Sequential Extensions of Causal and Evidential Decision Theory*

Tom Everitt Jan Leike Marcus Hutter

June 24, 2015

Abstract

Moving beyond the dualistic view in AI where agent and environment are separated incurs new challenges for decision making, as calculation of expected utility is no longer straightforward. The non-dualistic decision theory literature is split between *causal decision theory* and *evidential decision theory*. We extend these decision algorithms to the *sequential* setting where the agent alternates between taking actions and observing their consequences. We find that evidential decision theory has two natural extensions while causal decision theory only has one.

Keywords. Evidential decision theory, causal decision theory, causal graphical models, planning, dualism, physicalism.

1 Introduction

In artificial-intelligence problems an agent interacts sequentially with an environment by taking actions and receiving percepts [RN10]. This model is *dualistic*: the agent is distinct from the environment. It influences the environment only through its actions, and the environment has no other information about the agent. The dualism assumption is accurate for an algorithm that is playing chess, go, or other (video) games, which explains why it is ubiquitous in AI research. But often it is not true: real-world agents are embedded in (and computed by) the environment [OR12], and then a *physicalistic model*[1] is more appropriate.

This distinction becomes relevant in multi-agent settings with similar agents, where each agent encounters 'echoes' of its own decision making. If the other agents are running the same source code, then the agents' decisions are logically connected. This link can be used for uncoordinated cooperation [LFY+14]. Moreover, a physicalistic model is indispensable for self-reflection. If the agent is required to autonomously verify its integrity, and perform maintenance, repair, or upgrades, then the agent needs to be aware of its own functioning. For this, a reliable and accurate self-modeling is essential. Today, applications of this level of autonomy are mostly restricted to space probes distant from earth or robots navigating lethal situations, but in the future this might also become crucial for

*The final publication is available at `http://link.springer.com/`.

[1]Some authors also call this type of model *materialistic* or *naturalistic*.

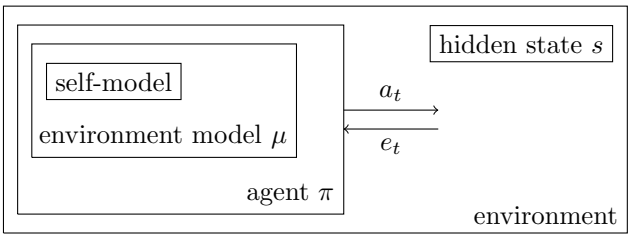


Figure 1: The physicalistic model. The hidden state s contains information about the agent that is unknown to it. The distribution μ is the agent's (subjective) *environment model*, and π its (deterministic) policy. The agent models itself through the beliefs about (future) actions given by its environment model μ. Interaction with the environment at time step t occurs through an action a_t chosen by the agent and a percept e_t returned by the environment.

sustained self-improvement in generally intelligent agents [Yud08, Bos14, SF14a, RDT+15].

In the physicalistic model the agent is embedded inside the *environment*, as depicted in Figure 1. The environment has a *hidden state* that contains information about the agent that is inaccessible to the agent itself. The agent has an *environment model* that describes the behavior of the environment given the hidden state and includes beliefs about the agent's own future actions (thus modeling itself).

Physicalistic agents may view their actions in two ways: as their selected output, and as consequences of properties of the environment. This leads to significantly more complex problems of inference and decision making, with actions simultaneously being both means to influence the environment and evidence about it. For example, looking at cat pictures online may simultaneously be a *means* of procrastination, and *evidence* of bad air quality in the room.

Dualistic decision making in a known environment is straightforward calculation of expected utilities. This is known as Savage decision theory [Sav72]. For non-dualistic decision making two main approaches are offered by the decision theory literature: *causal decision theory* (CDT) [GH78, Lew81, Sky82, Joy99, Wei12] and *evidential decision theory* (EDT) [Jef83, Bri14, Ahm14]. EDT and CDT both take actions that maximize expected utility, but differ in the way this expectation is computed: EDT uses the action under consideration as evidence about the environment while CDT does not. Section 2 formally introduces these decision algorithms.

Our contribution is to formalize and explore a decision-theoretic setting with a physicalistic reinforcement learning agent interacting *sequentially* with an environment that it is embedded in (Section 3). Previous work on non-dualistic decision theories has focused on *one-shot* situations. We find that there are two natural extensions of EDT to the sequential case, depending on whether the agent updates beliefs based on its next action or its entire policy. CDT has only one natural extension. We extend two famous *Newcomblike problems* to the sequential setting to illustrate the differences between our (generalized) decision theories.

Section 4 summarizes our results and outlines future directions. A list of

notation can be found on page 16 and Appendix A contains formal details to our examples.

2 One-Shot Decision Making

In a *one-shot decision problem*, we take one *action* $a \in \mathcal{A}$, receive a *percept* $e \in \mathcal{E}$ (typically called *outcome* in the decision theory literature) and get a *payoff* $u(e)$ according to the *utility function* $u : \mathcal{E} \to [0, 1]$. We assume that the set of actions $\mathcal{A}$ and the set of percepts $\mathcal{E}$ are finite. Additionally, the environment contains a *hidden state* $s \in \mathcal{S}$. The hidden state holds information that is inaccessible to the agent at the time of the decision, but may influence the decision and the percept. Formally, the environment is given by a probability distribution P over the hidden state, the action, and the percept that factors according to a causal graph [Pea09].

A *causal graph* over the random variables $x_1, \ldots, x_n$ is a directed acyclic graph with nodes $x_1, \ldots, x_n$. To each node x_i belongs a probability distribution $P(x_i \mid pa_i)$, where pa_i is the set of parents of x_i in the graph. It is natural to identify the causal graph with the factored distribution $P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i \mid pa_i)$. Given such a causal graph/factored distribution, we define the `do`-*operator* as

$$P(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n \mid \texttt{do}(x_j := b)) = \prod_{\substack{i=1 \\ i \neq j}}^{n} P(x_i \mid pa_i) \tag{1}$$

where x_j is set to b wherever it occurs in pa_i, $1 \leq i \leq n$. The result is a new probability distribution that can be marginalized and conditioned in the standard way. Intuitively, intervening on node x_j means ignoring all incoming arrows to x_j, as the effects they represent are no longer relevant when we intervene; the factor $P(x_j \mid pa_j)$ representing the ingoing influences to x_j is therefore removed in the right-hand side of (1). Note that the `do`-operator is only defined for distributions for which a causal graph has been specified. See [Pea09, Ch. 3.4] for details.

2.1 Savage Decision Theory

In the *dualistic* formulation of decision theory, we have a function P that takes an action a and returns a probability distribution P_a over percepts. *Savage decision theory* (SDT) [Sav72, Bri14] takes actions according to

$$\arg\max_{a \in \mathcal{A}} \sum_{e \in \mathcal{E}} P_a(e) u(e). \tag{SDT}$$

In the dualistic model it is usually conceptually clear what P_a should be. In the physicalistic model the environment model takes the form of a causal graph over a hidden state s, action a, and percept e, as illustrated in Figure 2. According to this causal graph, the probability distribution P factors causally into $P(s, a, e) = P(s)P(a \mid s)P(e \mid s, a)$. The hidden state is not independent of the decision maker's action and Savage's model is not directly applicable since we do not have a specification of P_a. How should decisions be made in this context? The literature focuses on two answers to this question: CDT and EDT.

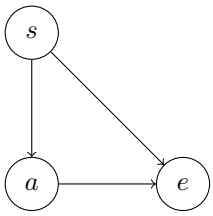


Figure 2: The causal graph $P(s, a, e) = P(s)P(a \mid s)P(e \mid s, a)$ for one-step decision making. The hidden state s influences both the decision maker's action a and the received percept e.

2.2 Causal and Evidential Decision Theory

The literature on causal and evidential decision theory is vast, and we give only a very superficial overview that is intended to bring the reader up to speed on the basics. See [Bri14, Wei12] and references therein for more detailed introductions.

Evidential decision theory (endorsed in [Jef83, Ahm14]) considers the probability of the percept e *conditional on* taking the action a:

$$\arg\max_{a \in \mathcal{A}} \sum_{e \in \mathcal{E}} P(e \mid a)\, u(e) \quad \text{with} \quad P(e \mid a) = \sum_{s \in \mathcal{S}} P(e \mid s, a) P(s \mid a) \tag{EDT}$$

Causal decision theory has several formulations [GH78, Lew81, Sky82, Joy99]; we use the one given in [Sky82], with Pearl's calculus of causality [Pea09]. According to CDT, the probability of a percept e is given by the *causal intervention* of performing action a on the causal graph from Figure 2:

$$\arg\max_{a \in \mathcal{A}} \sum_{e \in \mathcal{E}} P(e \mid \texttt{do}(a))\, u(e) \quad \text{with} \quad P(e \mid \texttt{do}(a)) = \sum_{s \in \mathcal{S}} P(e \mid s, a) P(s) \tag{CDT}$$

where $P(e \mid \texttt{do}(a))$ follows from (1) and marginalization over s.

The difference between CDT and EDT is how the action affects the belief about the hidden state. EDT assigns credence $P(s \mid a)$ to the hidden state s if action a is taken, while CDT assigns credence $P(s)$. A common argument for CDT is that an action under my direct control should not influence my belief about things that are not causally affected by the action. Hence $P(s)$ should be my belief in s, and not $P(s \mid a)$. (By assumption, the action does not *causally* affect the hidden state.) EDT might reply that if action a does not have the same likelihood under all hidden states s, then action a should indeed inform me about the hidden state, regardless of causal connection. The following two classical examples from the decision theory literature describe situations where CDT and EDT disagree. A formal definition of these examples can be found in Appendix A.

Example 1 (Newcomb's Problem [Noz69])**.** In Newcomb's Problem there are two boxes: an opaque box that is either empty or contains one million dollars and a transparent box that contains one thousand dollars. The agent can choose between taking only the opaque box ('one-boxing') and taking both boxes ('two-boxing'). The content of the opaque box is determined by a prediction about

the agent's action by a very reliable predictor: if the agent is predicted to one-box, the box contains the million, and if the agent is predicted to two-box, the box is empty. In Newcomb's problem EDT prescribes one-boxing because one-boxing is evidence that the box contains a million dollars. In contrast, CDT prescribes two-boxing because two-boxing dominates one-boxing: in either case we are a thousand dollars richer, and our decision cannot causally affect the prediction. Newcomb's problem has been raised as a critique to CDT, but many philosophers insist that two-boxing is in fact the rational choice,[2] even if it means you end up poor.

