utility.

[12] In RL, some have proposed alternative optimization targets that incorporate, e.g., risk aversion (García and Fernández, 2015, sect. 3).

[13] For preliminary work on this question, see Mayer, Feldmaier, and Shen (2016) and Oesterheld (2017a).

$U(H)$ primarily based on the action taken by the agent or on its observation of the environment.

For illustration, consider two versions of Newcomb's problem. In the first, the overseer eventually sees the payoff, i.e., how much money the agent has made. In this case, as soon as the money is observed, the overseer's estimate of $U(H)$ becomes independent of the agent's action. More generally, O may tell the overseer so much about $U(H)$ that it becomes independent of A even if $U(H)$ is not yet fully observed. That is,

$$\mathbb{E}\left[U(H) \mid e_r, a, o\right] = \mathbb{E}\left[U(H) \mid e_r, o\right] \tag{5}$$

and

$$\mathbb{E}\left[U(H) \mid e_r, do(a), o\right] = \mathbb{E}\left[U(H) \mid e_r, o\right] \tag{6}$$

for all e_r, a and o. Note that neither of these two implies the other.[14]

In the second version of Newcomb's problem, the monetary payoff is not observed but covertly invested into increasing the agent's utility function. Only the agent's choice can then inform the overseer about $U(H)$. Formally, it is both

$$\mathbb{E}\left[U(H) \mid e_r, a, o\right] = \mathbb{E}\left[U(H) \mid e_r, a\right] \tag{7}$$

and

$$\mathbb{E}\left[U(H) \mid e_r, do(a), o\right] = \mathbb{E}\left[U(H) \mid e_r, do(a)\right]. \tag{8}$$

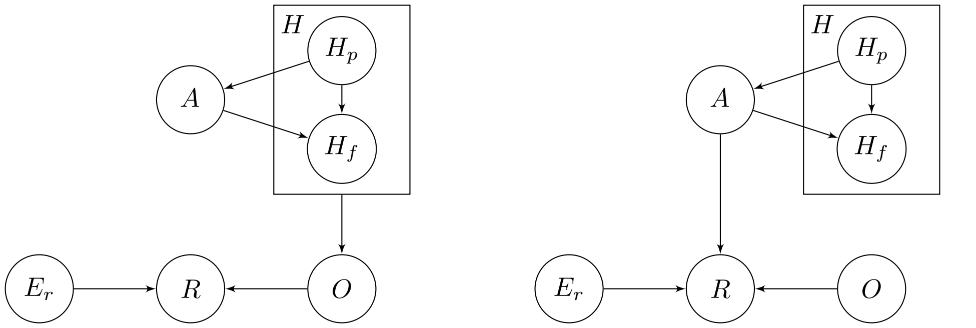
Again, we assume that this is known to the agent. An example class of cases is that in which the agent's decisions are correlated with those of agents in far-away parts of the environment (Oesterheld, 2017b; c.f. Treutlein and Oesterheld, 2017). The two versions are depicted in figure 2.

Of course, these are only the two extremes from the set of all possible situations. In real-world Newcomb-like scenarios, the overseer may also draw some information from both sources. Nonetheless, it seems useful to understand the extreme cases, as this may also help us understand mixed ones.

In the following subsections, we will show that in the first type, the decision theory of the agent is decisive, whereas in the second type, the overseer's decision theory is[15]. We will do so by considering all possible configurations of the type of the problem, the overseer's decision theory and agent decision theory. While we will limit our analysis to EDT and CDT, the results can easily be generalized to variants of these that arise from modifying the causal model or conditional credence distribution (e.g. Yudkowsky, 2010; Fisher, n.d.; Spohn, 2012; Dohrn, 2015). The analysis is summarized in table 1.

[14] We give a brief justification of this claim. If all of a's causal influence on H can be discerned from O, then, of course, a could still be diagnostically relevant for one's estimate of $U(H)$. The other direction is more complicated. The idea is that eq. 5 can be true if the causal and non-causal implications of a exactly cancel each other out. An example is a version of Newcomb's problem in which one-boxing ensures with certainty that both boxes contain the same amount of money. Then if O and E_r do not contain any information, the expected value of two-boxing and one-boxing is the same and so learning of the action is irrelevant for estimating $U(H)$. However, two-boxing is causally better than one-boxing, so eq. 6 is violated.

[15] The dominance of the overseer's decision theory in the second type of Newcomb's problem is mentioned (though not proven) by Christiano (2014, sect. "Avoid lock-in").



