Can *de se* choice be *ex ante* reasonable in games of imperfect recall?

Caspar Oesterheld and Vincent Conitzer

February 11, 2022

Contents

$$\begin{tikzcd} {} \arrow[r] & s_0 \arrow[r, "\mathrm{c}"] \arrow[d, "\mathrm{e}"] & s_1 \arrow[r, "\mathrm{c}"] \arrow[d, "\mathrm{e}"] & 1 \\ & 0 & 4 \end{tikzcd}$$

Figure 1: A graphical description of the absent-minded driver (Example 1) in our formalism.

1 Introduction

In this paper, we study single-player games of imperfect recall. In general, these are sequential games in which the agent does not remember past observations. We illustrate this idea and some of its consequences using a well-known example.

Example 1 (Absent-minded driver (Piccione and Rubinstein, 1997))**.** *"In order to get home [a driver] has to take the highway and get off at the second exit. Turning at the first exit leads into a disastrous area ([utility] 0). Turning at the second exit yields the highest reward ([utility] 4). If he continues beyond the second exit, he cannot go back and at the end of the highway he will find a motel where he can spend the night ([utility] 1). The driver is absentminded and is aware of this fact. At an intersection, he cannot tell whether it is the first or the second intersection and he cannot remember how many he has passed." We illustrate this problem graphically (in the formalism introduced in Section 2.1) in Figure 1.*

If the agent had perfect recall, the agent would know whether he is at the first or second exit and it would be clear what he should do: continue at the first exit and get off at the second exit. Under imperfect recall (the driver being "absentminded"), the problem becomes more interesting. The driver cannot distinguish between the two exits and therefore has to choose in the same way at both intersections. It is easy to see that if the agent can randomize the optimal policy must do so (to at least have a chance of arriving home). (Specifically, as we calculate in Section 2.1, the optimal policy is to continue with probability $2/3$.)

There are many reasons why one might be interested in games of imperfect recall. The assumption of imperfect recall is clearly realistic for humans. More generally, agents that interacts with its environment via high-bandwidth channels (such as a video stream from a camera) typically cannot have perfect memory. A second reason to consider games of imperfect recall is that even when it is possible to remember everything, one can make the search for optimal policies more tractable by considering imperfect-recall policies (Waugh et al., 2009; Ganzfried and Sandholm, 2014; Čermák, Bošanský, and Lisý, 2017).

A third motivation is that games of imperfect recall can be used as models of other types of games and situations. For example, there is a close connection between single-player games of imperfect recall and team games (Isbell, 1957; Piccione and Rubinstein, 1997; Binmore, 1997; Detwarasiti and Shachter, 2005; Conitzer, 2019, Section "Imperfect Recall"; Emmons et al., 2021) wherein a group of symmetric players maximize a shared objective but cannot freely communicate with one another and thus cannot coordinate on symmetry-breaking strategies.[1] In the philosophy of science, the field of *anthropics* (also referred to as the study of observation selection effects) has asked questions structurally similar to the ones studied in this paper (see Bostrom, 2010, for an overview). For example, if one cosmological theory posits a large (perhaps even infinite) universe with many observers and another predicts a small universe with few observers, should we find the large-universe theory more plausible, other things being equal? Other arguments in anthropics that hinge on theories of self-location include the doomsday argument (first made by Carter, 1983); fine-tuning arguments for the existence of god or a multiverse (for an overview, see Friederich, 2021); the anthropic shadow argument which purports to show that risk of human extinction is larger than we would naively expect (Ćirković, Sandberg, and Bostrom, 2010); the simulation argument (Bostrom, 2003).

Perhaps the simplest normative notion of choice in scenarios of imperfect recall is what we will call *ex ante* (Latin for "from before") optimality. That is, one might assume the perspective from the beginning of the scenario (a hypothetical *planning stage*, as Aumann, Hart, and Perry (1997) call it) and then simply optimize the parameters of the policy. For example, in the absentminded driver, we could imagine (in line with the original story of Piccione and Rubinstein, 1997) that before stepping in his car, the driver knows what decisions he will face and also knows that he will be absent-minded. Then he could reason about what probability of continuing is optimal from this *ex ante* perspective. We will make this mathematically precise in Section 2.1.

However, in this paper, we specifically study the more contentious *de se* (Latin for "of oneself") theories of choice in scenarios of imperfect recall. These theories address how the agent should reason about her individual choices (or, as Aumann, Hart, and Perry (1997) put it, how she should reason at the *action stage*). For example, how should the absent-minded driver reason about his options (continuing versus exiting) when facing an intersection? In particular, our theories will take a policy π as given and specify whether the individual choices in π are rational, given that the agent otherwise follows π (somewhat analogous to the concepts of Nash equilibrium and ratificationism in game and decision theory). We will take such theories to consist of two components:

1. The first is a method of assigning *self-locating probabilities*: given a policy (e.g., given that I continue with probability $1/2$ in the absent-minded driver) and some observation (e.g., seeing an exit on the highway), what is the probability that I am in a particular state (e.g., that I am at the first exit). People disagree about such questions in the absent-minded driver. In Section 2.3, we formally define three different theories for assigning self-locating probabilities. We call them generalized thirding (GT) (a.k.a. consistency (Piccione and Rubinstein, 1997), a.k.a. the self-indication assumption (Bostrom, 2010)), single-halfing (GSH) (a.k.a. the (non-minimal-reference-class) self-sampling assumption (Bostrom, 2010)), and double-halfing (GDH) (a.k.a. Z-consistency (Piccione and Rubinstein, 1997), a.k.a. compartmentalized conditionalization (Meacham, 2008), a.k.a. the minimal-reference-class self-sampling assumption

[1] A recent line of machine learning research has studied such symmetric team games with the aim of improving a system's ability to coordinate with new team mates (i.e., opponents it has not been trained with), including humans (Hu et al., 2020; Treutlein et al., 2021).

(Bostrom, 2010)), so named after the probabilities they assign in the so-called Sleeping Beauty problem (as discussed at the beginning of Section 2.3).

2. The second component is a method for reasoning about the consequences of a choice. Here we distinguish two theories that have been proposed in the literature: causal and evidential decision theory (CDT and EDT) (terms loosely inspired by the discussions in philosophical decision theory following the publication of Newcomb's problem by Nozick, 1969). To illustrate how these two differ, take some policy and imagine that based on this the agent assigned some probabilities to being at the two different exits – for example, $3/4$ to being at the first and $1/4$ to being at the second exit. Then CDT would assign an expected utility of $3/4 \cdot 0 + 1/4 \cdot 4 = 1$ to exiting. In contrast, EDT will generally revise its self-locating probabilities based on its actions. For example, it assesses exiting to yield an expected utility of 0 with certainty, reasoning: if I exit now, then this means that I exit (and have exited in the past) at every intersection. Hence, I cannot be at the second intersection. We formally define these two theories in Section 2.4.

In this paper, we evaluate each of the six possible combinations of these theories separately. For each of these combinations we then ask to what extent they agree with the *ex ante* optimal policy and to what extent they avoid *Dutch books*, i.e., to what extent do these theories avoid a *sure loss* when it is possible to walk away with a guaranteed non-negative payoff. We will specify the questions we ask in particular in more detail below.

Contributions As mentioned above, each of the six theories X specifies for each given policy π whether the agent acts in agreement with X in all decision situations (i.e., upon all observations). When this is the case, we call π compatible with theory X. In Sections 3 to 5, we primarily ask the following two questions: Is the *ex ante*-optimal policy compatible with X in all scenarios? Failing that, is there always a non-Dutch-book policy compatible with X? That is, in scenarios where it is possible to achieve a guaranteed non-negative payoff, is there a policy compatible with X that has a non-negative (or positive) payoff with positive probability? An overview of the answers to these questions can be found in Table 1. As indicated in the table, previous work has already given some of the answers. In this paper, we fill the gaps in the picture from the existing literature. Specifically, our contributions are the following.

- The *ex ante* optimal policy is always compatible with evidential decision theory plus generalized double-halfing (Corollary 10). We thereby substantially generalize a result by Briggs (2010).
- Draper and Pust (2008) give a Dutch book argument against (CDT/EDT+) generalized single-halfing (GSH). Indeed, we find that their argument succeeds against one version of GSH (Section 5.1). However, we take a second look at GSH and find the following.
  - Draper and Pust's Dutch Book fails against an alternative plausible version of GSH that we call GSH* (Section 5.2). Our diagnosis is that GSH's failure results from the following: When a random coin is flipped halfway

through the scenario, then GSH assigns different probabilities to this event before versus after the event occurs. GSH* works by imagining that all randomization in the scenario occurs at the beginning of the scenario. It then applies GSH probability as usual. The idea is to thereby assign probabilities to random events more consistently. Note that the unique CDT+GSH*- and EDT+GSH*-compatible policy in Draper and Pust's scenario is still *ex ante* suboptimal.
  - We show that in each scenario in which history length is independent of the agent's choices, there exists a CDT+GSH*-compatible non-Dutch-book policy (Section 5.3).
  - However, in general (i.e., if history length is allowed to depend on the agent's choices), there are scenarios in which there exist no CDT+GSH*-compatible policies (Section 5.4.2.1). There are also scenarios in which there exist CDT+GSH*-compatible policies but in which all CDT+GSH*-compatible policies are Dutch book policies (Section 5.4.2.2).
  - For EDT+GSH*, on the other hand, there is a scenario with choice-independent history length in which the only compatible policy is a Dutch book (Section 5.5).