Note how the decision depends on whether the action influences the belief about the hidden state (the contents of the opaque box) or not.

Newcomb's problem may appear as an unrealistic thought experiment. However, we argue that problems with similar structure are fairly common. The main structural requirement is that $P(s \mid a) \neq P(s)$ for some state or event s that is not causally affected by a. In Newcomb's problem the predictor's ability to guess the action induces an 'information link' between actions and hidden states. If the stakes are high enough, the predictor does not have to be much better than random in order to generate a *Newcomblike decision problem*. Consider for example spouses predicting the faithfulness of their partners, employers predicting the trustworthiness of their employees, or parents predicting their children's intentions. For AIs, the potential for accurate predictions is even greater, as the predictor may have access to the AI's source code. Although rarely perfect, all of these predictions are often substantially better than random.

To counteract the impression that EDT is generally superior to CDT, we also discuss the *toxoplasmosis problem*.

Example 2 (Toxoplasmosis Problem [Alt13])**.** [3] This problem takes place in a world in which there is a certain parasite that causes its hosts to be attracted to cats, in addition to uncomfortable side effects. The agent is handed an adorable little kitten and is faced with the decision of whether or not to pet it. Petting the kitten feels nice and therefore yields more utility than not petting it. However, people suffering from the parasite are more likely to pet the kitten. Petting the kitten is evidence of having the parasite, so EDT recommends against it. CDT correctly observes that petting the kitten does not *cause* the parasite, and is therefore in favor of petting.

Newcomb's problem and the toxoplasmosis problem cannot be properly formalized in SDT, because SDT requires the percept-probabilities P_a to be specified, but it is not clear what the right choice of P_a would be. However, both CDT and EDT can be recast in the context of SDT by setting P_a to be $P(\cdot \mid \texttt{do}(a))$ and $P(\cdot \mid a)$ respectively. Thus we could say that the formulation given by Savage needs a specification of the environment that tells us whether to act evidentially, causally, or otherwise.

[2] In a 2009 survey, 31.4% of philosophers favored two-boxing, and 21.3% favored one-boxing (931 responses); see `http://philpapers.org/surveys/results.pl`. Is that the reason there are so few wealthy philosophers?

[3] Historically, this problem has been known as the *smoking lesion problem* [Ega07]. We consider the smoking lesion formulation confusing, because today it is universally known that smoking *does* cause lung cancer.

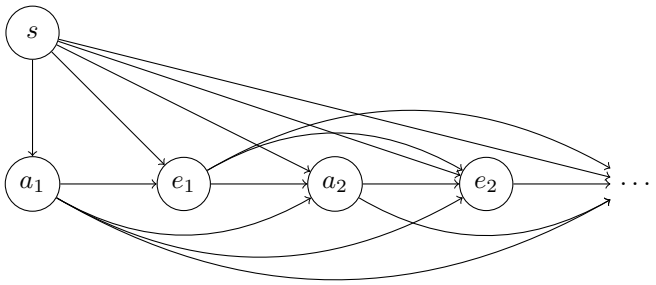


Figure 3: The (infinite) causal graph for a sequential environment. Each action a_t and each percept e_t is represented by a node in the causal graph. Actions and percepts affect all subsequent actions and percepts: causality follows time. The hidden state s is only ever indirectly (partially) observed.

3 Sequential Decision Making

In this section we extend CDT and EDT to the sequential case. We start by formally specifying the physicalistic model depicted in Figure 1 in the first subsection, and discuss problems with time consistency in Section 3.2, before defining the extensions proper in Section 3.3 and 3.4. The final subsection dissects the role of the hidden state.

3.1 The Physicalistic Model

For the remainder of this paper, we assume that the agent interacts sequentially with an environment. At time step t the agent chooses an *action* $a_t \in \mathcal{A}$ and receives a *percept* $e_t \in \mathcal{E}$ which yields a *utility* of $u(e_t) \in \mathbb{R}$; the cycle then repeats for $t+1$. A *history* is an element of $(\mathcal{A} \times \mathcal{E})^*$. We use $æ \in \mathcal{A} \times \mathcal{E}$ to denote one interaction cycle, and $æ_{<t}$ to denote a history of length $t-1$. The percepts between time t and time m are denoted $e_{t:m}$. A *policy* is a function that maps a history $æ_{<t}$ to the next action a_t. We only consider deterministic policies.

We assume that the agent is given an environment model μ, but knows neither the hidden state s nor its own future actions. The unknown hidden state may influence both percepts and actions. Actions and percepts in turn influence the entire future. The environment model μ is given by a probability distribution over hidden states and histories that factors as

$$\mu(s, æ_{<t}) = \mu(s) \prod_{i=1}^{t-1} \mu(a_i \mid s, æ_{<i}) \mu(e_i \mid s, æ_{<i} a_i) \tag{2}$$

for any $t \in \mathbb{N}$. While such a factorization is possible for any distribution, we additionally demand that this factorization is *causal* according to the causal graph in Figure 3. The distribution $\mu(a_t \mid s, æ_{<t})$ gives the likelihood of the agent's own actions provided a hidden state $s \in \mathcal{S}$ (for example, the prior probability of an infected agent petting the kitten in the toxoplasmosis problem above). For technical reasons, this distribution must always leave some uncertainty about the actions: if the environment model assigned probability zero for an action a', the agent could not deliberate taking action a' since a' could not be conditioned

on. Formally, we require $\mu(\cdot \mid s)$ to be *action-positive* for all $s \in \mathcal{S}$:

$$\forall æ_{<t} a_t \in (\mathcal{A} \times \mathcal{E})^* \times \mathcal{A}.\ \big(\mu(æ_{<t} \mid s) > 0 \implies \mu(a_t \mid s, æ_{<t}) > 0\big) \tag{3}$$

The distribution μ is a *model* of the environment, a belief held by the agent, but not the distribution from which the actual history is drawn. The actual history is distributed according to the true environment distribution. Because the environment contains the agent, the agent's algorithm might get modified by it and the actions that the agent actually ends up taking might not be the actions that were planned. In the end, model and reality will disagree: for example, we simultaneously assume the agent's policy π to be deterministic and the environment model to be action positive. Nevertheless, we assume the given environment model is *accurate* in the sense that it faithfully represents the environment in the ways relevant to the agent. In other words, we are interested in problems that arise during planning, not problems that arise due to poor modeling.

3.2 Time Consistency

When planning for the infinite future we need to make sure that utilities do not sum to infinity; typically this is achieved with discounting. Here, we simplify by fixing a finite $m \in \mathbb{N}$ to be the agent's *lifetime*: the agent cares about the sum of the utilities of all percepts $e_1 \dots e_m$ until and including time step m, but does not care what happens after that (presumably the agent is then retired).

In sequential decision theory we need to plan the next $m - t$ actions in time step t. We plan what we would do for all possible future percepts $e_{t:m}$ by choosing a policy $\pi : (\mathcal{A}\times\mathcal{E})^* \to \mathcal{A}$ that specifies which action we take depending on how the history plays out. For example, we take action a_t, and when we subsequently receive the percept e_t, we plan to take action a_{t+1}. Problems arise once we get to the next step and even tough we *did* take action a_t and the percept *did* turn out to be e_t, we change our mind and take a different action $\hat{a}_{t+1}$. This is called *time inconsistency*. Time inconsistency is an artifact of bad planning since the agent incorrectly anticipates her own actions. The choice of discounting can lead to time inconsistency: a sliding fixed-size horizon is time inconsistent, but a fixed finite lifetime is time consistent [LH14].

We achieve time consistency by using a fixed finite lifetime, and by calculating decisions recursively using value functions. A *value function* $V^\pi_{\mu,m}$ is a function of type $((\mathcal{A}\times\mathcal{E})^* \cup ((\mathcal{A}\times\mathcal{E})^* \times \mathcal{A})) \to \mathbb{R}$. It gives an estimate of future reward: $V^\pi_{\mu,m}(æ_{<t})$ and $V^\pi_{\mu,m}(æ_{<t}a_t)$ are estimates of how much reward the policy π will obtain in environment μ within lifetime m subsequent to history $æ_{<t}$ and $æ_{<t}a_t$ respectively. For any history $æ_{<t}$, we define $V^\pi_{\mu,m}(æ_{<t}) := V^\pi_{\mu,m}(æ_{<t}\pi(æ_{<t}))$. We say that a policy π is *optimal and time consistent for the value function* $V_{\mu,m}$ iff $\pi(æ_{<t}) = \arg\max_a V^\pi_{\mu,m}(æ_{<t}a)$ for all histories $æ_{<t} \in (\mathcal{A} \times \mathcal{E})^{t-1}$ and all $t \leq m$.

3.3 Sequential Evidential Decision Theory

Evidential decision theory assigns probability $P(e \mid a)$ to action a resulting in percept e (Section 2.2). There are two ways to generalize this to the sequential setting, depending on whether we use only the next action or the whole future policy as evidence for the next percept.

Definition 3 (Action-Evidential Decision Theory)**.** The *action-evidential value of a policy π with lifetime m in environment μ given history $æ_{<t}a_t$* is

$$V_{\mu,m}^{\text{aev},\pi}(æ_{<t}a_t) := \sum_{e_t} \mu(e_t \mid æ_{<t}a_t)\Big(u(e_t) + V_{\mu,m}^{\text{aev},\pi}(æ_{<t}a_te_t)\Big) \qquad \text{(SAEDT)}$$

and $V_{\mu,m}^{\text{aev},\pi}(æ_{<t}a_t) := 0$ for $t > m$. *Sequential Action-Evidential Decision Theory (SAEDT)* prescribes adopting an optimal and time consistent policy π for $V_{\mu,m}^{\text{aev}}$.