(a) The reward is computed based on observing the state of the world.

(b) The reward is computed based on observing the agent's action.

Figure 2: Two different ways in which the overseer can calculate the reward.

Type of Newcomb problem	Agent's DT	Reward module's DT	Resulting DT
First type	CDT	EDT	CDT
		CDT	CDT
	EDT	EDT	EDT
		CDT	EDT
Second type	CDT	EDT	EDT
		CDT	CDT
	EDT	EDT	EDT
		CDT	CDT

Table 1: An overview of the results of the calculations in sect. 3.

3.1 First type

3.1.1 The EDT agent

The EDT agent judges its action by

$$\mathbb{E}\left[R \mid a\right]. \tag{9}$$

If the overseer calculates regular conditional expectation, then it is

$$\begin{aligned}\mathbb{E}\left[R \mid a\right] &= \mathbb{E}\left[\mathbb{E}\left[U(H) \mid E_r, O, a\right] \mid a\right] && (10)\\ &= \mathbb{E}\left[U(H) \mid a\right], && (11)\end{aligned}$$

where the last line is due to what is sometimes called the law of total expectation (LTE) or the tower rule (see, e.g., Ross, 2007, sect. 3.4; Billingsley, 1995, Theorem 34.4). Intuitively, you cannot expect that gaining more evidence (i.e., E_r and O in addition to a) moves your expectation of $U(H)$ into any particular direction.

Because the overseer knows more than the agent, we will need this rule in all of the following derivations. Its application makes it hard to generalize these results to other decision theories. If the two decision theories do not both compute a form of expected utility, then LTE does not apply.

Equations 10 and 11 show that if the overseer computes regular expected value and the agent maximizes the reward according to EDT, then the agent as a whole maximizes u according to EDT.

If the overseer computes CDT's expected value, it is

$$\begin{aligned}\mathbb{E}\left[R \mid a\right] &= \mathbb{E}\left[\mathbb{E}\left[U(H) \mid E_r, do(a), O\right] \mid a\right] && (12)\\ &= \sum_{e_r,o} P(e_r, o \mid a) \cdot \mathbb{E}\left[U(H) \mid e_r, do(a), o\right] && (13)\\ &\underset{\text{eq. 5 and 6}}{=} \sum_{e_r,o} P(e_r, o \mid a) \cdot \mathbb{E}\left[U(H) \mid e_r, a, o\right] && (14)\\ &= \mathbb{E}\left[\mathbb{E}\left[U(H) \mid E_r, a, O\right] \mid a\right] && (15)\\ &\underset{\text{LTE}}{=} \mathbb{E}\left[U(H) \mid a\right] && (16)\end{aligned}$$

3.1.2 The CDT agent

The CDT agent judges its action by

$$\mathbb{E}\left[R \mid do(a)\right]. \tag{17}$$

If the overseer uses regular expected value (EDT), then

$$\begin{aligned}\mathbb{E}\left[R \mid do(a)\right] &= \mathbb{E}\left[\mathbb{E}\left[U(H) \mid a, O, E_r\right] \mid do(a)\right] && (18)\\ &= \sum_{e_r,o} P(e_r, o \mid do(a)) \cdot \mathbb{E}\left[U(H) \mid a, o, e_r\right] && (19)\\ &\underset{\text{eq. 5 and 6}}{=} \sum_{e_r,o} P(e_r, o \mid do(a)) \cdot \mathbb{E}\left[U(H) \mid do(a), o, e_r\right] && (20)\\ &= \mathbb{E}\left[\mathbb{E}\left[U(H) \mid do(a), O, E_r\right] \mid do(a)\right] && (21)\\ &\underset{\text{LTE}}{=} \mathbb{E}\left[U(H) \mid do(a)\right] && (22)\end{aligned}$$

Learning about an intervention $do(a)$ cannot always be treated in the same way as learning about other events. Hence, the application of the law of total expectation is not straightforward. However, $P(\cdot \mid do(x))$ is always a probability distribution. Because the law of total expectation applies to all probability distributions, it also applies to ones resulting from the application of the do-calculus.