In Section 6, we discuss various conceptual issues raised by our results:

- Our positive results for CDT assume and hinge on the agent's ability to independently randomize at each decision point; our positive results for EDT do not. We discuss this difference in Section 6.1 and relate it analogous points made in the literature on Newcomb-like problems.
- In Section 6.2, we discuss Conitzer's (2015) Dutch book against EDT+GDH and how our version of EDT+GDH avoids it. Again we draw parallels to ideas from the literature on Newcomb-like decision problems.
- In general, any of the six *de se* theories of choice might permit multiple policies. It is easy to see that in some scenarios (e.g., Example 4, which we already give in Section 2.4.1), even a Dutch book policy is compatible with all six theories. In Section 6.3, pose the question of whether an agent can avoid Dutch books or perhaps even reliably choose the *ex ante* optimal policy while relying purely on *de se* reasoning. This question has received little attention in the literature. To show that the problem is hard, we demonstrate the failure of one natural approach to this problem. All three of our theories for assigning self-locating beliefs (GDH, GSH, GT) define for each policy and each observation o observed with positive probability an expected utility. If say the expected utility of π_1 is higher than the expected utility of π_2 from all decision perspectives, we might expect the agent to never follow policy π_2. Unfortunately, as we show in Section 6.3, there are scenarios in which some CDT+GT, EDT+GDH and CDT+GSH*-compatible Dutch book policy has strictly higher (GT/GDH/GSH) expected utility from all decision perspectives than all other (CDT+GT/EDT+GDH/CDT+GSH-)compatible policies. We thereby generalize results by Aumann, Hart, and Perry (1997, Section 5) and Korzukhin (2020).

GT	CDT	✓ *ex ante* opt. (Piccione and Rubinstein, 1997)
	EDT	✗ Dutch book (Briggs, 2010)
GDH	CDT	✗ Dutch book (Hitchcock, 2004)
	EDT	✓ *ex ante* opt. (Corollary 10)
GSH	CDT EDT	✗ Dutch book (Draper and Pust, 2008, Section 5)
GSH*	CDT	(✓) no DB if actions don't affect observations (Corollary 15) ✗ no compatible policy (Section 5.4.2.1) ✗ Dutch book (Section 5.4.2.2)
	EDT	✗ Dutch book (Section 5.5)

Table 1: Is the *ex ante* optimal policy compatible with *de se* choice? If not, is some non-Dutch-book policy compatible with *de se* choice? The answers depend on what theories of *de se* choice we use and (for CDT+GSH*) on what type of scenarios we consider. This table summarizes the answers given in the literature and in Sections 3 to 5 of the present paper.

2 Preliminaries

2.1 Single-player games of imperfect recall

Definition 1. *A (single-player) game of imperfect recall or scenario is a tuple* $(S, S_T, P_0, O, \omega, A, T, u)$*, consisting of:*

- *a finite set* S *of states;*
- *a set* $S_T \subseteq S$ *of terminal states;*
- *an initial state distribution* $P_0 \in \Delta(S - S_T)$*;*
- *a set of possible observations* O*;*
- *a function* $\omega\colon S - S_T \to O$ *that maps each non-terminal state onto the observation made by the agent in that state;*
- *a finite set* A *of possible actions that the agent can choose from;*
- *a probabilistic, conditional transition mapping* $T\colon S - S_T \times A \to \Delta(S)$ *that provides for any current non-terminal state* s *and any action* $a \in A$ *taken by the agent, a probability distribution* $T(\cdot \mid s, a)$ *over successor states; and*
- *a utility function* $u\colon S_T \to \mathbb{R}$ *that maps terminal states onto real numbers.*

As an example of a scenario, see the formalization of the absent-minded driver in Figure 1. Throughout this paper, we will provide many more examples.

Note that, in line with existing work, our formalism assumes the agent's goal to be constant throughout the scenario. In particular, we do not allow the goal to be relative to the current time step or observation. For instance, we do not allow agents whose goal it is to achieve some favorable outcome for the next time step. Since "the next time step" means different things to different instances of the agent, such a goal would mean that the agent's goals change across the scenario.

A *(memoryless) policy* is a probabilistic mapping $\pi\colon O \to \Delta(A)$ that determines for each observation $o \in O$ a probability distribution $\pi(\cdot \mid o)$ over actions.

We call a *history* of the scenario any finite sequence $s_0 \ldots s_n$ of states where $s_n \in S_T$ is a terminal state of the scenario, and $s_0, \ldots, s_{n-1} \notin S_T$, i.e., $s_0, \ldots, s_{n-1}$ are all non-terminal. Given a policy, the probability of a history is defined as

$$P(s_0 \ldots s_n \mid \pi) = P_0(s_0) \prod_{i=1}^{n} \left(\sum_{a \in A} \pi(a \mid \omega(s_{i-1})) T(s_i \mid s_{i-1}, a) \right).$$

Our scenarios are allowed to loop, i.e., a state may be visited multiple times. This creates the possibility of infinite histories. Infinite histories create a few problems. Most importantly, it is unclear how to generalize our theories of self-locating beliefs to infinite histories. Since addressing some of these problems is difficult and would distract from the main issues and ideas of this paper, we will generally assume infinite histories away. Specifically, we assume throughout this paper that for all policies π, all actions a and all states s, the game terminates with probability 1 if in state s, action a is played once and then policy π is followed, i.e.,

$$\sum_{n=1}^{k} \sum_{s_0 \ldots s_n :\, s_0 = s} P(s_1 \mid s_0, a) P(s_2 \ldots s_n \mid s_1, \pi) \to 1 \text{ as } k \to \infty,$$

where the inner sum is over all histories of length n that start in s and $s_n \in S_T$ is a terminal state. For simplicity, this assumption is a little stronger than needed. Interestingly, evidential decision theory requires weaker assumptions than causal decision theory. We discuss this briefly in Appendix B.

Let $\mathscr{E}$ be a scenario, π be a policy for $\mathscr{E}$, s be a state of $\mathscr{E}$, and a be an action for $\mathscr{E}$. Then define $Q_\pi(s,a)$ to be the expected utility of being in state s, choosing action a and then following policy π. Define $Q_\pi(s) := \sum_{a \in A} \pi(a \mid \omega(s)) Q_\pi(s,a)$ to analogously be the expected utility given that the current state is s and policy π is used. Finally, we use $Q_\pi(P_0) := \sum_{s \in S} P_0(s) Q_\pi(s)$ for the expected utility if an initial state is sampled from P_0 and the agent then follows π.

There are many alternative ways to specify games of imperfect recall. The present one resembles the formalism of episodic, partially observable Markov decision processes (POMDPs) as studied in machine learning. The substantial difference to episodic POMDPs is that we restrict consideration to memoryless policies, i.e., policies that only depend on the current observation (as opposed to the history of observations) (cf. Li, Yin, and Xi, 2011). Probably the most widespread representation of games of imperfect recall is the tree representation of extensive-form games used in game theory. More interestingly, there is a close analogy between games of imperfect recall (as studied in this paper), and common-payoff games with symmetry constraints on strategies. To our knowledge, this similarity was first pointed out by Isbell (1957) (and noted again by, for example, Piccione and Rubinstein, 1997; Binmore, 1997; Detwarasiti and Shachter, 2005; Conitzer, 2019, Section "Imperfect Recall"; Emmons et al., 2021).

2.2 Two *ex ante* standards of rational choice

2.2.1 *Ex ante* optimal policies

Let $\Pi \subseteq \Delta(A)^O$ be some set of policies. We call a policy $\pi^* \in \Pi$ *ex ante optimal in* Π if $\pi^* \in \arg\max_{\pi \in \Pi} Q_\pi(P_0)$. The two most important cases of Π are $\Pi = \Delta(A)^O$, which allows all mixed policies; and (using sloppy notation) $\Pi = A^O$, the set of *deterministic* policies which for each observation choose some action with probability 1. When considering the set of all mixed policies, we will refer to policies as (*ex ante*) *locally* optimal, if they are local optima of the function $\Delta(A)^O \to \mathbb{R} \colon \pi \mapsto Q_\pi(P_0)$.

For illustration, we now calculate the *ex ante* optimal policies in the absent-minded driver scenario (Example 1). First, among the two deterministic policies, the optimal one is to always continue for a reward of 1. Next, we calculate the optimal policy if mixing is allowed. For any p, let π_p be the policy that continues with probability p and exits with probability $1-p$. The *ex ante* expected utility of π_p is

$$Q_{\pi_p}(s_0) = P(s_0 s_1 1 \mid \pi_p) + 4P(s_0 s_1 4) = p^2 + 4p(1-p).$$

The unique local and global optimum of this polynomial is $p = 2/3$. Hence, the unique (globally and locally) optimal policy is $\pi_{2/3}$, to play continue with probability $2/3$.

For some sets Π there might not be a policy that is *ex ante* optimal in Π (e.g., for the absent-minded driver consider $\Pi = \Delta(A)^O - \{\pi_{2/3}\}$), but for the two most important cases, an *ex ante* optimal policy exists.

Proposition 1. *Let $\mathscr{E}$ be a scenario and let $\Pi = \Delta(A)^O$ or $\Pi = A^O$. Then $\mathscr{E}$ has at least one policy that is ex ante optimal in Π, if $\Pi = \Delta(A)^O$ or $\Pi = A^O$.*

Ex ante optimality is the simplest plausible normative concept for scenarios of imperfect recall. That is, one may simply posit that agents should choose according to some *ex ante* optimal policy π^*.[2] Like a number of previous works, the question we ask in this paper is whether *ex ante* optimality is consistent with what we will call *de se* theories of choice – we imagine that the agent finds herself within the scenario, forms some (probabilistic) belief about the current state and chooses based on this belief.