It may be argued that SAEDT does not take all available (deliberative) information into account. When considering the consequences of an action, future developments of the environment-policy interactions could also be used as evidence. That is, we could condition not only on the next action, but on the future policy as a whole (within the lifetime). In order to define conditional probabilities with respect to (deterministic) policies, we define the following events. For a given policy π, let $\Pi_{t:m}$ be the set of all strings consistent with π between time step t and m:

$$\Pi_{t:m} := \{æ_{1:\infty} \mid \forall t \leq i \leq m.\ \pi(æ_{<i}) = a_i\}$$

The likelihood of a next percept e_t provided a history $æ_{<t}$ and a (future) policy π followed from time step t until lifetime m (denoted $\pi_{t:m}$) is then defined as

$$\mu(e_t \mid æ_{<t}, \pi_{t:m}) := \mu(e_t \mid æ_{<t} \cap \Pi_{t:m}). \tag{4}$$

This is an *atemporal* conditional because we are conditioning on future actions up until the end of the agent's lifetime. The conditional (4) is well-defined because we only take the actions from time step t to m into account; conditioning on policies with infinite lifetime leads to technical problems because such policies typically have μ-measure zero.

Definition 4 (Policy-Evidential Decision Theory)**.** The *policy-evidential value of a policy π with lifetime m in environment μ given history $æ_{<t}a_t$* is

$$V_{\mu,m}^{\text{pev},\pi}(æ_{<t}a_t) := \sum_{e_t} \mu(e_t \mid æ_{<t}a_t, \pi_{t+1:m}) \cdot \Big(u(e_t) + V_{\mu,m}^{\text{pev},\pi}(æ_{<t}a_te_t)\Big) \qquad \text{(SPEDT)}$$

and $V_{\mu,m}^{\text{pev},\pi}(æ_{<t}) := 0$ for $t > m$. *Sequential Policy-Evidential Decision Theory (SPEDT)* prescribes adopting an optimal and time consistent policy π for $V_{\mu,m}^{\text{pev}}$.

For one-step decisions ($m = t + 1$), SAEDT and SPEDT coincide.

To all our embedded agents, past actions constitute evidence about the hidden state. For evidential agents, this principle is extended to future actions. SAEDT and SPEDT differ in how far they extend it. The action-evidential agent only updates his belief on the action about to take place. In that sense, he only updates his belief about the next percept on events taking place *before* this percept. The policy-evidential agent takes the principle much further, using "thought-experiments" of what action he *would take in hypothetical situations*, most of which will never be realized. This is illustrated in the next example.

Example 5 (Sequential Toxoplasmosis)**.** In our sequential variation of the toxoplasmosis problem the agent has some probability of encountering a kitten.

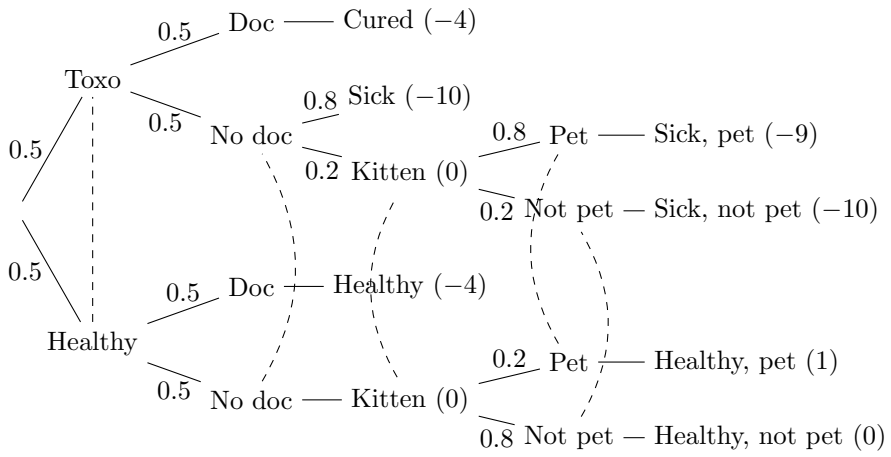


Figure 4: One formalization of the sequential toxoplasmosis problem. Dashed lines connect states indistinguishable to the agent. The numbers on the edges indicate probabilities of the environment model μ, and the numbers in parenthesis indicate utilities of the associated percepts. In the first step, the environment selects the hidden state that is unknown to the agent. The agent then decides whether to go to the doctor. If he does not go, he may encounter a kitten which he can choose to pet or not. SAEDT and SPEDT will disagree whether going to the doctor is the best option in this scenario. Appendix A contains the full calculations.

Additionally, the agent has the option of seeing a doctor (for a fee) and getting tested for the parasite, which can then be safely removed. In the very beginning, an SPEDT agent updates his belief on the fact that if he encountered a kitten, he would not pet it, which lowers the probability that he has the parasite and makes seeing the doctor unattractive. An SAEDT agent only updates his belief about the parasite when he actually encounters a kitten, and thus prefers seeing the doctor. See Figure 4 for more details and a graphical illustration.

The observant reader may ask whether SPEDT could be enticed to make some percepts unlikely by choosing improbable actions subsequent to them. For example, could an SPEDT agent decide on a policy of selecting highly improbable actions in case it rained to make histories with rain less likely? The answer is no, as most such policies would not be time consistent. If it does rain, the highly improbable action would usually not the best one, and so the policy would not be prescribed by Definition 4.

3.4 Sequential Causal Decision Theory

In sequential causal decision theory we ask what would happen if we causally intervened on the node a_t of the next action and fix it to $\pi(æ_{<t})$ according to the policy π. This is expressed by the notation $\texttt{do}(a_t := \pi(æ_{<t}))$, or $\texttt{do}(\pi(æ_{<t}))$ for short.

Definition 6 (Sequential Causal Decision Theory)**.** The *causal value of a policy*

π with lifetime m in environment μ given history $æ_{<t}a_t$ is

$$V^{\text{cau},\pi}_{\mu,m}(æ_{<t}a_t) := \sum_{e_t \in \mathcal{E}} \mu(e_t \mid æ_{<t}, \texttt{do}(a_t))\Big(u(e_t) + V^{\text{cau},\pi}_{\mu,m}(æ_{<t}a_t e_t)\Big) \quad \text{(SCDT)}$$

and $V^{\text{cau},\pi}_{\mu,m}(æ_{<t}a_t) := 0$ for $t > m$. *Sequential Causal Decision Theory (SCDT)* prescribes adopting an optimal and time consistent policy π for $V^{\text{cau}}_{\mu,m}$.

For sequential evidential decision theory we discussed two versions (SAEDT) and (SPEDT), based on next action and future policy respectively. In SCDT we perform the causal intervention $\texttt{do}(a_t := \pi(æ_{<t}))$. We could also consider a policy-causal decision theory by replacing $\mu(e_t \mid æ_{<t}, \texttt{do}(a_t))$ with $\mu(e_t \mid æ_{<t}, \texttt{do}(\pi_{t:m}))$ in Definition 6. The causal intervention $\texttt{do}(\pi_{t:m}))$ of a policy π between time step t and time step m is defined as as

$$\mu(e_t \mid æ_{<t}, \texttt{do}(\pi_{t:m})) := \sum_{e_{t+1:m}} \mu(e_{t:m} \mid æ_{<t}, \texttt{do}(a_t := \pi(æ_{<t}), \ldots, a_m := \pi(æ_{<m}))). \tag{5}$$

However, since the interventions are causal, we do not get any extra evidence from the future interventions. Therefore policy-causal decision theory is the same as action-causal decision theory:

Proposition 7 (Policy-Causal = Action-Causal)**.** *For all histories $æ_{<t} \in (\mathcal{A} \times \mathcal{E})^*$ and all $e_t \in \mathcal{E}$, we have $\mu(e_t \mid æ_{<t}, \texttt{do}(\pi_{t:m})) = \mu(e_t \mid æ_{<t}, \texttt{do}(\pi(æ_{<t})))$.*

We defer the proof to the end of this section. The following two examples illustrate the difference between SCDT and SAEDT/SPEDT in sequential settings.

Example 8 (Newcomb with Precommitment)**.** In this variation to Newcomb's problem the agent first has the option to pay \$300,000 to sign a contract that binds the agent to pay \$2000 in case of two-boxing. An SAEDT or SPEDT agent knows that he will one-box anyways and hence has no need for the contract. An SCDT agent knows that she favors two-boxing, but signs the contract only if this occurs before the prediction is made (so it has a chance of causally affecting the prediction). With the contract in place, one-boxing is the dominant action, and thus the SCDT agent is predicted to one-box.

Example 9 (Newcomb with Looking)**.** In this variation to Newcomb's problem the agent may look into the opaque box before making the decision which box to take. An SCDT agent is indifferent towards looking because she will take both boxes anyways. However, an SAEDT or SPEDT agent will avoid looking into the box, because once the content is revealed he two-boxes.

3.5 Expansion over the Hidden State

The difference between sequential versions of EDT and CDT is how they update their prediction of a next percept e_t (Definitions 3, 4 and 6). The following proposition expands the different beliefs in terms of the hidden state.