If the overseer uses CDT's expected value, then

$$\begin{aligned} \mathbb{E}\left[R \mid do(a)\right] &= \mathbb{E}\left[\mathbb{E}\left[U(H) \mid E_r, O, do(a)\right] \mid do(a)\right] && (23) \\ &\underset{\text{LTE}}{=} \mathbb{E}\left[U(H) \mid do(a)\right]. && (24) \end{aligned}$$

3.2 Second type

3.2.1 The EDT agent

The EDT agent judges its actions by

$$\mathbb{E}\left[R \mid a\right]. \tag{25}$$

If the overseer is based on regular conditional expectation (EDT), then it is again

$$\begin{aligned} \mathbb{E}\left[R \mid a\right] &= \mathbb{E}\left[\mathbb{E}\left[U(H) \mid E_r, a\right] \mid a\right] && (26) \\ &\underset{\text{LTE}}{=} \mathbb{E}\left[U(H) \mid a\right]. && (27) \end{aligned}$$

If the overseer is based on CDT-type expectation, then

$$\begin{aligned} \mathbb{E}\left[R \mid a\right] &= \mathbb{E}\left[\mathbb{E}\left[U(H) \mid E_r, do(a)\right] \mid a\right] && (28) \\ &= \sum_{e_r} P(e_r \mid a) \cdot \mathbb{E}\left[U(H) \mid do(a), e_r\right] && (29) \\ &= \sum_{e_r} P(e_r) \cdot \mathbb{E}\left[U(H) \mid do(a), e_r\right] && (30) \\ &= \sum_{e_r} P(e_r \mid do(a)) \cdot \mathbb{E}\left[U(H) \mid do(a), e_r\right] && (31) \\ &= \mathbb{E}\left[\mathbb{E}\left[U(H) \mid E_r, do(a)\right] \mid do(a)\right] && (32) \\ &\underset{\text{LTE}}{=} \mathbb{E}\left[U(H) \mid do(a)\right]. && (33) \end{aligned}$$

3.2.2 The CDT agent

The CDT agent judges actions by

$$\mathbb{E}\left[R \mid do(a)\right]. \tag{34}$$

Because of Rule 2 in Theorem 3.4.1 of Pearl (2009, sect. 3.4.2) applied to the causal graph of figure 2b, it is

$$\mathbb{E}\left[R \mid do(a)\right] = \mathbb{E}\left[R \mid a\right]. \tag{35}$$

Thus, the analysis of the CDT agent is equivalent to that of the EDT agent.

4 Conclusion

In this paper, we have taken a step to map reinforcement learning architectures onto decision theories. We found that in Newcomb-like problems, if the overseer rewards the agent purely on the basis of the agent's action, then the overall system's behavior is determined by the decision theory implicit in the overseer's reward function. If the overseer judges the agent based on looking at the world, however, then the agent's decision theory is decisive.

This has implications for how we should design approval-directed agents. For instance, if we would like to leave decision-theoretical judgements to the overseer, we must ensure that the overseer assigns rewards before making new observations about the world state (cf. Christiano, 2014, sect. "Avoid lock-in"). Of course, this makes the reward less accurate and may thus slow down the agent's learning process. If we want the overseer to look at both the world and the agent's action, then we need to align both the overseer's and the agent's decision theory.

Much more research is left to be done at the intersection of decision theory and artificial intelligence. For instance, what (if any) decision theories describe the way modern RL algorithms maximize reward? Do the results of this paper generalize to sequential decision problems? Moving away from the RL framework, what decision theories do other frameworks in AI implement? What about decision theories other than CDT and EDT?

Acknowledgements

I am indebted to Max Daniel, Johannes Treutlein, Tom Everitt, Lukas Gloor, Brian Tomasik and Tobias Baumann for valuable comments and discussions.

References

Achen, Christopher H. and Larry M. Bartels (2016). *Democracy For Realists. Why Elections Do Not Produce Responsive Government.* Princeton University Press.

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

Alexander, Larry and Michael Moore (2016). "Deontological Ethics". In: *The Stanford Encyclopedia of Philosophy.* Ed. by Edward N. Zalta. Winter 2016. Metaphysics Research Lab, Stanford University. URL: `https://plato.stanford.edu/archives/win2016/entries/ethics-deontological/`.

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`.

Armstrong, Stuart (2011). *Anthropic Decision Theory.* Future of Humanity Institute. URL: `https://arxiv.org/abs/1110.6437`.

Arntzenius, Frank (2008). "No Regrets, or: Edith Piaf Revamps Decision Theory". In: *Erkenntnis* 68.2, pp. 277–297. DOI: `10.1007/s10670-007-9084-8`.