2.2.2 Dutch books

Definition 2. *Let $\mathscr{E}$ be a game of imperfect recall in which there is a policy that has a non-negative payoff with probability 1. Then we call a policy π for $\mathscr{E}$ a Dutch book policy if it has a negative payoff with probability 1, i.e., if for all histories $s_0 \ldots s_n$ of $\mathscr{E}$, $P(s_0 \ldots s_n \mid \pi) > 0 \implies u(s_n) < 0$.*

We can modify this definition in various slight ways. For instance, if π yields a negative reward with positive probability and a positive reward with probability 0, we might still call it a Dutch book (assuming $\mathscr{E}$ has policies that ensure a non-negative payoff). Moreover, we could generalize the no-Dutch-book principle by saying that for any number y, if it is possible to guarantee a payoff of at least y, one should not accept a policy that has a payoff less than y with certainty.[3] None of these small modifications would substantially affect any of our results or arguments.

[2] Armstrong (2011) explicitly argues that even *de se*, agents should choose by finding and following an *ex ante* optimal policy.

[3] Spencer (2021) calls this the Guaranteed principle.

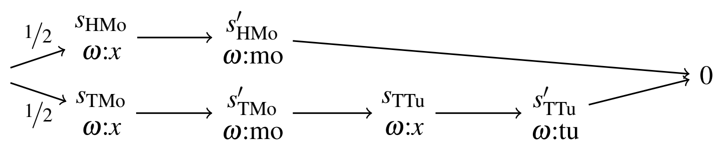

Figure 2: A graphical description of Lewis' variant of the Sleeping Beauty scenario (Example 2) in our formalism.

2.3 Three theories of assigning self-locating probabilities

In this section, we describe three theories of assigning self-locating probabilities. We start with a canonical example for distinguishing them. The basic version of the example – which resembles an absent-minded driver scenario with policy fixed to continuing with probability $1/2$ – was first given by Piccione and Rubinstein (1997, Example 5); the Sleeping Beauty story was introduced to the literature by Elga (2000). The specific variant of the below scenario (in which Sleeping Beauty is told in the afternoon which day it is) was introduced by Lewis (2001).

Example 2 (Sleeping Beauty)**.** *Sleeping Beauty falls asleep on Sunday. A group of researchers conduct the following experiment on her. First, they flip a fair coin. Regardless of the outcome of the coin, Sleeping Beauty is woken up on Monday morning. However, if the coin comes up Tails, they put her back to sleep in the evening, erase any memory of her awakening on Monday, and then wake her up for a second time on Tuesday. Upon waking up, Sleeping Beauty cannot tell whether it is Monday or Tuesday. Later she is then told what day it is (but not how the coin came up). We assume that when Sleeping Beauty consented to be part of this experiment, the setup as described above was explained to her. Two questions arise: 1. Upon waking up, what probability should Sleeping Beauty assign to the coin having come up Heads? 2. Upon being told in the afternoon that it is Monday, what probability should Sleeping Beauty assign to the coin having come up Heads? A graphical description of this problem in our formalism is given in Figure 2; notice that the agent faces no choices and we therefore omit actions from the graph.*

Upon first exposure to the first question, people are generally split, roughly evenly, into those who believe the answer is $1/2$ – "halfers" – and those who believe the answer is $1/3$ – "thirders". The simplest argument for the halfer position is that Sleeping Beauty wakes up regardless – hence, waking up is no evidence about the coin flip's outcome. The simplest argument for the thirder position is that in expectation exactly one third of the times that Beauty is awakened, she is awakened after the coin has come up Heads.

What about the second question? Here, too, two different answers have been given: $1/2$ and $2/3$. Thirders on the first question tend to give the $1/2$ answer based on the same argument as before: in expectation, exactly one half of the times that Beauty is told that it is Monday, the coin has come up Heads. Halfers, on the other hand, are split between the two positions. Halfers who give the $1/2$ answer for a second time are called double-halfers; we call halfers who give the $2/3$ answer single-halfers. The simplest argument for double-halfing is the same as the above argument for halfing: regardless of how the coin comes up, Beauty is told at some point that it is Monday. Hence, observing that it is Monday is no evidence either way. The argument for single-halfing is that the hypothesis that the coin came up Heads predicts twice as well that Beauty observes that it is Monday than the hypothesis that the coin came up Tails.

Much has subsequently been written about how to assign self-locating beliefs and in particular about what the correct answer (especially to the first question) should be. Like a few other authors (whom we will cite throughout this paper), we take an exclusively pragmatist approach to this question. That is, we will analyze the two positions in terms of whether they successfully guide choice.

In the rest of this section, we provide procedures that generalize (the arguments behind) double-halfing, single-halfing and thirding, respectively, to assign self-locating probabilities in other scenarios of imperfect recall.

2.3.1 Generalized double-halfing

We start by describing a principle that generalizes double-halfing. This way of belief formation was first introduced as "Z-consistency" by Piccione and Rubinstein (1997, Section 5). It is sometimes also referred to as compartmentalized conditionalization (following Meacham, 2008), as the minimal-reference class self-sampling assumption (following Bostrom, 2010), or simply as "The Halfer Rule" (e.g., Briggs, 2010)

We first describe generalized double-halfing (GDH) verbally in reference to our model. Our goal will be to assign – given some observation o – a probability to any statement of the form,

the true history is $s_0...s_n$ and the current time step is i

for any history $s_0...s_n$ and $i = 0, 1, ..., n-1$. For short, we write this statement as: i-th in $s_0...s_n$. It is immediately clear that to assign any such probability we usually (e.g., in the absent-minded driver, though not in Sleeping Beauty) need to know the agent's policy – otherwise we cannot even assign a non-self-locating probability to $s_0...s_n$. Overall the probabilities of interest are therefore of the form $P_{\text{GDH}}(i\text{-th in } s_0...s_n \mid \pi, o)$. Alternatively, we can think of GDH as assigning probabilities $P_{\text{GDH}}(s_0...s_n \mid \pi, o)$, since in any history $s_0...s_n$, GDH (uncontroversially) splits the probability mass $P_{\text{GDH}}(s_0...s_n \mid \pi, o)$ equally among all time steps in $s_0, ..., s_n$ in which o is observed.[4]

The crux of GDH is that $P_{\text{GDH}}(s_0...s_n \mid \pi, o)$ is simply the non-self-locating probability of $s_0...s_n$ conditional on the fact that o is observed at least once. Thus relative to the prior probabilities $P(s_0...s_n \mid \pi)$, $P_{\text{GDH}}(s_0...s_n \mid \pi, o)$ simply eliminates the histories in which o is never observed and then renormalizes the probabilities of the remaining histories. So in contrast to thirders, GDH updates the probability of all *histories* in which o is observed by the same factor, regardless of how many times o is observed in them. And in contrast to single-halfers, GDH does not take into account the number of observations *other* than o made in $s_0...s_n$.

[4]For some discussion of the principle of equally splitting probability mass in this way, see, e.g., Briggs (2010, Section 2.3) and references therein.

Before giving the formal definition, we need some more notation. Define $\#(o, s_0...s_n) := \sum_{i=0}^{n} \mathbb{1}[\omega(s_i) = o]$ to be the number of times o is observed in $s_0...s_n$. Further, define $P(s_0...s_n \mid \pi, o)$ to be the (regular, non-self-locating) probability of the sequence of states $s_0...s_n$, given that π is played and that o is observed at least once. That is, by the definition of conditional probability, if $\omega(s_i) = o$ for some $i \in \{0, ..., n-1\}$, then

$$P(s_0...s_n \mid \pi, o) = \frac{P(s_0...s_n \mid \pi)}{\sum_{s'_0...s'_k : \exists j : \omega(s'_j) = o} P(s'_0...s'_k \mid \pi)},$$

and else $P(s_0...s_n \mid \pi, o) = 0$.

Definition 3. *Let o be observed with positive probability under policy π. Then for all $s_0...s_n$ and $i = 0, 1, ..., n$ define $P_{\text{GDH}}(i\text{-th in } s_0...s_n \mid \pi, o) = 0$ if $\omega(s_i) \neq o$, and*

$$P_{\text{GDH}}(i\text{-th in } s_0...s_n \mid \pi, o) = \frac{P(s_0...s_n \mid \pi, o)}{\#(o, s_0...s_n)}$$

otherwise.

2.3.2 Generalized single-halfing

Single-halfers assign a probability of 1/2 to Heads before being told what day it is, but assign a probability of 2/3 once they are told that it is Monday. Among others, Lewis (2001) has argued for single-halfing in the Sleeping Beauty problem. We here present generalized single-halfing (GSH). Bostrom (2010) argues for it as a non-minimal reference class version of the self-sampling assumption. (Our version of GSH is, in our formalism, a *maximum* reference class version of the self-sampling assumption. We discuss reference classes more at the end of this section.) Generalized single-halfing is also assumed in the so-called *doomsday argument* (Carter, 1983). Thus, most versions of the doomsday argument contain GSH-like calculations, though usually without fully acknowledging their contentiousness.

The justification for single-halfing is that the hypothesis that the coin came up Heads predicts twice as well that Beauty observes that it is Monday than the hypothesis that the coin came up Tails. This is because the fraction of Monday observations is twice as large in the Heads branch than in the Tails branch (1/2 versus 1/4). Therefore, so the argument goes, she should update from the 50-50 prior toward the Heads branch.

We now describe *generalized single-halfing* (GSH). Like GDH, it is most naturally described as assigning probabilities to statements of the form, i-th in $s_0...s_n$. A natural interpretation of GSH is that it uses a prior $P(i\text{-th in } s_0...s_n \mid \pi) = 1/n P(s_0...s_n \mid \pi)$, which resembles the GDH probabilities. But then GSH performs a Bayes-like update. Recall that Bayes' theorem states that $P(x \mid y, z) = P(x)P(y \mid x, z) / \sum_{x'} P(x')P(y \mid x', z)$. Replacing x with the hypothesis i-th in $s_0...s_n$, y with the fact that I am observing o, and z with the fact that the agent follows π, we obtain the following definition of GSH.