Proposition 10. *For all histories $æ_{<t}a_te_t \in (\mathcal{A} \times \mathcal{E})^*$ the following holds for the next-percept beliefs of SAEDT, SPEDT and SCDT respectively:*

$$\mu(e_t \mid æ_{<t}a_t) = \sum_{s \in \mathcal{S}} \mu(s \mid æ_{<t}a_t)\mu(e_t \mid s, æ_{<t}a_t) \tag{6}$$

$$\mu(e_t \mid æ_{<t}, \pi_{t:m}) = \sum_{s \in \mathcal{S}} \mu(s \mid æ_{<t}, \pi_{t:m})\mu(e_t \mid s, æ_{<t}, \pi_{t:m}) \tag{7}$$

$$\mu(e_t \mid æ_{<t}, \texttt{do}(a_t)) = \sum_{s \in \mathcal{S}} \mu(s \mid æ_{<t})\mu(e_t \mid s, æ_{<t}a_t) \tag{8}$$

Proof. For the action-evidential conditional we take the joint distribution with s, and then split off e_t:

$$\begin{aligned}\mu(e_t \mid æ_{<t}a_t) = \frac{\sum_{s \in \mathcal{S}} \mu(s, æ_{<t}a_te_t)}{\mu(æ_{<t}a_t)} &= \frac{\sum_{s \in \mathcal{S}} \mu(s, æ_{<t}a_t)\mu(e_t \mid s, æ_{<t}a_t)}{\mu(æ_{<t}a_t)} \\ &= \sum_{s \in \mathcal{S}} \mu(s \mid æ_{<t}a_t)\mu(e_t \mid s, æ_{<t}a_t)\end{aligned}$$

Similarly for the policy-evidential conditional:

$$\begin{aligned}\mu(e_t \mid æ_{<t}, \pi_{t:m}) &= \frac{\sum_{s \in \mathcal{S}} \mu(s, æ_{<t}\pi(æ_{<t})e_t, \pi_{t+1:m})}{\mu(æ_{<t}, \pi_{t:m})} \\ &= \frac{\sum_{s \in \mathcal{S}} \mu(s, æ_{<t}\pi(æ_{<t}), \pi_{t+1:m})\mu(e_t \mid s, æ_{<t}\pi(æ_{<t}), \pi_{t+1:m})}{\mu(æ_{<t}, \pi_{t:m})} \\ &= \frac{\sum_{s \in \mathcal{S}} \mu(s, æ_{<t}, \pi_{t:m})\mu(e_t \mid s, æ_{<t}\pi(æ_{<t}), \pi_{t+1:m})}{\mu(æ_{<t}, \pi_{t:m})} \\ &= \sum_{s \in \mathcal{S}} \mu(s \mid æ_{<t}, \pi_{t:m})\mu(e_t \mid s, æ_{<t}\pi(æ_{<t}), \pi_{t+1:m}) \\ &= \sum_{s \in \mathcal{S}} \mu(s \mid æ_{<t}, \pi_{t:m})\mu(e_t \mid s, æ_{<t}, \pi_{t:m})\end{aligned}$$

For the causal conditional we turn to the rules of the do-operator [Pea09, Thm. 3.4.1]. The first equality below holds by definition. In the denominator of the second equality we can use Rule 3 (deletion of actions) to remove $\texttt{do}(a_t)$ because the do-operator removes all incoming edges to a_t and makes a_t independent of the history $æ_{<t}$. In the numerator of the second equality we use the definition of do (1):

$$\begin{aligned}\mu(e_t \mid æ_{<t}, \texttt{do}(a_t)) &= \frac{\mu(æ_{<t}, e_t \mid \texttt{do}(a_t))}{\mu(æ_{<t} \mid \texttt{do}(a_t))} \\ &= \frac{\sum_{s \in \mathcal{S}} \mu(s, æ_{<t})\mu(e_t \mid s, æ_{<t}a_t)}{\mu(æ_{<t})} \\ &= \sum_{s \in \mathcal{S}} \mu(s \mid æ_{<t})\mu(e_t \mid s, æ_{<t}a_t)\end{aligned}$$

□

Proposition 10 shows that between SCDT and SAEDT, the difference in opinion about e_t only depends on differences in their (acausal) *posterior belief* $\mu(s \mid \ldots)$ about the hidden state. SCDT and SAEDT thus become equivalent in scenarios where there is only one hidden state s^* with $\mu(s^*) = 1$, as this renders

$\mu(s^* \mid æ_{<t}) = \mu(s^* \mid æ_{<t}a_t) = \mu(s^*) = 1$. SPEDT, on the other hand, may disagree with the other two also after a hidden state has been fixed.

From a problem modeler's perspective, it is also instructive to consider the effect of moving uncertainty between the hidden state and environmental stochasticity. For two different environment models μ and μ', the action and percept probabilities may be identical (i.e., $\mu(a_t \mid æ_{<t}) = \mu'(a_t \mid æ_{<t})$ and $\mu(e_t \mid æ_{<t}a_t) = \mu'(e_t \mid æ_{<t}a_t)$) even though μ and μ' have non-isomorphic sets of hidden states $\mathcal{S}$ and $\mathcal{S}'$. For example, given any μ, an environment model μ' with a single hidden state s_0, $\mu'(s_0) = 1$, may be constructed from μ by $\mu'(s_0, æ_{<t}) := \sum_{s\in\mathcal{S}} \mu(s, æ_{<t})$. The transformation will not affect SAEDT and SPEDT, as the definitions of their value functions only depends on the 'observable' action- and percept-probabilities $\mu(a_t \mid æ_{<t})$ and $\mu(e_t \mid æ_{<t}a_t)$ which are preserved between μ and μ'. But the transformation will change SCDT's behavior in any μ where SCDT disagrees with SAEDT, as SCDT and SAEDT are equivalent in μ' that only has a single hidden state. That SCDT depends on what uncertainty is captured by the hidden state is unsurprising given that the hidden state has a special place in the causal structure of the problem. Ultimately, the modeler must decide what uncertainty to put in the hidden state, and what to attribute to environmental stochasticity. A general principle for how to do this is still an open question [SF14b].

The value functions of SAEDT, SPEDT and SCDT can be rewritten in the following *iterative forms*, where the latter form uses Proposition 10. Numbers above equality signs reference a justifying equation. Let $a_i := \pi(æ_{<i})$ for $i \geq t$:

$$V^{\text{aev},\pi}_{\mu,m}(æ_{<t}) = \sum_{k=t}^{m} \sum_{e_{t:k}} u(e_k) \prod_{i=t}^{k} \mu(e_i \mid æ_{<i}a_i) \tag{9}$$

$$\overset{(6)}{=} \sum_{k=t}^{m} \sum_{e_{t:k}} u(e_k) \prod_{i=t}^{k} \sum_{s\in\mathcal{S}} \mu(s \mid æ_{<i}a_i)\mu(e_i \mid s, æ_{<i}a_i) \tag{10}$$

$$V^{\text{pev},\pi}_{\mu,m}(æ_{<t}) = \sum_{k=t}^{m} \sum_{e_{t:k}} u(e_k) \prod_{i=t}^{k} \mu(e_i \mid æ_{<i}, \pi_{i:m}) \tag{11}$$

$$\overset{(7)}{=} \sum_{k=t}^{m} \sum_{e_{t:k}} u(e_k) \prod_{i=t}^{k} \sum_{s\in\mathcal{S}} \mu(s \mid æ_{<i}\pi_{i:m})\mu(e_i \mid s, æ_{<i}, \pi_{i:m}) \tag{12}$$

$$V^{\text{cau},\pi}_{\mu,m}(æ_{<t}) = \sum_{k=t}^{m} \sum_{e_{t:k}} u(e_k) \prod_{i=t}^{k} \mu(e_i \mid æ_{<i}, \texttt{do}(a_i)) \tag{13}$$

$$\overset{(8)}{=} \sum_{k=t}^{m} \sum_{e_{t:k}} u(e_i) \prod_{i=t}^{k} \sum_{s\in\mathcal{S}} \mu(s \mid æ_{<i})\mu(e_i \mid s, æ_{<i}a_i) \tag{14}$$

	SAEDT	SPEDT	SCDT
Nwcb	*1-box*	*1-box*	2-box
Nwcb w/ precommit	*not commit, 1-box*	*not commit, 1-box*	commit, 1-box
Nwcb w/ looking	not look, 1-box	not look, 1-box	indifferent, 2-box
Toxoplasmosis	not pet	not pet	*pet*
Seq. Toxoplasmosis	doc, not pet	no doc, not pet	*doc, pet*

Table 1: Decisions made by SAEDT, SPEDT and SCDT in Example 1, Example 2, Example 5, Example 8, and Example 9. The latter three examples are sequential. Winning moves are in italics; in Newcomb with looking the winning move is to be indifferent and one-box. Because Savage decision theory is dualistic, these problems cannot be properly formalized in it.

Proof of Proposition 7. By the definition (5) of $\mathtt{do}(\pi_{t:m})$,

$$\begin{aligned}
\mu(e_t \mid æ_{<t}, \mathtt{do}(\pi_{t:m})) &= \sum_{e_{t+1:m}} \mu(e_{t:m} \mid æ_{<t}, \mathtt{do}(a_t := \pi(æ_{<t}), \ldots, a_m := \pi(æ_{<m}))) \\
&= \sum_{s, e_{t+1:m}} \mu(s \mid æ_{<t}) \mu(e_{t:m} \mid s, æ_{<t}, \mathtt{do}(\pi(æ_{<t}), \ldots, \pi(æ_{<m}))) \\
&\overset{(1)}{=} \sum_{s, e_{t+1:m}} \mu(s \mid æ_{<t}) \prod_{i=t}^{m} \mu(e_i \mid s, æ_{<i} \pi(æ_{<i})) \\
&= \sum_{s} \mu(s \mid æ_{<t}) \mu(e_t \mid s, æ_{<t} \pi(æ_{<t})) \\
&\overset{(8)}{=} \mu(e_t \mid æ_{<t}, \mathtt{do}(\pi(æ_{<t})))
\end{aligned}$$

The second equality follows from the equivalence $P(\,\cdot\,) = \sum_s P(s) P(\,\cdot \mid s)$ applied to the distribution $\mu(\,\cdot \mid æ_{<t}, \mathtt{do}(a_t := \pi(æ_{<t}), \ldots, a_m := \pi(æ_{<m})))$, and the third equality by (repeated) application of (1) to $\mu(æ_{t:m} \mid s, æ_{<t}) = \prod_{i=t}^{m} \mu(a_i \mid s, æ_{<i}) \mu(e_i \mid s, æ_{<i} a_i)$. □

4 Discussion

Our paper is a first stab at the problem of how physicalistic agents should make sequential decisions. CDT and EDT provide an existing basis for non-dualistic decision making, which we extended to the sequential setting. There are two natural ways for making sequential evidential decisions: do I update my beliefs about the hidden state based on my next action ('what I do next', SAEDT) or my whole policy ('the kind of agent I am', SPEDT)? By Proposition 7, this distinction does not exist for causal decision theory, because with that theory the agent does not consider its own actions evidence at all. Therefore we have only one version of sequential causal decision theory, SCDT.