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

Aumann, Robert J., Sergiu Hart, and Motty Perry (1997). "The Absent-Minded Driver". In: *Games and Economic Behavior* 20, pp. 102–116.

Billingsley, Patrick (1995). *Probability and Measure*. 3rd ed. Wiley.

Bostrom, Nick (2014a). *Hail Mary, Value Porosity, and Utility Diversification*. URL: `http://www.nickbostrom.com/papers/porosity.pdf`.

— (2014b). *Superintelligence. Paths, Dangers, Strategies*. 1st ed. Oxford University Press.

Briggs, Rachael (2017). *Real-Life Newcomb Problems?* Talk at the 1st Workshop on Decision Theory & the Future of Artificial Intelligence in Cambridge, UK.

Cavalcanti, Eric G. (2010). "Causation, Decision Theory, and Bell's Theorem: A Quantum Analogue of the Newcomb Problem". In: *The British Journal for the Philosophy of Science* 61.3, pp. 569–597. DOI: `10.1093/bjps/axp050`.

Christiano, Paul (2014). *Model-free decisions*. URL: `https://ai-alignment.com/model-free-decisions-6e6609f5d99e`.

— (2016). *Adequate oversight*. URL: `https://ai-alignment.com/adequate-oversight-25fadf1edce9`.

Dohrn, Daniel (2015). "Egan and agents: How evidential decision theory can deal with Egan's dilemma". In: *Synthese* 192.6, pp. 1883–1908.

Doyle, Jon (1992). "Rationality and its Roles in Reasoning". In: *Computational Intelligence* 8.2, pp. 376–409. URL: `https://www.csc2.ncsu.edu/faculty/jdoyle2/publications/rrrp92.pdf`.

Eells, Ellery (1981). "Causality, Utility, and Decision". In: *Synthese* 48.2, pp. 295–329.

Everitt, Tom, Jan Leike, and Marcus Hutter (2015). "Sequential Extensions of Causal and Evidential Decision Theory". In: *Algorithmic Decision Theory: 4th International Conference, ADT 2015, Lexington, KY, USA, September 27-30, 2015, Proceedings*. Ed. by Toby Walsh. Springer, pp. 205–221. DOI: `10.1007/978-3-319-23114-3_13`.

Fisher, Justin C. (n.d.). "Disposition-based decision theory". URL: `http://www.justin-fisher.com/papers/DBDT.pdf`.

García, Javier and Fernando Fernández (2015). "A Comprehensive Survey on Safe Reinforcement Learning". In: *Journal of Machine Learning Research* 16, pp. 1437–1480.

Gibbard, Allan and William L. Harper (1981). "Counterfactuals and Two Kinds of Expected Utility". In: *Ifs. Conditionals, Belief, Decision, Chance and Time*. Ed. by William L. Harper, Robert Stalnaker, and Glenn Pearce. Vol. 15. The University of Western Ontario Series in Philosophy of Science. A Series of Books in Philosophy of Science, Methodology, Epistemology, Logic, History of Science, and Related Fields. Springer, pp. 153–190. DOI: `10.1007/978-94-009-9117-0_8`.

Greene, Preston (forthcoming). "Success-First Decision Theories". In: *Newcomb's Problem*. Ed. by Arif Ahmed. Cambridge University Press. URL: `http://prestongreene.org/s/Greene-Success-First-Decision-Theories.pdf`.

Gustafsson, Johan E. (2011). "A Note in Defence of Ratificationism". In: *Erkenntnis* 75.1, pp. 147–150.

Hintze, Daniel (2014). *Problem Class Dominance in Predictive Dilemmas*. URL: `http://intelligence.org/files/ProblemClassDominance.pdf`.

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

Hutter, Marcus (2005). *Universal Artificial Intelligence. Sequential Decision Based on Algorithmic Probability.* Ed. by Wilfried Brauer, Grzegorz Rozenberg, and Arto Salomaa. Texts in Theoretical Computer Science. Springer.

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

Kuhn, Steven (2017). "Prisoner's Dilemma". In: *The Stanford Encyclopedia of Philosophy.* Ed. by Edward N. Zalta. Spring 2017. Metaphysics Research Lab, Stanford University. URL: `https://plato.stanford.edu/archives/spr2017/entries/prisoner-dilemma/`.

Kumar, Ramana (2017). *New Work for Decision Theorists.* Talk at the 1st Workshop on Decision Theory & the Future of Artificial Intelligence in Cambridge, UK.