Definition 4. *Let o be observed with positive probability under policy π. Then for all histories $s_0...s_n$ and $i = 0, 1, ..., n-1$ define $P_{\text{GSH}}(i\text{-th in } s_0...s_n \mid \pi, o) = 0$ if $\omega(s_i) \neq o$, and*

$$P_{\text{GSH}}(i\text{-th in } s_0...s_n \mid \pi, o) = \frac{\frac{1}{n} P(s_0...s_n \mid \pi)}{\sum_{s'_0...s'_k :\ \exists j : \omega(s'_j) = o} \frac{1}{k} P(s'_0...s'_k \mid \pi)}.$$

Further, define

$$P_{\text{GSH}}(s \mid \pi, o) := \sum_{s_0 \ldots s_n} \sum_{i:\ s_i = s} P_{\text{GSH}}(i\text{-}th\ in\ s_0 \ldots s_n \mid \pi, o).$$

From the definition, it is easy to verify that GDH and GDH are equivalent in single-observation scenarios such as the absent-minded driver.

Proposition 2. *In any single-observation ($|O| = 1$) scenario, GDH and GSH are equivalent, i.e., for all histories $s_0, \ldots, n$, policies π, observations o, and time steps $i \in \{0, \ldots, n-1\}$, $P_{\text{GDH}}(i\text{-}th\ in\ s_0 \ldots s_n \mid \pi, o) = P_{\text{GSH}}(i\text{-}th\ in\ s_0 \ldots s_n \mid \pi, o)$.*

2.3.2.1 Immediate concerns about generalized single-halfing Of our three theories, GSH is the only one that assigns unusual probabilities even in scenarios of *perfect* recall. For example, consider a version of Example 2 in which Beauty is told *immediately* upon waking up whether it is Monday or Tuesday, such that the observations of x disappear. Then it remains the case that $P(s'_{\text{HMo}} \mid \text{mo}) = 2/3$. But the scenario now has perfect recall in that every observation reveals to the agent what observations have been made in the past. This is a first hint that it might be difficult to square GSH with *ex ante* optimal policies.

Moreover, GSH is the only theory whose probabilities conditional on o depend on how many observations other than o are made in different histories (even if no relevant decision is ever made upon making these other observations). For instance, in Example 2 as formalized in Figure 2, Beauty's probability conditional on mo is sensitive to the fact that the scenario includes the states s_{TTu} and s'_{TTu} with observations x and tu, respectively. If we removed these states from the scenario and instead had the state transition directly from s'_{TMo} to s_T, then the GSH probability of s'_{HMo} conditional on mo (i.e., the probability of Heads conditional on it being Monday) would be $1/2$ (not $2/3$). Moreover, if we took Figure 2 as a base scenario and added, say, 100 states with some observation $\tilde{o}$ in both of the two branches of the scenario, then the probability of Heads (s'_{HMo}) conditional on it being Monday becomes close to $1/2$ ($52/103$, to be precise). This is another hint that GSH might be incompatible with *ex ante* optimal policies. But it also means that GSH probabilities are sensitive to more features of a given scenario than GDH and GT probabilities are – for GSH it matters, for example, for how long Beauty lives after the scenario, which is generally not specified. Fortunately, our formal models of these scenarios will clarify these things.

2.3.2.2 On reference classes We now explain the concept of *reference classes* to relate GSH (and GDH) to, e.g., Bostrom's self-sampling assumption, and to explain why we analyze this specific version of GSH. See Bostrom (2010) for a more detailed discussion of the self-sampling assumption and reference classes. The concept won't play a role in the analysis given in this paper.

Roughly, the main distinguishing feature of GSH (relative to generalized thirding and generalized double-halfing) is that it assigns lower probability to i-th in $s_0 \ldots s_n$ if n is large, even if the observations in s_j for $j \neq i$ are mostly different from the observation in s_i. The intuition behind it is that in addition to $s_0 \ldots s_n$ occurring in the first place, a further coincidence is required for i to be sampled from the n possible observer moments in this history. But now imagine, for example, that in states $s_0...s_{i-1}$ the agent's observations are similar to each other in some ways, while in $s_i...s_n$, the observations are very different. For example, you might imagine that $s_0 s_1$ are Monday and Tuesday that Sleeping Beauty spends in the experiment, and that $s_2...s_n$ is time spent after the experiment, dealing with completely different issues. Then perhaps upon observing $\omega(s_0)$ or $\omega(s_1)$, the agent should not reason that she could have also been one of the later observer moments. She should then not divide by n but by i. This is what is meant by saying that the observations made in time steps $1,...,i$ are in o's reference class. This is not so natural in our formalism and our typical stories of imperfect memory perhaps, but in other applications of self-locating beliefs this might be more natural. (Again, see Bostrom's book for examples.)

With the concept of reference classes in hand, we could provide different versions of reference classes. Roughly, we could specify for each observation $o \in O$ a set $r(o) \subseteq O$ as o's reference class. We could then define

$$\tilde{P}^r_{\mathrm{GSH}}(i\text{-th in } s_0...s_n \mid \pi, o) = \frac{\frac{1}{\#(r(o),s_0...s_n)} P(s_0...s_n \mid \pi)}{\sum_{s'_0...s'_k, j:\omega(s'_j)=o} \frac{1}{\#(r(o),s'_0...s'_k)} P(s'_0...s'_k \mid \pi)},$$

where $\#(r(o), s_0...s_n) \coloneqq \{i \in \{0,...,n-1\} \mid \omega(s_i) \in r(o)\}$ is the number of times that an observation from o's reference class is observed in the history $s_0...s_n$. GSH as given in Definition 4 is the special case where $r(o) = O$ for all $o \in O$. Note also that GDH is the special case where $r(o) = \{o\}$ for all $o \in O$.

Why do we only consider the maximum- and minimal-reference-class versions of GSH / the self-sampling assumption? First note that for single-observation scenarios (scenarios with $|O| = 1$), there is only one sensible choice of reference class, and in particular GSH and GDH are equivalent. Our negative results for CDT+GSH (and GDH) can be proved by using only single-observation scenarios. Second, we are not aware of any procedure (other than GSH and GDH) for determining r that is specified with sufficient rigor to discuss here. Lacking any specific candidate procedure, one could try to analyze $\tilde{P}^r_{\mathrm{GSH}}$ in general. However, without restrictions on what type of r we consider, it would be difficult to say anything interesting about $\tilde{P}^r_{\mathrm{GSH}}$, given that the two extremes – GDH and GSH as per Definitions 3 and 4, respectively – have very different properties in general, as we will see in this paper.

2.3.3 Generalized thirding

Finally, we describe generalized thirding (GT). GT was first given by Piccione and Rubinstein (1997, Section 5) as "consistency"; Bostrom (2010) calls it the self-indication assumption. In contrast to the halfer's position, we are not aware of any disputes as to how thirding is to be generalized in the present model. To calculate $P_{\mathrm{GT}}(s \mid \pi, o)$, the agent asks: what fraction of the times I observe o am I in state s? In other words, if the agent observes o and wants to assign a probability to being in a particular state s with $\omega(s) = o$, then GT dictates that she divide the expected number of times that s occurs by the expected number of times that o is observed.

To give a formal definition, we need some additional notation. Define $\#(s, s_0...s_n) := \sum_{i=0}^{n} \mathbb{1}[s_i = s]$ to be the number of occurrences of s in the history $s_0...s_n$, and $C_\pi(s) := \sum_{s_0...s_n} P(s_0...s_n \mid \pi)\#(s, s_0...s_n)$ to be the frequency of s under policy π, i.e., the expected number of times s occurs under policy π. Analogously, $C_\pi(o) := \sum_{s \in S - S_T :\, \omega(s)=o} C_\pi(s)$ is defined as the expected number of times that o is observed.

Definition 5. *Let $C_\pi(o) > 0$. Then $P_{\mathrm{GT}}(s \mid \pi, o) := 0$ if $s \in S_T$ or $\omega(s) \neq o$, and $P_{\mathrm{GT}}(s \mid \pi, o) := C_\pi(s)/C_\pi(o)$ otherwise.*

Contrary to our definitions of GSH and GDH, the above definition does not assign probabilities $P_{\mathrm{GT}}(i\text{-th in } s_0...s_n \mid \pi, o)$. Such probabilities can easily be defined.[5] However, we will not need them throughout the rest of this paper. Conversely, we could define GDH probabilities $P_{\mathrm{GDH}}(s \mid \pi, o)$ but we will not need these either. This asymmetry is due the fact that we will couple GT only with causal and GDH only with evidential decision theory.

2.4 *De se* choice using self-locating beliefs

How an agent chooses in a scenario of imperfect memory depends on how she assigns probabilities (whether she uses GDH, GSH, GT, or something else). But as others – including Piccione and Rubinstein (1997) – have pointed out it also depends on how an agent reasons about her choices. In particular, when choosing in response to some observation o, should she take into account that her choice will determine how she will choose (and has chosen in the past) in response to observing o? This can be shown in the absent-minded driver case (Piccione and Rubinstein, 1997; cf. Schwarz, 2015), but we will use the following simpler example.

Example 3. *Consider a variant of the Sleeping Beauty problem in which we skip the states in which Sleeping Beauty is told what day it is ($s'_{\mathrm{HMo}}, s'_{\mathrm{TMo}}, s'_{\mathrm{TTu}}$). Also, on each awakening, the agent is offered a bet that pays -1 if the coin came up Heads (the single-awakening branch) and $2/3$ if the coin came up Tails. We give a graphical description of this scenario in our framework in Figure 3.*

Clearly, the *ex ante* optimal policy in this problem is to (always) accept the bet, because *ex ante*, accepting pays -1 with 50% probability and $2 \cdot 2/3 = 4/3$ with 50%. But what happens if the agent reasons not *ex ante*, but using self-locating beliefs?