To illustrate the differences between the decision theories, we discussed three variants of Newcomb's problem (Example 1, Example 8, and Example 9) and two variants of the toxoplasmosis problem (Example 2 and Example 5). The formal specification of these examples can be found in Appendix A. We implemented

SCDT, SAEDT, and SPEDT; Table 1 shows their behavior on those examples.[4]

So which decision theory is better? The answer to this question depends on which decision you consider to be *correct* (or even *rational*) in each of the problems. We posit that ultimately, what counts is not whether your decision algorithm is theoretically pleasing, but *whether you win*. Winning means getting the most utility. If maximizing utility involves making crazy decisions, then this is what you should do!

In Newcomb's problem, winning means one-boxing, because you end up richer. In the toxoplasmosis problem, winning means petting the kitten, because that yields more utility. (S)CDT performs suboptimally in the Newcomb variations, while the evidential decision theories perform suboptimally in the toxoplasmosis variations. This entails that neither CDT nor EDT are the final answer to the problem of non-dualistic decision making.

Furthermore, neither CDT nor EDT agents are fully physicalistic: they do not model the environment to contain themselves [SF14b]. For example, when playing a prisoner's dilemma against your own source code [SF15], your opponent defects if and only if you defect. This *logical* connection between your action and your opponent's is disregarded in the formalization based on causal graphical models that we discuss here because it is not causal.

Timeless decision theory [Yud10] and *updateless decision theory* [SF14b] are recent attempts of more physicalistic decision theories. However, so far both have eluded explicit formalization [SF15]. We conclude that finding a physicalistic decision theory remains an important open problem in artificial intelligence research.

Acknowledgements. This work was in part supported by ARC grant DP120100950. It started at a MIRIxCanberra workshop sponsored by the Machine Intelligence Research Institute. Mayank Daswani and Daniel Filan contributed in the early stages of this paper and we thank them for interesting discussions and helpful suggestions. We also thank Nate Soares for useful feedback.

References

[Ahm14] Arif Ahmed. *Evidence, Decision and Causality*. Cambridge University Press, 2014.

[Alt13] Alex Altair. A comparison of decision algorithms on Newcomblike problems. Technical report, Machine Intelligence Research Institute, 2013. `http://intelligence.org/files/Comparison.pdf`.

[Bos14] Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014.

[Bri14] Rachael Briggs. Normative theories of rational choice: Expected utility. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Fall 2014 edition, 2014.

[4] Source code available at `http://jan.leike.name/`.

[Ega07] Andy Egan. Some counterexamples to causal decision theory. *The Philosophical Review*, pages 93–114, 2007.

[GH78] Allan Gibbard and William L Harper. Counterfactuals and two kinds of expected utility. In *Foundations and Applications of Decision Theory*, pages 125–162. Springer, 1978.

[Jef83] Richard C Jeffrey. *The Logic of Decision*. University of Chicago Press, 2nd edition, 1983.

[Joy99] James M Joyce. *The Foundations of Causal Decision Theory*. Cambridge University Press, 1999.

[Lew81] David Lewis. Causal decision theory. *Australasian Journal of Philosophy*, 59(1):5–30, 1981.

[LFY+14] Patrick LaVictoire, Benja Fallenstein, Eliezer Yudkowsky, Mihaly Barasz, Paul Christiano, and Marcello Herreshoff. Program equilibrium in the prisoner's dilemma via Löb's theorem. In *AAAI Workshop on Multiagent Interaction without Prior Coordination*, 2014.

[LH14] Tor Lattimore and Marcus Hutter. General time consistent discounting. *Theoretical Computer Science*, 519:140–154, 2014.

[Noz69] Robert Nozick. Newcomb's problem and two principles of choice. In *Essays in honor of Carl G. Hempel*, pages 114–146. Springer, 1969.

[OR12] Laurent Orseau and Mark Ring. Space-time embedded intelligence. In *Artificial General Intelligence*, pages 209–218. Springer, 2012.

[Pea09] Judea Pearl. *Causality*. Cambridge University Press, 2nd edition, 2009.

[RDT+15] Stuart Russell, Daniel Dewey, Max Tegmark, Janos Kramar, and Richard Mallah. Research priorities for robust and beneficial artificial intelligence. Technical report, Future of Life Institute, 2015. `http://futureoflife.org/static/data/documents/research_priorities.pdf`.

[RN10] Stuart J Russell and Peter Norvig. *Artificial Intelligence. A Modern Approach*. Prentice Hall, 3rd edition, 2010.

[Sav72] Leonard J Savage. *The Foundations of Statistics*. Dover Publications, 1972.

[SF14a] Nate Soares and Benja Fallenstein. Aligning superintelligence with human interests: A technical research agenda. Technical report, Machine Intelligence Research Institute, 2014. `http://intelligence.org/files/TechnicalAgenda.pdf`.

[SF14b] Nate Soares and Benja Fallenstein. Toward idealized decision theory. Technical report, Machine Intelligence Research Institute, 2014. `http://intelligence.org/files/TowardIdealizedDecisionTheory.pdf`.

[SF15] Nate Soares and Benja Fallenstein. Counterpossibles as necessary for decision theory. In *Artificial General Intelligence*. Springer, 2015.

[Sky82] Brian Skyrms. Causal decision theory. *The Journal of Philosophy*, pages 695–711, 1982.

[Wei12] Paul Weirich. Causal decision theory. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Winter 2012 edition, 2012.

[Yud08] Eliezer Yudkowsky. Artificial intelligence as a positive and negative factor in global risk. In Nick Bostrom and Milan M Ćirković, editors, *Global Catastrophic Risks*, pages 308–345. Oxford University Press, 2008.

[Yud10] Eliezer Yudkowsky. Timeless decision theory. Technical report, Machine Intelligence Research Institute, 2010. `http://intelligence.org/files/TDT.pdf`.

List of Notation

$:=$	defined to be equal
$\mathbb{N}$	the natural numbers, starting with 0
$\mathbb{R}$	the real numbers
ε	a small positive real number
$\mathcal{A}$	the (finite) set of possible actions
$\mathcal{E}$	the (finite) set of possible percepts
$\mathcal{S}$	the set of hidden states
u	the utility function $u : \mathcal{E} \to [0, 1]$
a_t	the action in time step t
e_t	the percept in time step t
$æ_{<t}$	the first $t-1$ interactions, $a_1 e_1 a_2 e_2 \dots a_{t-1} e_{t-1}$
$æ_{i:k}$	the interactions between and including time step i and time step k, $a_i e_i a_{i+1} e_{i+1} \dots a_k e_k$
$æ_{1:\infty}$	a history of infinite length
s	a hidden state
π	a deterministic policy, i.e., a function $\pi : (\mathcal{A} \times \mathcal{E})^* \to \mathcal{A}$
$\pi_{t:k}$	policy π restricted to the time steps between and including t and k
$V_{\mu,m}^{\text{aev},\pi}$	action-evidential value of policy π in environment μ up to time step m, defined in (SAEDT)
$V_{\mu,m}^{\text{pev},\pi}$	policy-evidential value of policy π in environment μ up to time step m, defined in (SPEDT)
$V_{\mu,m}^{\text{cau},\pi}$	causal value of policy π in environment μ up to time step m, defined in (SCDT)
k, i	time steps, natural numbers
t	(current) time step
m	lifetime of the agent
P_a	distribution over percepts induced by action a in SDT
P	distribution over percepts and actions in one-shot decision making
μ	an accurate environment model

A Examples

This section contains the formal calculations for Example 1, Example 2, Example 5, Example 8, and Example 9. These calculations are also available as Python code at `http://jan.leike.name/`.

Example 11 (Newcomb's Problem)**.** This is a formalization of Example 1.

- $\mathcal{S} := \{E, F\}$ where E means the opaque box is empty and F means the opaque box is full
- $\mathcal{A} := \{B_1, B_2\}$ where B_1 means one-boxing and B_2 means two-boxing
- $\mathcal{E} := \{O_0, O_T, O_M, O_{MT}\}$
- $u(O_0) := 0$, $u(O_T) := 1{,}000$, $u(O_M) := 1{,}000{,}000$, $u(O_{MT}) := 1{,}001{,}000$

Let $\varepsilon > 0$ be a small constant denoting the accuracy of the predictor. Because the environment has to assign non-zero probability to all actions, ε must be strictly positive. The environment's distribution μ is defined as follows.

$$\begin{aligned} \mu(E) = \mu(F) &= 0.5 & \mu(O_T \mid E, B_2) &= 1 \\ \mu(B_1 \mid F) = \mu(B_2 \mid E) &= 1-\varepsilon & \mu(O_0 \mid E, B_1) &= 1 \\ \mu(B_1 \mid E) = \mu(B_2 \mid F) &= \varepsilon & \mu(O_{MT} \mid F, B_2) &= 1 \\ & & \mu(O_M \mid F, B_1) &= 1 \end{aligned}$$

By Bayes' rule,

$$\mu(F \mid B_1) = \frac{\mu(B_1 \mid F)\mu(F)}{\sum_{s\in\mathcal{S}} \mu(B_1 \mid s)\mu(s)} = \frac{\frac{1}{2}(1-\varepsilon)}{\frac{1}{2}(1-\varepsilon) + \frac{1}{2}\varepsilon} = (1-\varepsilon)$$

which also gives $\mu(E \mid B_1) = \varepsilon$. Similarly, $\mu(F \mid B_2) = \varepsilon$ and $\mu(E \mid B_2) = 1-\varepsilon$.