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

Legg, Shane and Marcus Hutter (2007). "Universal Intelligence: A Definition of Machine Intelligence". In: *Minds and Machines* 17.4, pp. 391–444.

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

Mayer, Dominik, Johannes Feldmaier, and Hao Shen (2016). "Reinforcement Learning in Conflicting Environments for Autonomous Vehicles". In: *International Workshop on Robotics in the 21st century: Challenges and Promises.* URL: `https://arxiv.org/abs/1610.07089`.

Meacham, Christopher J. G. (2010). "Binding and its consequences". In: *Philosophical Studies* 149.1, pp. 49–71. DOI: `10.1007/s11098-010-9539-7`.

Muehlhauser, Luke and Louie Helm (2012). *Intelligence Explosion and Machine Ethics.* Machine Intelligence Research Institute. URL: `https://intelligence.org/files/IE-ME.pdf`.

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

Oesterheld, Caspar (2017a). *Doing what has worked well in the past leads to evidential decision theory.* URL: `https://casparoesterheld.files.wordpress.com/2017/09/learningdt.pdf`.

— (2017b). *Multiverse-wide cooperation via correlated decision-making.* URL: `https://foundational-research.org/files/Multiverse-wide-Cooperation-via-Correlated-Decision-Making.pdf`.

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

Piccione, Michele and Ariel Rubinstein (1997). "On the Interpretation of Decision Problems with Imperfect Recall". In: *Games and Economic Behavior* 20, pp. 3–24.

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

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

Price, Huw (2012). "Causation, Chance, and the Rational Significance of Supernatural Evidence". In: *Philosophical Review* 121.4, pp. 483–538.
Price, Huw and Richard Corry, eds. (2007). *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited.* Oxford University Press.
Ross, Sheldon M. (2007). *Introduction to Probability Models.* 9th ed. Academic Press.
Russell, Stuart and Peter Norvig (2010). *Artificial Intelligence. A modern approach.* 3rd ed. Pearson Education, Inc.
Skyrms, Brian (1982). "Causal Decision Theory". In: *The Journal of Philosophy,* 79.11, pp. 695–711.
Soares, Nate (2014a). *Newcomblike problems are the norm.* URL: `http://mindingourway.com/newcomblike-problems-are-the-norm/`.
— (2014b). *Why Ain't You Rich?* URL: `https://intelligence.org/2014/10/07/nate-soares-talk-aint-rich/`.
Soares, Nate and Benja Fallenstein (2014a). *Aligning Superintelligence with Human Interests: A Technical Research Agenda.* Tech. rep. 2014-8. Machine Intelligence Research Institute. URL: `https://intelligence.org/files/TechnicalAgenda.pdf`.
— (2014b). *Toward Idealized Decision Theory.* Tech. rep. 2014-7. Machine Intelligence Research Institute. URL: `https://arxiv.org/abs/1507.01986`.
Soares, Nate and Benjamin A. Levinstein (2017). "Cheating Death in Damascus". URL: `https://intelligence.org/files/DeathInDamascus.pdf`.
Sorg, Jonathan Daniel (2011). "The Optimal Reward Problem: Designing Effective Reward for Bounded Agents". PhD thesis. University of Michigan. URL: `https://deepblue.lib.umich.edu/bitstream/handle/2027.42/89705/jdsorg_1.pdf`.
Spohn, Wolfgang (2003). "Dependency Equilibria and the Causal Structure of Decision and Game Situation". In: *Homo Oeconomicus* 20, pp. 195–255.
— (2012). "Reversing 30 years of discussion: why causal decision theorists should one-box". In: *Synthese* 187.1, pp. 95–122.
Sutton, Richard S. and Andrew G. Barto (1998). *Reinforcement Learning: An Introduction.* MIT Press.
Treutlein, Johannes and Caspar Oesterheld (2017). "A Wager for Evidential Decision Theory". Unpublished manuscript.
Wedgwood, Ralph (2013). "Gandalf's solution to the Newcomb problem". In: *Synthese* 190.14, pp. 2643–2675. DOI: `10.1007/s11229-011-9900-1`.
Weirich, Paul (2016). "Causal Decision Theory". In: *The Stanford Encyclopedia of Philosophy.* Spring 2016.
Yudkowsky, Eliezer (2010). *Timeless Decision Theory.* The Singularity Institute. URL: `http://intelligence.org/files/TDT.pdf`.