We focus on the case where the agent uses GDH or GSH, which (regardless of the agent's policy) assign a probability of $1/2$ to Heads–Monday (s_{HMo}), and $1/4$ to Tails–Monday (s_{TMo}) and Tails–Tuesday (i.e. to being either in s_{TMo} or s_{TTu}). Given these probabilities, how should the agent choose? There seem to be two plausible-looking but conflicting lines of reasoning, which we will associate with causal and evidential decision theory (CDT and EDT), respectively:

1. (CDT) With probability $1/2$ the coin came up Heads, in which case accepting the bet costs me 1. With the remaining probability $1/2$, I'm in the Tails branch, in which

[5] As usual, $P_{\mathrm{GT}}(i\text{-th in } s_0...s_n \mid \pi, o) = 0$ if $\omega(s_i) \neq o$. Otherwise, $P_{\mathrm{GT}}(i\text{-th in } s_0...s_n \mid \pi, o) = P(s_0...s_n/\pi)/C_\pi(o)$.

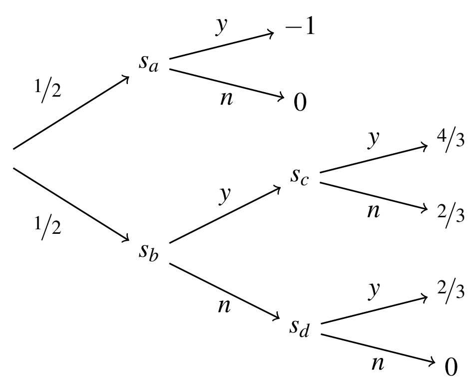

Figure 3: A graphical description of Example 3 in our formalism.

case – regardless of what I do in the other Tails branch awakening – accepting the bet earns me an extra $2/3$ relative to not accepting it. Since a 50% probability loss of 1 outweighs a 50% probability gain of $2/3$, I should reject the bet.

2. (EDT) With probability $1/2$ the coin came up Heads, in which case accepting the bet gives a payoff of -1. With the remaining probability $1/2$, I'm one of the two instances in the Tails branch. If I accept the bet, then the other instance of myself will also accept. Similarly, if I reject, my other instance will also reject. Hence, in the Tails branch, accepting earns me $4/3$. Since a 50% probability gain of $4/3$ outweighs a 50% probability loss of 1, I should accept the bet.

Piccione and Rubinstein (1997) discuss these two different styles of reasoning (though not under the names CDT and EDT) in their seminal work on games of imperfect recall. In the more philosophical literature following Elga (2000), Arntzenius (2002) is to our knowledge the first to make the connection to CDT versus EDT. But a few papers since have implicitly assumed (what we call) CDT or EDT; see, for example, Draper and Pust's (2008, Section 4) and Briggs' (2010, Section 3.2) discussions of Hitchcock's (2004) Dutch book.

With the relevance of CDT versus EDT, and thirding versus single-halfing versus double-halfing, we have six theories for choice under imperfect recall to consider. However, below we omit definitions and discussions of EDT plus thirding and of CDT plus double-halfing. This is because earlier work has shown, conclusively in our view, that these two combinations are vulnerable to Dutch books. In particular, Hitchcock (2004, Section 6) gives a scenario in which the only policy consistent with CDT plus double-halfing is a Dutch book policy (though, as noted, Hitchcock does not explicitly state that he considers CDT agents); Briggs (2010, Section 3.3) gives a scenario in which the only policy compatible with EDT plus GT is a Dutch book policy.Under the present agenda, we therefore have little more to say about CDT+GDH and EDT+GT. With a different agenda, these other combinations may still be interesting, as argued by Schwarz (2015).

2.4.1 Evidential decision theory plus generalized double- and single-halfing

First, it is helpful to notice that our methods of assigning self-locating beliefs immediately allow us to assign expected utilities conditional on an observation and a policy.

Definition 6. *For any policy π, we define the GDH/GSH expected utility, conditional on o being observed as*

$$\mathrm{EU}_{\mathrm{GDH/GSH}}(\pi, o) := \sum_{s_0 \ldots s_n} \sum_{i=0}^{n-1} P_{\mathrm{GDH/GSH}}(i\text{-}th \text{ in } s_0 \ldots s_n \mid \pi, o)u(s_n).$$

Upon observing o, how does an evidential decision theorist decide between different distributions $\alpha \in \Delta(A)$ over actions? The crux of EDT relative to CDT is that it evaluates a distribution $\alpha \in \Delta(A)$ by the expected utility $\mathrm{EU}_{\mathrm{GDH/GSH}}(\pi_{o \to \alpha}, o)$ for some policy $\pi_{o \to \alpha}$ that chooses α upon observing o. After all, choosing α upon the current observation of o is (definitive) evidence that the agent chooses α whenever she observes o.

But what happens if our scenario has multiple observations occurring with positive probability (e.g., some version of Lewis' Sleeping Beauty problem with decisions)? Then to evaluate a distribution α to play upon o, we need to calculate some expected utility $\mathrm{EU}_{\mathrm{GDH/GSH}}(\pi_{o \to \alpha}, o)$ where $\pi_{o \to \alpha}(\cdot \mid o) = \alpha$ – but what should be the rest of π? In principle, upon observing o, the agent might form beliefs about her choice for other observations o'. However, upon observing o' she also needs to form beliefs about what she would do in o. We thus run into a specific version of the circularity problem of multiple agents reasoning about one another. In principle, a rational agent should be able to deal with such problems in general, including in game-theoretic cases where the other agent has different goals. Unfortunately, it would be difficult and contentious to define and analyze such a procedure here.[6]

To avoid the circularity, we take inspiration from the concept of Nash equilibrium in game theory and ratificationism in decision theory (Jeffrey [1965] 1983, Section 1.7; Weirich, 2016, Section 3.5). For now, we take a specific policy π as given and then merely ask for each observation (that occurs with positive probability given π): assuming the agent follows π for all observations other than o, is it optimal as judged by EDT plus GDH/GSH to play $\pi(\cdot \mid o)$ upon observing o? This gives us a necessary condition for a policy π to be knowingly followed by an EDT+GDH/GSH agent.

For the formal definition, we need the following bit of notation. For any policy π, any $o \in O$ and distribution $\alpha \in \Delta(H)$, define the policy $\pi_{o \to \alpha}$ to be the one that is equal to π, except that upon observation o, it chooses according to distribution α. Formally, we let $\pi_{o \to \alpha}(a \mid o) = \alpha(a)$ and for all $o' \neq o$, $\pi_{o \to \alpha}(a \mid o') = \pi(a \mid o')$.

Definition 7. *We say that a policy $\pi \in \Pi$ is compatible with EDT+GDH/EDT+GSH as restricted to Π if for all $o \in O$ that are observed with positive probability under π,*

$$\pi(\cdot \mid o) \in \operatorname*{arg\,max}_{\alpha \in \Delta(A): \pi_{o \to \alpha} \in \Pi} \mathrm{EU}_{\mathrm{GDH/GSH}}(\pi_{o \to \alpha}, o).$$

[6] Skyrms (1990) work on deliberational dynamics addresses this general problem. Another well-known approach is no-regret learning (e.g. Cesa-Bianchi and Lugosi, 2006).

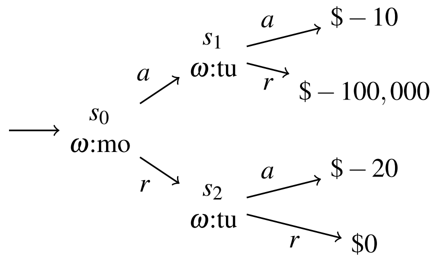

Figure 4: A graphical formalization of Example 4.

Our calculations w.r.t. Example 3 provide a first example of EDT+GDH reasoning. To provide a second example, we calculate the EDT+GDH-compatible policies in the absent-minded driver (Example 1). (By Proposition 2, the calculation for EDT+GSH yields the same result.) This calculation will also give an intuition for the proof of our positive results in Section 4. Letting π_p again be the policy that continues with probability p, EDT+GDH requires that the agent continues with a probability from $\arg\max_{p\in[0,1]} \mathrm{EU}_{\mathrm{GDH}}(\pi_p)$. By definition, $\mathrm{EU}_{\mathrm{GDH}}(\pi_p)$ is equal to

$$\begin{aligned}
& P_{\mathrm{GDH}}(\text{0-th in } s_0 0 \mid \pi_p)\cdot 0 + P_{\mathrm{GDH}}(\text{0-th in } s_0 s_1 4 \mid \pi_p)\cdot 4 + P_{\mathrm{GDH}}(\text{1st in } s_0 s_1 4 \mid \pi_p)\cdot 4 \\
& \quad + P_{\mathrm{GDH}}(\text{0-th in } s_0 s_1 1 \mid \pi_p)\cdot 1 + P_{\mathrm{GDH}}(\text{1st in } s_0 s_1 1 \mid \pi_p)\cdot 1 \\
= \;& P(s_0 s_1 4 \mid \pi_p)\cdot 4 + P(s_0 s_1 1 \mid \pi_p)\cdot 1.
\end{aligned}$$

Clearly, this is exactly, the *ex ante* expected utility of π_π, which we have seen (in Section 2.2.1) is uniquely maximized at $p = 2/3$. Thus, the *ex ante* optimal policy $\pi_{2/3}$ is the unique EDT+GDH/GSH-compatible policy.