For EDT we use equation (EDT) to compute the value of an action. Since the percept e_1 is generated deterministically, $\mu(e \mid s, a)$ only attains values 0 or 1. We therefore omit it in the calculation below. For action B_1 we get

$$\begin{aligned} V_{\mu,1}^{\text{evi},B_1} := \sum_{e\in\mathcal{E}} \mu(e \mid B_1)u(e) &= \sum_{e\in\mathcal{E}}\sum_{s\in\mathcal{S}} \mu(e \mid s, B_1)\mu(s \mid B_1)u(e) \\ &= \mu(E \mid B_1)u(O_0) + \mu(F \mid B_1)u(O_M) \\ &= \varepsilon\cdot 0 + (1-\varepsilon)\cdot 1{,}000{,}000 \end{aligned}$$

For action B_2 we get

$$\begin{aligned} V_{\mu,1}^{\text{evi},B_2} := \sum_{e\in\mathcal{E}} \mu(e \mid B_2)u(e) &= \sum_{e\in\mathcal{E}}\sum_{s\in\mathcal{S}} \mu(e \mid s, B_2)\mu(s \mid B_2)u(e) \\ &= \mu(E \mid B_2)u(O_T) + \mu(F \mid B_2)u(O_{MT}) \\ &= (1-\varepsilon)\cdot 1{,}000 + \varepsilon\cdot 1{,}001{,}000 \\ &= 1{,}000 + \varepsilon\cdot 1{,}000{,}000 \end{aligned}$$

For $\varepsilon < 49.95$ (just slightly better than random guessing), we get that EDT favors B_1 over B_2:

$$V_{\mu,1}^{\text{evi},B_1} = (1-\varepsilon)\cdot 1{,}000{,}000 > 500{,}500 > 1{,}000 + \varepsilon\cdot 1{,}000{,}000 = V_{\mu,1}^{\text{evi},B_2}$$

For CDT we use equation (CDT) to compute the value of an action. For action B_1 we get

$$\begin{aligned} V_{\mu,1}^{\text{cau},B_1} := \sum_{e\in\mathcal{E}} \mu(e \mid \texttt{do}(B_1))u(e) &= \sum_{e\in\mathcal{E}}\sum_{s\in\mathcal{S}} \mu(e \mid s, B_1)\mu(s)u(e) \\ &= \mu(E)u(O_0) + \mu(F)u(O_M) \\ &= 0.5\cdot 0 + 0.5\cdot 1{,}000{,}000 = 500{,}000 \end{aligned}$$

For action B_2 we get

$$\begin{aligned} V_{\mu,1}^{\text{cau},B_2} := \sum_{e\in\mathcal{E}} \mu(e \mid \texttt{do}(B_2))u(e) &= \sum_{e\in\mathcal{E}}\sum_{s\in\mathcal{S}} \mu(e \mid s, B_2)\mu(s)u(e) \\ &= \mu(E)u(O_T) + \mu(F)u(O_{MT}) \\ &= 0.5\cdot 1{,}000 + 0.5\cdot 1{,}001{,}000 = 500{,}500 \end{aligned}$$

We get that CDT favors B_2 over B_1 regardless of the prediction accuracy ε:

$$V_{\mu,1}^{\text{evi},B_1} = 500{,}000 < 500{,}500 = V_{\mu,1}^{\text{evi},B_2}$$

Moreover, CDT prefers B_2 *regardless of the prior over* $\mu(E)$. Two-boxing is the dominant action because it yields \$1,000 more regardless of the hidden state.

Example 12 (Newcomb with Looking)**.** This is a formalization of Example 9; it extends Example 11.

In the first time step, the agent gets to choose between looking into the box (L) and not looking (N). If the agent looks, the subsequent percept will be E or F, depending on whether the box is empty (E) or full (F). If the agent does not look, the subsequent percept will be 0. All three of these percepts E, F, and 0 have zero utility.

In the second time step the agent chooses to one-box (B_1) or to two-box (B_2). The payoffs are then based on the boxes' contents as in Example 11.

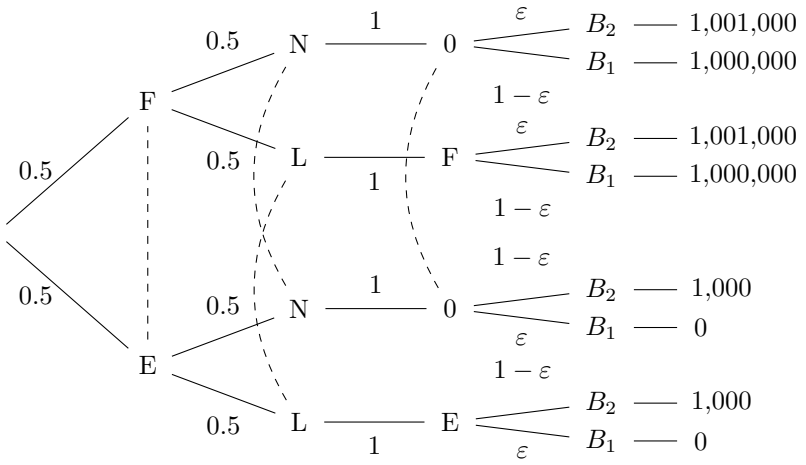
- $\mathcal{S} := \{E, F\}$ where E means the opaque box is empty and F means the opaque box is full
- $\mathcal{A} := \{B_1, B_2\}$ where B_1 means one-boxing and B_2 means two-boxing, $L := B_1$ means looking into the box and $N := B_2$ means not looking (the set of actions has to be the same for all time steps)
- $\mathcal{E} := \{E, F, 0, O_0, O_T, O_M, O_{MT}\}$
- $u(O_0) := 0$, $u(O_T) :=$ 1,000, $u(O_M) :=$ 1,000,000, $u(O_{MT}) :=$ 1,001,000, $u(E) := u(F) := u(0) := 0$

Let $\varepsilon > 0$ be a small constant denoting the prediction accuracy. Because the environment has to assign non-zero probability to all actions, ε must be strictly positive. The environment's distribution μ is defined as follows. Question marks

stand for single actions or percepts whose value is irrelevant.

$$\begin{aligned}
\mu(E) = \mu(F) &= 0.5 & \mu(E \mid E, L) &= 1 \\
\mu(L \mid F) = \mu(L \mid E) &= 0.5 & \mu(0 \mid E, N) &= 1 \\
\mu(N \mid F) = \mu(N \mid E) &= 0.5 & \mu(F \mid F, L) &= 1 \\
\mu(B_1 \mid E, ??) &= \varepsilon & \mu(0 \mid F, N) &= 1 \\
\mu(B_1 \mid F, ??) &= 1 - \varepsilon & \mu(O_0 \mid E, ??B_1) &= 1 \\
\mu(B_2 \mid E, ??) &= 1 - \varepsilon & \mu(O_T \mid E, ??B_2) &= 1 \\
\mu(B_2 \mid F, ??) &= \varepsilon & \mu(O_M \mid F, ??B_1) &= 1 \\
& & \mu(O_{MT} \mid F, ??B_2) &= 1
\end{aligned}$$

The environment's game tree is given as follows, where dashed lines connect states indistinguishable by the agent (also known as *information sets*):



Using Bayes' rule, we calculate the following conditional probabilities of the hidden state given a history a_1 or $a_1e_1a_2$:

$$\begin{aligned}
0.5 = \mu(E \mid L) = \mu(F \mid L) = \mu(E \mid N) &= \mu(F \mid N) \\
1 = \mu(E \mid LEB_1) = \mu(E \mid LEB_2) &= \mu(F \mid LFB_1) = \mu(F \mid LFB_1) \\
\varepsilon = \mu(E \mid N0B_1) &= \mu(F \mid N0B_2) \\
1 - \varepsilon = \mu(E \mid N0B_2) &= \mu(F \mid N0B_1)
\end{aligned}$$

Next, we write out the formula for SAEDT for a horizon of 2 based on (10). The first percept has no utility, which simplifies the equation.

$$V_{\mu,2}^{\text{aev},\pi} = \sum_{e_{1:2}} u(e_2) \left(\sum_{s \in \mathcal{S}} \mu(s \mid a_1) \mu(e_1 \mid s, a_1) \right) \left(\sum_{s \in \mathcal{S}} \mu(s \mid æ_1 a_2) \mu(e_2 \mid s, æ_1 a_2) \right)$$

where $a_1 = \pi(\epsilon)$ and $a_2 = \pi(æ_1)$. The formula for SPEDT for a horizon of 2 based on (12) is as follows.

$$V_{\mu,2}^{\text{pev},\pi} = \sum_{e_{1:2}} u(e_2) \frac{\sum_{s \in \mathcal{S}} \mu(sa_1e_1\pi(a_1e_1))}{\sum_{s \in \mathcal{S}} \sum_{e \in \mathcal{E}} \mu(sa_1e\pi(a_1e))} \sum_{s \in \mathcal{S}} \mu(s \mid æ_1\pi_2) \mu(e_2 \mid s, æ_1 a_2)$$

with $\pi_{1:2}$ and π_2 defined according to (4). The formula for SCDT for a horizon of 2 based on (14) is as follows.

$$V_{\mu,2}^{\text{cau},\pi} = \sum_{e_{1:2}} u(e_2) \left(\sum_{s \in \mathcal{S}} \mu(s)\mu(e_1 \mid s, a_1) \right) \left(\sum_{s \in \mathcal{S}} \mu(s \mid æ_1)\mu(e_2 \mid s, æ_1 a_2) \right)$$

where $a_1 = \pi(\epsilon)$ and $a_2 = \pi(æ_1)$.

There are six different possible policies:

- Look and always one-box (curious one-boxer)
- Look and always two-box (curious two-boxer)
- Don't look and one-box (incurious one-boxer)
- Don't look and two-box (incurious two-boxer)
- Look and one-box iff the box is empty (paradox-lover)
- Look and one-box iff the box full (fatalistic)

Using the formulas above we can calculate their value. We use $\varepsilon := 0.01$.

	$V_{\mu,2}^{\text{aev},\pi}$	$V_{\mu,2}^{\text{pev},\pi}$	$V_{\mu,2}^{\text{cau},\pi}$
Curious one-boxer	500,000	*990,000*	500,000
Curious two-boxer	501,000	11,000	*501,000*
Incurious one-boxer	*990,000*	*990,000*	500,000
Incurious two-boxer	11,000	11,000	*501,000*
Paradox-lover	500,500	500,500	500,500
Fatalistic	500,500	500,500	500,500

The highest values are displayed in italics. The incurious one-boxer has the highest action-evidential value. The curious one-boxer and the incurious one-boxer have the highest policy-evidential value. However, of these two policies only the incurious one-boxer is a time-consistent policy for SPEDT, because the agent wants to two-box after looking into the box:

$$\begin{aligned}
V_{\mu,1}^{\text{aev},B_1}(LF) &= V_{\mu,1}^{\text{pev},B_1}(LF) = 1,000,000 \\
V_{\mu,1}^{\text{aev},B_2}(LF) &= V_{\mu,1}^{\text{pev},B_2}(LF) = 1,001,000 \\
V_{\mu,1}^{\text{aev},B_1}(LE) &= V_{\mu,1}^{\text{pev},B_1}(LE) = 0 \\
V_{\mu,1}^{\text{aev},B_2}(LE) &= V_{\mu,1}^{\text{pev},B_2}(LE) = 1,000
\end{aligned}$$

The curious two-boxer and the incurious two-boxer have the highest causal value, and they are both time-consistent for SCDT.