In both the absent-minded driver and Example 3, only the *ex ante* optimal policy is EDT+GDH/GSH compatible policy. We now give a simple scenario in which there are multiple compatible policies, one of which is a Dutch book (and thus in particular *ex ante* suboptimal).

Example 4. *On Monday, Alice is offered* \$10*. However, if she accepts, this causes a time bomb to be hidden somewhere in her house. On Tuesday, Alice does not remember whether she accepted the offer on Monday (and therefore does not know whether a bomb is hidden in her house). Alice can now decide whether to buy equipment at a price of* \$20 *to find and defuse the bomb. (This equipment is 100% effective.) If a bomb was placed and she does not defuse it, the explosion will cause damages to Alice's house worth* \$100,000*. A graphical description of this problem in our formalism is given in Figure 4.*

Clearly, the optimal policy for this problem is to reject the offer on Monday and to not buy the equipment on Tuesday, thus resulting in a certain payoff of 0. However, consider the policy $\tilde{\pi}$ that accepts the offer on Monday and buys the equipment on Tuesday. This policy loses money with certainty, but is EDT+GDH/GSH compatible. Intuitively, if Alice on Monday believes that on Tuesday she will buy the equipment anyway, then by accepting the offer she earns an extra \$10; and if on Tuesday Alice believes that on Monday she accepts the bomb, then she better defuse the bomb to avoid the loss of \$100,000.[7] We thus conclude the following.

Proposition 3. *In Example 4, there exists a EDT+GDH/GSH-compatible Dutch book policy.*

For most of this paper, we will not be concerned with whether there exist compatible Dutch book policies; we will primarily consider whether there are good (i.e., *ex ante* optimal or at least non-Dutch-book) policies. However, since multiple policies might be EDT+GDH/GSH compatible, our definition will in general not fully answer the question of what policy an EDT+GDH/GSH agent does or should follow. We will discuss this multiplicity of compatible policies more in Section 6.3.

2.4.2 Causal decision theory plus generalized thirding/single-halfing

We now define causal decision theory plus generalized thirding and causal decision theory plus generalized single-halfing; the two definitions are the same up to the use of different probabilities. Many of the ideas here are analogous to those in the previous section. First we will define $\mathrm{EU}_{\mathrm{GT/GSH}}(\pi, o, a)$ to be the expected utility under GT/GSH probabilities of observing o, choosing a now and otherwise following π, including in other instances of observing o. We take a policy as given and then ask whether for all o that occur with positive probability, $\pi(\cdot \mid o)$ only chooses actions that are optimal in terms of $\mathrm{EU}_{\mathrm{GT/GSH}}(\pi, o, a)$. Note that, in contrast to our definition of EDT, it is sufficient to define the causal expected value of taking some action $a \in A$ deterministically. We could define the causal expected utility of a probability distribution $\alpha \in \Delta(A)$ as $\mathrm{EU}_{\mathrm{GT/GSH}}(\pi, o, \alpha) = \sum_{a\in A} \alpha(a)\mathrm{EU}_{\mathrm{GT/GSH}}(\pi, o, a)$.[8] However, it is then easy to see that choosing a distribution α maximizes $\mathrm{EU}_{\mathrm{GT/GSH}}(\pi, o, \alpha)$ if and only if all $a \in A$ with $\alpha(a) > 0$ maximize $\mathrm{EU}_{\mathrm{GT/GSH}}(\pi, o, a)$.

Definition 8. *For any policy π, any observation $o \in O$ observed with positive probability, and any $a \in A$, define*

$$\mathrm{EU}_{\mathrm{GT/GSH}}(\pi, o, a) := \sum_{s\in S} P_{\mathrm{GT/GSH}}(s \mid \pi, o) Q_\pi(s, a).$$

Definition 9. *We say that a policy $\pi \in \Delta(A)^O$ is CDT+GT/GSH compatible if for all $o \in O$ that are observed with positive probability under π, and all a^* s.t. $\pi(a^* \mid o) > 0$, $a^* \in \arg\max_{a\in A} \mathrm{EU}_{\mathrm{GT/GSH}}(\pi, o, a)$.*

We do not define CDT under a restricted set of policies for reasons we will describe in Section 6.1.

[7] Note that there is also a third EDT+GDH/GSH-compatible policy: accepting on Monday with probability $1/5{,}000$, and defusing on Tuesday with probability $1 - 1/10{,}000$. When following this policy, the EDT+GDH/GSH agent agent is indifferent between all probability distributions on both Monday and Tuesday. Readers familiar with game theory will notice a similarity to the structure of the set of Nash equilibria in many games. This is no coincidence, as we will see in Section 4.

[8] The analogous equality does not hold for EDT, i.e., in general $\mathrm{EU}_{\mathrm{GT/GSH}}(\pi_{o\to\alpha}, o)$ is not necessarily equal to $\sum_{a\in A} \alpha(a)\mathrm{EU}_{\mathrm{GT/GSH}}(\pi_{o\to a}, o)$. The absent-minded driver serves as an example.

Proposition 4. *In Example 4, a Dutch book policy is compatible with CDT+GT and with CDT+GSH.*

A similar result is due to Korzukhin (2020). Specifically, he gives a variant of Sleeping Beauty in which some policy $\tilde{\pi}$ is compatible with CDT+GT (but neither with EDT+GDH nor with CDT+GSH) and loses money with certainty. We also give a single-observation case with a CDT+GSH- and CDT+GT-compatible Dutch book policy in Section 6.1, based on Conitzer's (2015) "Three Awakenings" case. (As we will see in Corollary 11, in single-observation scenarios, only (and exactly) the *ex ante*-optimal policies are EDT+GDH/GSH compatible.)

3 *Ex ante* optimal policies are compatible with causal decision theory plus generalized thirding

In this section, we give Piccione and Rubinstein's (1997, Proposition 3) result that in every scenario the *ex ante* optimal policy is compatible with causal decision theory plus generalized thirding. While Piccione and Rubinstein give a monolithic proof of this result, we first show a lemma that encapsulates the key idea of the proof of this result. From this lemma we obtain a characterization of CDT+GT-compatible policies (Theorem 6). From the characterization, Piccione and Rubinstein's result then follows (Corollary 7). We prove these results in Appendix D. All proofs follow the main ideas in Piccione and Rubinstein's proof.

We first define the derivatives of $Q_\pi(P_0)$ w.r.t. $\pi(a \mid o)$. These aren't uniquely defined by regular multivariable calculus, because (infinitesimally) increasing $\pi(a \mid o)$ must be accompanied by (infinitesimally) decreasing some of the other probabilities in $\pi(\cdot \mid o)$ to make sure that $\pi(\cdot \mid o)$ remains a probability distribution. We here let all other probabilities in $\pi(\cdot \mid o)$ decrease infinitesimally and uniformly. We can then define the derivative as follows.

Definition 10. *Let π be a policy, a be an action and o be an observation. Then for all $\varepsilon > 0$ define*

$$\pi_{\varepsilon,a,o}(a' \mid o') = \begin{cases} \pi(a' \mid o'), & \textit{if } o' \neq o \\ (1-\varepsilon)\pi(a' \mid o), & \textit{if } o' = o \textit{ and } a' \neq a \\ (1-\varepsilon)\pi(a' \mid o) + \varepsilon, & \textit{if } o' = o \textit{ and } a' = a \end{cases} .$$

Then define

$$\frac{d}{d\pi(a \mid o)} Q_\pi(P_0) := \lim_{\varepsilon \downarrow 0} \frac{Q_{\pi_{\varepsilon,a,o}}(P_0) - Q_\pi(P_0)}{\varepsilon}.$$

Note that in particular if $\pi(a \mid o) = 1$, then $\frac{d}{d\pi(a|o)} Q_\pi(P_0) = 0$. It turns out that the derivatives of $Q_\pi(P_0)$ are closely related to the causal expected utilities. For the following, define $\mathrm{EU}_{\mathrm{GT}}(\pi, o) := \sum_{a \in A} \pi(a \mid o) \mathrm{EU}_{\mathrm{GT}}(\pi, o, a)$ to be the CDT+GT expected utility of following π upon observing o.

Lemma 5. *For o observed with positive probability,*

$$\frac{d}{d\pi(a \mid o)} Q_\pi(P_0) = C_\pi(o)(\mathrm{EU}_{\mathrm{GT}}(\pi, o, a) - \mathrm{EU}_{\mathrm{GT}}(\pi, o)).$$

We here give some rough intuition for why this result holds. The left-hand side of the equation asks: What happens if we increase the probability of playing a in o by some small but positive ε? It is helpful to focus on the case where $\pi(a \mid o) = 0$, i.e., where π would otherwise never take action a when observing o. The crucial idea is that as $\varepsilon \to 0$, the probability that a is played multiple times in o (being on the order $O(\varepsilon^n)$ for $n \geq 2$) is negligible compared to the probability that a is played once in o (being on the order $O(\varepsilon)$). Thus, the effect of infinitesimally increasing the probability of playing a in o is dominated by the effect of playing *a exactly* once, while otherwise following π, as compared to always following π. Assuming that there is a single deviation, the probability that such a deviation happens at any particular state s with $\omega(s) = o$ is proportional to $C_\pi(s)$, the frequency with which s occurs under π, as $\varepsilon \to 0$. Thus, the expected effect of selecting a once in o is $\mathrm{EU}_{\mathrm{GT}}(\pi, o, a) - \mathrm{EU}_{\mathrm{GT}}(\pi, o)$. The factor of $C_\pi(o)$ reflects the fact that for any given ε the probability that there is a deviation at all – and thus the size of the derivative – is proportional to the expected number of times that o is observed under π.

From Lemma 5, we directly obtain the following result.

Theorem 6. *A policy $\pi \in \Pi = \Delta(A)^O$ is CDT+GT compatible if and only if for all $o \in O$ and $a \in A$, $\frac{d}{d\pi(a|o)} Q_\pi(P_0) \leq 0$.*

This characterization in turn implies the following important corollary.

Corollary 7 (Piccione and Rubinstein, 1997)**.** *Let π be a globally ex ante optimal strategy from $\Pi = \Delta(A)^O$. Then π is CDT+GT compatible.*

To our knowledge, Piccione and Rubinstein (1997, Proposition 3) are the first to prove Corollary 7. In a blog post, Taylor (2016) also proves Corollary 7, in a formalism closer to ours. In the domain of multi-agent symmetric games, a related result (with a similar proof idea) has been given by Emmons et al. (2021, Theorem 3.2).

Briggs (2010, Sections 3.4 and 3.5) and Conitzer (2015b, Section 4) give related results. Their results require some restrictions on the game structure but allow the stronger conclusion that the CDT+GT-compatible policies are exactly the *ex ante* optimal ones (cf. the discussion of these results by Korzukhin, 2020).

4 *Ex ante* optimal policies are compatible with EDT plus generalized double-halfing

In this section, we prove positive results for evidential decision theory plus generalized double-halfing that are analogous to those in Section 3 for CDT+GT. We first give a characterization of EDT+GDH-compatible policies (Theorem 8). From this characterization it follows directly that every *ex ante* optimal policy is EDT+GDH compatible. We then give two further interesting corollaries. The first is that in scenarios with

$|O| = 1$, the EDT+GDH-compatible policies are exactly the *ex ante* optimal ones. The second is that all EDT+GDH-compatible policies are also CDT+GT compatible. To our knowledge, all results of this section are novel.

Theorem 8. *Let $\Pi \subseteq \Delta(A)^O$. A policy $\pi \in \Pi$ is EDT+GDH compatible in Π if and only if for all $o \in O, \alpha \in \Delta(A)$ s.t. $\pi_{o\to\alpha} \in \Pi$, $Q_\pi(P_0) \geq Q_{\pi_{o\to\alpha}}(P_0)$.*

Theorem 8 can be interpreted game-theoretically.[9] In order to keep it brief and because we have not defined any game-theoretic terminology anyway, we only state this informally:

Corollary 9 (informal)**.** *Let $\Pi \subseteq \Delta(A)^O$ be a set of policies. Let the (game-theoretic) common-payoff (i.e., fully cooperative) game Γ be constructed as follows. For each $o \in O$ there is a player. Each player chooses from $\Delta(A)$. For any (pure) strategy profile $(\alpha_o \in \Delta(A))_{o\in O}$ and corresponding policy $\pi\colon o \mapsto \alpha_o$, we define the utility of $(\alpha_o)_{o\in O}$ as $u((\alpha_o)_{o\in O}) = Q_\pi(s_0)$ if $\pi \in \Pi$ and $u((\alpha_o)_{o\in O}) = -\infty$ otherwise. Then any $\pi \in \Pi$ is EDT+GDH compatible if and only if the corresponding pure strategy profile $(\pi(\cdot \mid o))_{o\in O}$ is a Nash equilibrium of Γ.*

This model of reasoning about scenarios of imperfect recall – setting up a game with one player for each $o \in O$ choosing $\pi(\cdot \mid o)$ – has been proposed as the multiself approach by Piccione and Rubinstein (1997, Section 7). However, they do not show this to be equivalent to any method of choice under self-locating beliefs.

As a consequence of Theorem 8, *ex ante* optimal policies are EDT+GDH compatible:

Corollary 10. *Let Π be any set of policies for $\mathscr{E}$ and let π be ex ante optimal in Π. Then π is compatible with EDT+GDH restricted to Π.*

As with CDT+GT, Briggs (2010, Section 3) gives a related result. Again, Briggs' result assumes some restrictions on the game structure and under these restrictions, the EDT+GDH policy (like the CDT+GT policy) is unique (cf., again, the discussion of these results by Korzukhin, 2020).

Theorem 8 and the equivalence between GDH and GSH on single-observation scenarios (Proposition 2) also directly implies the following.

Corollary 11. *Let $\mathscr{E}$ be a scenario that has only one observation (i.e., $|O| = 1$). Let $\Pi \subseteq A^O$. Then a policy $\pi \in \Pi$ is ex-ante optimal if and only if π is compatible with EDT+GDH/GSH restricted to Π.*

Note that the same could not be said of CDT, as shown by, e.g., Aumann, Hart, and Perry (1997, Section 5), Conitzer's (2015) "Three Awakenings" (cf. our Example 9) and Korzukhin (2020).

With the help of Theorems 6 and 8, we can also obtain the following result.

[9] As we have noted, there is a similar relation between CDT+GT and Nash equilibrium in symmetric games (cf. Emmons et al., 2021). The main difference is that for the following result about EDT+GDH we specify one player per observation, whereas for CDT+GT we need one player per observation. We do not give the result in detail in this paper, because it is even more cumbersome to state formally.

Corollary 12. *If a policy is EDT+GDH compatible (without any policy restriction), it is CDT+GT compatible.*

Since Corollary 11 does not hold true for CDT, the converse of Corollary 12 also does not hold.

5 On generalized single-halfing

5.1 Draper and Pust's Dutch book argument against single-halfing

We start by describing a version of Draper and Pust's Dutch book argument against single-halfing.

Example 5 (adapted from Draper and Pust, 2008, Section 5)**.** *As a base scenario, take Example 2 (Lewis' variant of the Sleeping Beauty problem) but imagine that Beauty is told immediately upon waking up what day it is. On Sunday, Beauty is offered a bet that costs* \$15 *and pays* $\$30+\varepsilon$ *if the coin comes up Tails. On Monday afternoon (regardless of how the coin came up), Beauty is offered a bet that costs* \$18 *and pays* $\$30+\varepsilon$ *if the coin came up Heads (i.e., if she awakes only on Monday and not on Tuesday). If the coin comes up Tails, she is awakened again on Tuesday, with no relevant decision to make.*

This version differs from Draper and Pust's in two ways. First, we added small extra payoffs ($+\varepsilon$) to make sure that single-halfers must *strictly* prefer the choices that lead them to get Dutch-booked. Second, the bet offered on Monday costs \$18 as opposed to \$20. As we will see below, this is just low enough for generalized single-halfers to have to accept this bet. If the bet cost \$20, generalized single-halfers would reject it and the Dutch book would not work against our version of GSH (as per Definition 4). Adding an extra decision point on Sunday changes the GSH probabilities of Heads and Tails (at least as defined in Definition 4) on Monday from the usual $(2/3, 1/3)$ to $(3/5, 2/5)$.[10]

If Beauty accepts both bets, then she loses $\$3-\varepsilon$ with certainty. But Draper and Pust argue that in both decision situations, single-halfers prefer accepting the bets (independent of what is chosen at the other decision point). Indeed, Draper and Pust's argument is compatible with our formalism and the definition of CDT+GSH in Definition 9. Consider the version of Example 5 in Figure 5. Upon observing that it is Sunday, Beauty knows that the current state is s_{Su}, and it is easy to see that regardless of π, $Q_\pi(s_{\mathrm{Su}}, \mathrm{accept}) > Q_\pi(s_{\mathrm{Su}}, \mathrm{reject})$. Upon observing that it is Monday, things are slightly more complicated, since Beauty's beliefs depend on π and for mixed π she might assign positive probability to multiple states. For simplicity assume that π accepts the offer with probability 1 on both Monday and Sunday. Then omitting the normalizing constants, we obtain $P_{\mathrm{GSH}}(s_{\mathrm{T,Mo},a} \mid \mathrm{mo}, \pi_{\mathrm{accept}}) = P_{\mathrm{GSH}}(\text{1st in } s_{\mathrm{Su}} s_{\mathrm{H,Mo},a} - 5 \mid \mathrm{mo}, \pi_{\mathrm{accept}}) \sim 1/2 \cdot 1/2 = 1/4$ and $P_{\mathrm{GSH}}(s_{\mathrm{T,Mo},a} \mid \mathrm{mo}, \pi_{\mathrm{accept}}) = P_{\mathrm{GSH}}(\text{1st in } s_{\mathrm{Su}} s_{\mathrm{T,Mo},a} s_{\mathrm{T,Mo},a,a} -$

[10] Is this an oversight on Draper and Pust's part? It may be, but it might also be the case that Draper and Pust assume a version of GSH that uses a smaller reference class (see Section 2.3.2.2). Specifically, their analysis is accurate if we assume that the awakenings on Monday and Tuesday form one reference class, while the awakening on Sunday forms its own reference class.

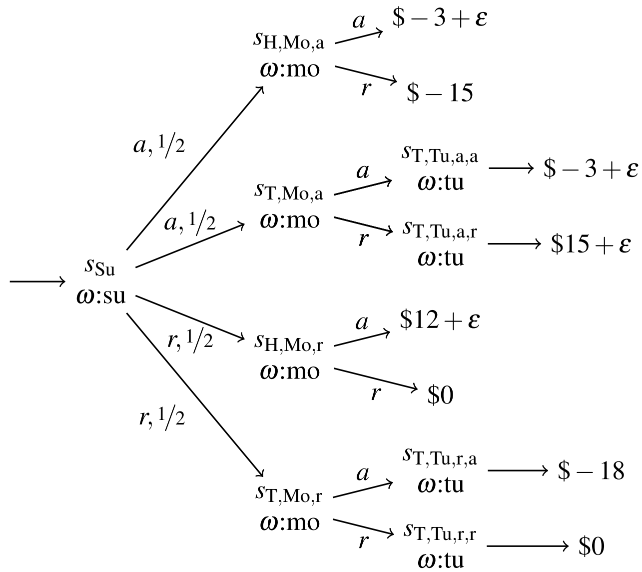

Figure 5: A graphical formalization of Example 5 (adapted from Draper and Pust, 2008, Section 5). In this formalization, the only CDT+GSH-compatible policy is accepting the bet on both Sunday and Monday, which is a Dutch book.