Example 13 (Newcomb with Precommitment)**.** This is a formalization of Example 8, it extends Example 11.

In the first time step, the agent gets to choose between signing the contract (S) and not signing (N). If the agent signs, the subsequent percept will be C, which costs \$300,000, and the prediction will be updated to one-boxing. If the agent does not sign, the subsequent percept will be 0 with zero utility.

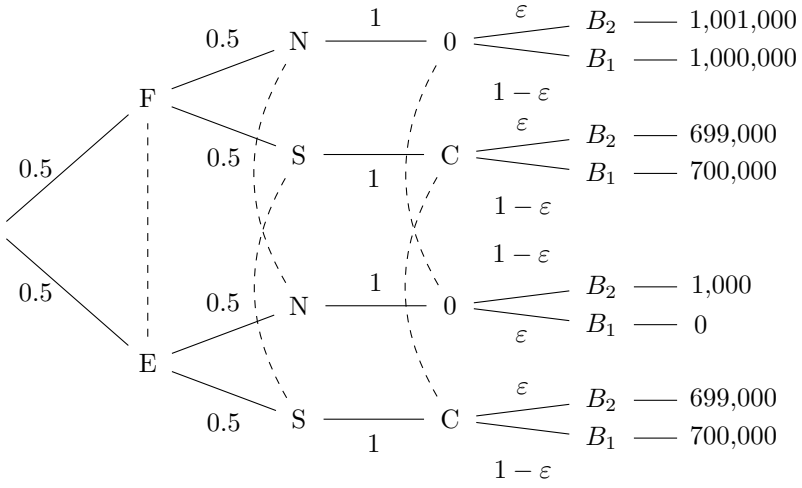
In the second time step the agent chooses to one-box (B_1) or to two-box (B_2). The payoffs are then based on the boxes' contents as in Example 11. If the agent signed the contract and choses two boxes, this incurs an additional cost of \$2,000.

- $\mathcal{S} := \{E, F\}$ where E means the opaque box is empty and F means the opaque box is full
- $\mathcal{A} := \{B_1, B_2\}$ where B_1 means one-boxing and B_2 means two-boxing, $S := B_1$ means signing the contract and $N := B_2$ means not signing (the set of actions has to be the same for all time steps)
- $\mathcal{E} := \{C, 0, O_0, O_T, O_{-T}, O_M, O_{MT}, O_{M-T}\}$
- $u(O_0) := 0$, $u(O_T) := 1,000$, $u(O_{-T}) := -1,000$ $u(O_M) := 1,000,000$, $u(O_{MT}) := 1,001,000$, $u(O_{M-T}) := 999,000$, $u(C) := -300,000$, $u(0) := 0$

Let $\varepsilon > 0$ be a small constant denoting the prediction accuracy. Because the environment has to assign non-zero probability to all actions, ε must be strictly positive. The environment's distribution μ is defined as follows. Question marks stand for single actions or percepts whose value is irrelevant.

$$\begin{aligned}
\mu(E) = \mu(F) &= 0.5 & \mu(C \mid E, S) &= 1 \\
\mu(S \mid F) = \mu(S \mid E) &= 0.5 & \mu(0 \mid E, N) &= 1 \\
\mu(N \mid F) = \mu(N \mid E) &= 0.5 & \mu(C \mid F, S) &= 1 \\
\mu(B_1 \mid E, N0) &= \varepsilon & \mu(0 \mid F, N) &= 1 \\
\mu(B_1 \mid F, N0) &= 1 - \varepsilon & \mu(O_0 \mid E, N0B_1) &= 1 \\
\mu(B_2 \mid E, N0) &= 1 - \varepsilon & \mu(O_T \mid E, N0B_2) &= 1 \\
\mu(B_2 \mid F, N0) &= \varepsilon & \mu(O_M \mid F, N0B_1) &= 1 \\
\mu(B_2 \mid ?, SC) &= \varepsilon & \mu(O_{MT} \mid F, N0B_2) &= 1 \\
\mu(B_1 \mid ?, SC) &= 1 - \varepsilon & \mu(O_M \mid E, SCB_1) &= 1 \\
& & \mu(O_{M-T} \mid E, SCB_2) &= 1
\end{aligned}$$

The environment's game tree is given as follows:



There are four different possible policies:

- Sign the contract and one-box (signing one-boxer)
- Sign the contract and two-box (signing two-boxer)
- Don't sign the contract and one-box (refusing one-boxer)
- Don't sign the contract and two-box (refusing two-boxer)

Using the formulas from Example 12 we can calculate their value. We use $\varepsilon := 0.01$.

	$V_{\mu,2}^{\text{aev},\pi}$	$V_{\mu,2}^{\text{pev},\pi}$	$V_{\mu,2}^{\text{cau},\pi}$
Signing one-boxer	700,000	700,00	*700,000*
Signing two-boxer	699,000	699,000	699,000
Refusing one-boxer	*990,000*	*990,000*	500,000
Refusing two-boxer	11,000	11,000	501,000

The highest values are displayed in italics. Both SAEDT and SPEDT refuse the contract: the refusing one-boxer has the highest action-evidential and the highest policy-evidential value. SCDT signs the contract and then one-boxes: the signing one-boxer has the highest causal value.

Example 14 (Toxoplasmosis)**.** This is a formalization of Example 2.

- $\mathcal{S} := \{T, H\}$ where T means having the toxoplasmosis parasite and H means being healthy
- $\mathcal{A} := \{P, N\}$ where P means petting and N means not petting
- $\mathcal{E} := \{P\&T, N\&T, P\&H, N\&H\}$ where the percepts just reflect the action and hidden state

- $u(P\&T) := -9$, $u(N\&T) := -10$, $u(P\&H) := 1$, $u(N\&H) := 0$ where petting gives a utility of 1 and suffering from the parasite gives a utility of -10

The environment's distribution μ is defined as follows.

$$\begin{aligned} \mu(T) = \mu(H) &= 0.5 & \mu(P\&T \mid P, T) &= 1 \\ \mu(P \mid T) &= 0.8 & \mu(N\&T \mid N, T) &= 1 \\ \mu(N \mid T) &= 0.2 & \mu(P\&H \mid P, H) &= 1 \\ \mu(P \mid H) &= 0.2 & \mu(N\&H \mid N, H) &= 1 \\ \mu(N \mid H) &= 0.8 \end{aligned}$$

Using Bayes' rule, we calculate the following conditional probabilities.

$$\mu(T \mid P) = 0.8 \qquad \mu(H \mid P) = 0.2 \qquad \mu(T \mid N) = 0.2 \qquad \mu(H \mid N) = 0.8$$

We consider EDT first. Since the percept e_1 is generated deterministically, $\mu(e \mid s, a)$ only attains values 0 or 1. We therefore omit it in the calculation below. For action P (petting) we get

$$\begin{aligned} V_{\mu,1}^{\text{evi},P} := \sum_{e \in \mathcal{E}} \mu(e \mid P) u(e) &= \sum_{e \in \mathcal{E}} \sum_{s \in \mathcal{S}} \mu(e \mid s, P) \mu(s \mid P) u(e) \\ &= \mu(T \mid P) u(T\&P) + \mu(H \mid P) u(P\&H) \\ &= 0.8 \cdot (-9) + 0.2 \cdot 1 = -7 \end{aligned}$$

For action N (not petting) we get

$$\begin{aligned} V_{\mu,1}^{\text{evi},N} := \sum_{e \in \mathcal{E}} \mu(e \mid N) u(e) &= \sum_{e \in \mathcal{E}} \sum_{s \in \mathcal{S}} \mu(e \mid s, N) \mu(s \mid N) u(e) \\ &= \mu(T \mid N) u(T\&N) + \mu(H \mid N) u(H\&N) \\ &= 0.2 \cdot (-10) + 0.8 \cdot 0 = -2 \end{aligned}$$

Therefore we get that EDT favors N over P:

$$V_{\mu,1}^{\text{evi},P} = -7 < -2 = V_{\mu,1}^{\text{evi},N}$$

For CDT we get for action P (petting)

$$\begin{aligned} V_{\mu,1}^{\text{cau},P} := \sum_{e \in \mathcal{E}} \mu(e \mid \texttt{do}(P)) u(e) &= \sum_{e \in \mathcal{E}} \sum_{s \in \mathcal{S}} \mu(e \mid s, P) \mu(s) u(e) \\ &= \mu(T) u(T\&P) + \mu(N) u(N\&P) \\ &= 0.5 \cdot (-9) + 0.5 \cdot 1 = -4 \end{aligned}$$

For action N (not petting) we get

$$\begin{aligned} V_{\mu,1}^{\text{cau},N} := \sum_{e \in \mathcal{E}} \mu(e \mid \texttt{do}(N)) u(e) &= \sum_{e \in \mathcal{E}} \sum_{s \in \mathcal{S}} \mu(e \mid s, N) \mu(s) u(e) \\ &= \mu(T) u(T\&N) + \mu(H) u(H\&N) \\ &= 0.5 \cdot (-10) + 0.5 \cdot 0 = -5 \end{aligned}$$

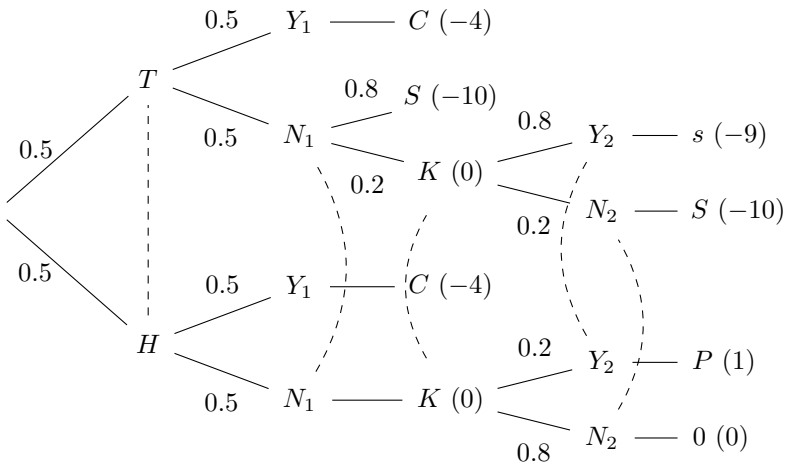
We get that CDT favors P over N:

$$V_{\mu,1}^{\text{evi},P} = -4 > -5 = V_{\mu,1}^{\text{evi},N}$$

Example 15 (Sequential Toxoplasmosis)**.** We here formalize a version of Example 5. First the agent chooses whether to go to the doctor. Going to the doctor incurs a fee, but removes the risk of getting sick. Agents that do not go to the doctor have a chance of meeting a kitten. If they meet it, they can choose to pet it or not; infected agents are more likely to pet the kitten. The example is intended to elucidate the difference between SAEDT and SPEDT, whose decisions we will calculate in detail. We will not calculate the action of SCDT.

- $\mathcal{S} := \{T(\text{oxoplasmosis}), H(\text{ealthy})\}$.
- $\mathcal{A} := \{Y(\text{es}), N(\text{o})\}$. In this example, an action is taken twice. We use Y_1 and Y_2, and N_1 and N_2, to distinguish between the first and the second action.
- $\mathcal{E} := \{C(\text{ured}), K(\text{itten}), S(\text{ick, not pet kitten}), s(\text{ick, pet kitten}), P(\text{et, not sick}), 0(\text{neutral})\}$
- $u(C) = -4$, $u(K) := 0$, $u(S) := -10$, $u(s) := -9$, $u(P) := 1$, and $u(0) = 0$.

The environment's game tree is given as follows, where dashed lines connect states indistinguishable by the agent.



First, the environment chooses whether to infect the agent or not with the parasite with probability 0.5. The agent then decides whether to see the doctor. If the agent sees the doctor, this incurs a (utility) fee of -4, but the agent will not be sick. If the agent does not see the doctor, there will be a kitten with probability 0.2 (or 1) and the agent will pet it with probability 0.8 (or 0.2) if the parasite is present (or not). If there is no kitten, the next percept is S or 0 depending on whether the agent is infected or not. The agent gets -10 utility if infected and did not see the doctor, and gets $+1$ utility for petting the kitten.

We want to compare the choices of SAEDT and SPEDT. Their two-step value functions are

$$V_{\mu,2}^{\text{aev},\pi} = \sum_{e_1} \mu(e_1 \mid a_1) \left(u(e_1) + V_{\mu,2}^{\text{aev},\pi}(a_1 e_1)\right)$$

$$V_{\mu,2}^{\text{pev},\pi} = \sum_{e_1} \mu(e_1 \mid \pi_{1:2}) \left(u(e_1) + V_{\mu,2}^{\text{pev},\pi}(a_1 e_1)\right)$$

where the second step value functions

$$V_{\mu,2}^{\text{aev},\pi}(a_1 e_1) = V_{\mu,2}^{\text{pev},\pi}(a_1 e_1) = \sum_{e_2} \mu(e_2 \mid a_1 e_1 a_2) \cdot u(e_2)$$

are the same for both decision theories. They only differ by assigning probability $\mu(e_1 \mid a_1)$ and $\mu(e_1 \mid \pi_{1:2})$ to the first percept, respectively.

Since not petting is always better than petting for evidential agents (the evidence towards not having the disease weighs stronger than the extra utility), the only policies that are potentially optimal and time consistent are $\pi_1 := N_1 N_2$ and $\pi_2 := Y_1$.

First percept. For π_1 the occurring action-evidential quantities $\mu(e_1 \mid a_1)$ are

$$\begin{aligned}
\mu(N_1) &= \sum_{s \in \mathcal{S}} \mu(s, N_1) = \mu(T, N_1) + \mu(H, N_1) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2} \\
\mu(e_1 = S \mid N_1) &= \frac{\sum_{s \in \mathcal{S}} \mu(s, N_1 S)}{\mu(N_1)} = \frac{\mu(T, N_1 S)}{\mu(N_1)} = \frac{\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{4}{5}}{\frac{1}{2}} = \frac{2}{5} \\
\mu(e_1 = K \mid N_1) &= 1 - \mu(S \mid N_1) = \frac{3}{5}
\end{aligned}$$

and the occurring policy-evidential quantities $\mu(e_1 \mid \pi_{1:2})$ are

$$\begin{aligned}
\mu(N_1 N_2) &= \sum_{s, e_1, e_2} \mu(s, N_1 e_1 N_2 e_2) \\
&= \mu(T, N_1 K N_2 S) + \mu(T, N_1 S N_2 0) + \mu(H, N_1 K N_2 0) \\
&= \frac{1}{100} + \frac{1}{10} + \frac{1}{5} = \frac{31}{100} \\
\mu(e_1 = K \mid N_1 N_2) &= \frac{\sum_{s, e_2} \mu(s, N_1 K N_2 e_2)}{\mu(N_1, N_2)} \\
&= \frac{\mu(T, N_1 K N_2 S) + \mu(H, N_1 K N_2 0)}{\mu(N_1 N_2)} = \frac{\frac{1}{100} + \frac{1}{5}}{\frac{31}{100}} = \frac{21}{31} \\
\mu(e_1 = S \mid N_1 N_2) &= 1 - \mu(K \mid N_1 N_2) = \frac{20}{31}
\end{aligned}$$

The policy $\pi_2 = \{Y_1\}$ always goes to the doctor for the treatment, and so

$$\mu(e_1 = C \mid Y_1) = 1$$

for both AESDT and PESDT.

Second percept. With the policy π_2, the second percept is always empty. Under π_1, the only action sequence that can reach the second percept is N_1KN_2

$$\begin{aligned}\mu(N_1KN_2) &= \sum_s \mu(s, N_1KN_2) = \mu(T, N_1KN_2) + \mu(H, N_1KN_2)\\ &= \frac{1}{100} + \frac{1}{5} = \frac{21}{100}\\ \mu(e_2 = S \mid N_1KN_2) &= \frac{\sum_s \mu(s, N_1KN_2S)}{\mu(N_1KN_2)} = \frac{\mu(T, N_1KN_2S)}{\mu(N_1KN_2)} = \frac{\frac{1}{100}}{\frac{21}{100}} = \frac{1}{21}.\end{aligned}$$

Value Functions. We start by evaluating the recursive definition from the second time step. The second step value functions are 0 for π_1 and for the history N_1S for π_2. For the history N_1K, both SAEDT and PAEDT assign the following identical value to π_2:

$$\begin{aligned}V_{\mu,2}^{\text{aev},\pi_1}(N_1K) = V_{\mu,2}^{\text{pev},\pi}(N_1K) &= \sum_{e_2} \mu(e_2 \mid N_1KN_2) \cdot u(e_2)\\ &= \mu(e_2 = S \mid N_1KN_2) \cdot u(S) + \mu(e_2 = 0 \mid N_1KN_2) \cdot u(0)\\ &= \frac{1}{21} \cdot (-10) + \frac{20}{21} \cdot 0 = -\frac{10}{21}\end{aligned}$$

The first step value functions now evaluates to:

$$\begin{aligned}V_{\mu,2}^{\text{aev},\pi_1} &= \sum_{e_1} \mu(e_1 \mid N_1) \cdot \big(u(e_1) + V_{\mu,2}^{\text{aev},\pi_1}(N_1e_1)\big)\\ &= \mu(S \mid N_1) \cdot (u(S) + V_{\mu,2}^{\text{aev},\pi_1}(N_1S))\\ &\quad + \mu(K \mid N_1) \cdot (u(K) + V_{\mu,2}^{\text{aev},\pi_1}(N_1K))\\ &= \frac{2}{5} \cdot (-10 + 0) + \frac{3}{5} \cdot (0 - \frac{10}{21}) = -\frac{30}{7} \approx -4.3\end{aligned}$$

$$\begin{aligned}V_{\mu,2}^{\text{pev},\pi_1} &= \sum_{e_1} \mu(e_1 \mid N_1) \cdot \big(u(e_1) + V_{\mu,2}^{\text{pev},\pi_1}(N_1e_1)\big)\\ &= \mu(S \mid N_1N_2) \cdot (u(S) + V_{\mu,2}^{\text{pev},\pi_1}(N_1S))\\ &\quad + \mu(K \mid N_1N_2) \cdot (u(K) + V_{\mu,2}^{\text{pev},\pi_1}(N_1K))\\ &= \frac{10}{31} \cdot (-10 + 0) + \frac{21}{31} \cdot (0 - \frac{10}{21}) = -\frac{110}{31} \approx -3.5\end{aligned}$$

Meanwhile, the value of π_2 is

$$\begin{aligned}V_{\mu,2}^{\text{aev},\pi_2} = V_{\mu,2}^{\text{aev},\pi_2} &= \sum_{e_1} \mu(e_1 \mid N_1) \big(u(e_1) + V_{\mu,2}^{\text{aev},\pi_2}(N_1e_1)\big)\\ &= \mu(C \mid Y_1)(u(C) + V_{\mu,2}^{\text{aev},\pi_2}(Y_1C)) = 1 \cdot (-4 + 0) = -4\end{aligned}$$

That is, $V_{\mu,2}^{\text{aev},\pi_1} < V_{\mu,2}^{\text{aev},\pi_2} = V_{\mu,2}^{\text{pev},\pi_2} < V_{\mu,2}^{\text{pev},\pi_1}$. So SPEDT but not SAEDT prefers π_1 to π_2. In other words, an SAEDT agent considers himself sufficiently likely to have the parasite to adopt policy π_2 of seeing the doctor. The SPEDT agent relies on the fact that he would pet the cat in case he saw it, and takes that as evidence of not being sick. Hence he will instead adopt policy π_1 of not seeing the doctor.