$5 \mid \text{mo}, \pi_{\text{accept}}) \sim 1/2 \cdot 1/3 = 1/6$; all other histories have probability zero under this policy. To renormalize, we have to divide by the sum, i.e., by $1/4 + 1/6 = 5/12$. We thus get that the single-halfer's probabilities of Heads and Tails given that it is Monday are $P_{\text{GSH}}(s_{\text{H,Mo},a} \mid \text{mo}, \pi_{\text{accept}}) = 3/5$ and $P_{\text{GSH}}(s_{\text{T,Mo},a} \mid \text{mo}, \pi_{\text{accept}}) = 2/5$, respectively. Hence, the causal (and evidential) expected utility of accepting relative to rejecting is $3/5 \cdot (12+\varepsilon) - 2/5 \cdot 18 = 3\varepsilon/5$. Hence, the generalized single-halfer accepts the second bet as well.

With the help of Draper and Pust we have thus shown the following.

Proposition 13 (Draper and Pust, 2008, Section 5)**.** *There is a scenario in which the only CDT+GSH/EDT+GSH-compatible policy is a (deterministic) Dutch book policy.*

5.2 How single-halfers can avoid Draper and Pust's Dutch book

In this section, we present a defense of single-halfing against Draper and Pust's Dutch book. In particular, we present an argument to the extent that single-halfers should strictly prefer to reject the bet on Sunday! This approach may be surprising, since accepting the Sunday bet seems unproblematic. For instance, accepting the Sunday bet is *ex ante* optimal. This approach will therefore not address the concern of *ex ante* suboptimality: our argument will have the agent reject the Sunday bet and accept the Monday bet, which is *ex ante* suboptimal and which in particular yields a voluntary *expected* loss from the *ex ante* perspective. In the following we first give the argument informally and then make it formal. We will specifically discuss CDT+GSH – while the same argument applies to EDT+GSH in this particular scenario, we show in Section 5.5 that EDT+GSH is vulnerable to other simple Dutch books.

Imagine you are the subject of Example 5 and that it is Sunday. You will be put to sleep momentarily and the fair coin will be flipped in a few hours. Should you believe that the probability of Heads/Tails is 50%? Following the general pattern of the single-halfer's argument, you might think the following: Under the hypothesis that the coin will come up Heads, I will have two observation moments, one of which is that I observe that it is Sunday. If the coin will come up Tails, I will have *three* observation moments, one of which is that I observe that it is Sunday. Thus, the Heads hypothesis better predicts that it is Sunday. Thus, observing that it is Sunday should update me towards the Heads hypothesis. In particular, I should not bet at (close to) even odds that the coin will come up Tails.[11]

We now want to make the argument formal. First, we can easily verify that GSH (as defined in Definition 4) indeed updates toward the Heads histories, even upon observing that it is Sunday. For simplicity, consider again the policy π that accepts both bets with probability 1. Then in the same way as above, we obtain that the single-halfer's probabilities of Heads and Tails given that it is Sunday are $P_{\text{GSH}}(\text{0-th in } s_{\text{Su}}s_{\text{H,Mo},a} - 5 \mid \text{su}, \pi_{\text{accept}}) = 3/5$ and $P_{\text{GSH}}(\text{0-th in } s_{\text{Su}}s_{\text{T,Mo},a}s_{\text{T,Mo},a,a} - 5 \mid \text{su}, \pi_{\text{accept}}) = 2/5$, respectively.

So why does Draper and Pust's Dutch book work against CDT+GSH? The problem lies in how CDT as defined in Definition 9 uses the GSH probabilities. When CDT observes that it is Sunday, it uses GSH probabilities to determine the probabilities of different *states* (not histories). In this particular case with the formalization in Figure 5, the observation that it is Sunday uniquely determines the current state to be s_{Su}, regardless of whether we use GSH or something else. For calculating the expected utilities of different actions in this state, CDT+GSH (like CDT+GT) simply uses $Q_\pi(s_{\text{Su}}, a)$, which does not take any further input from GSH and in particular does not take into account GSH's belief that the coin will come up Heads with probability $3/5$ not $1/2$.

One approach to fix this would be to try to modify the values $Q_\pi(s,a)$ in such a way that they incorporate the single-halfer's probabilities. We here use a different approach. Since CDT does take into account GSH's probabilities over states, we will modify the formal representation of the scenario in such a way that all probability judgments made by GSH are reflected in the GSH probabilities assigned to states conditional on the current observation (while in the above scenario, some of them are only visible in the probabilities assigned to histories). In particular we will do this (both in this specific example and in general) by giving a formalization of the scenario in which all randomization happens in the beginning. We will then apply CDT+GSH as per Definition 9.

So consider the alternative version of the scenario in Figure 6. In this version it is determined at random at the very beginning of the scenario whether the coin comes up Heads or Tails. Thus, when the agent chooses whether to accept the Sunday bet or not, the outcome of the coin flip is already encoded as part of the state. It is easy to verify

[11] This line of argument would not work if we considered a version of the self-sampling assumption, as explained in Section 2.3.2.2, in which the Sunday observation is in its own reference class. As noted in footnote Footnote 10, this is plausibly what Draper and Pust had in mind.

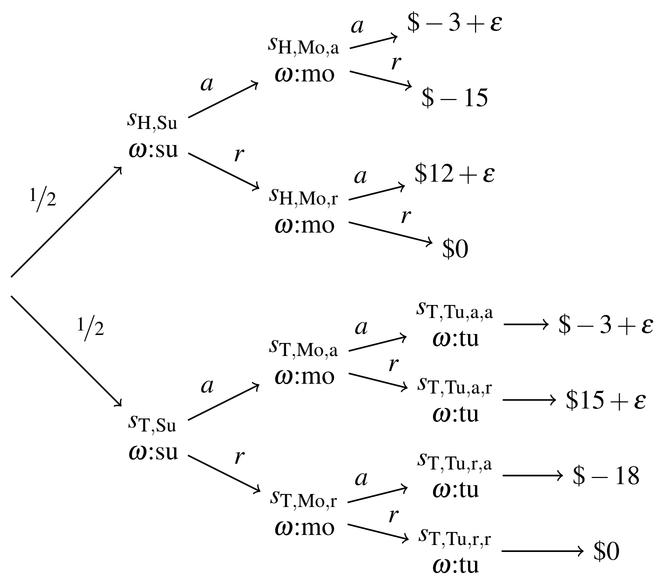

Figure 6: An alternative formalization of Example 5 (adapted from Draper and Pust, 2008, Section 5), in which all randomization (in the environment) happens at the very beginning. In this formalization, the only CDT+GSH-compatible policy is to reject the bet on Sunday and accept the bet on Monday. While this policy is *ex ante* suboptimal, it is not a Dutch book.

that $P_{\text{GSH}}(s_{\text{H,Su}} \mid \text{su}, \pi_{\text{accept}}) = 3/5$ and $P_{\text{GSH}}(s_{\text{T,Su}} \mid \text{su}, \pi_{\text{accept}}) = 2/5$. Thus, in this new scenario, CDT+GSH strictly prefers rejecting the bet on Sunday and thereby avoids the Dutch book.

5.3 Characterization and partial Dutch-book immunity of CDT+GSH*

We now generalize the insight of the previous section. That is, we describe a general theory CDT+GSH* that works by assuming that all (relevant) randomization happens at the beginning of the scenario. We show that this version avoids Dutch books when the agent cannot affect the length of the history. That is, in every scenario that randomizes only in the beginning and in which the agent's actions do not affect the length of the history, there is a non-Dutch-book policy that is compatible with CDT+GSH*. As we will see in the next section, CDT+GSH*'s success (relative to CDT+GSH) does not generalize to all scenarios.

We first define formally what it means for a scenario to randomize only in the beginning.

Definition 11. *We say that a scenario* $\mathscr{E} = (S, S_T, P_0, O, \omega, A, T, u)$ *randomizes only in the beginning if* $T(s' \mid s, a)$ *is* 1 *or* 0 *for all* $s, s' \in S$ *and all* $a \in A$.

For any given scenario, we could now define CDT+GSH* as the application of CDT+GSH to a version of the given scenario that randomizes only in the beginning. To do so, we would need to specify how to turn a given scenario into one that only randomizes in the beginning. (Is this even possible? Are there different ways to do this? Under what conditions do different methods yield the same result?) We do not do this here, because it is intuitive and straightforward but at the same time formally cumbersome.[12] Instead, we will assume that the scenario is already provided in a format that randomizes in the beginning. Thus, for now the discussion of CDT+GSH* is, in effect, a discussion of CDT+GSH as applied to scenarios that only randomize in the beginning.

An immediate question might be how CDT+GSH* is supposed to deal with policy randomization. If having random state transitions causes problems, does having the agent choose randomly cause the same problems? The answer is yes and we will address this in detail in Section 5.4. The positive results in this section assume that either the policy is deterministic, or that the agent's choices do not affect the length of the history (which makes policy randomization unproblematic for GSH).

Definition 12. *Let $\mathscr{E}$ be a scenario that randomizes only in the beginning. We say that history length is choice independent in $\mathscr{E}$ if for each s_0 in $\mathrm{supp}(P_0)$, there is a natural number m s.t. for all π, $\sum_{s_1...s_m} P(s_0 s_1 ... s_m \mid \pi, s_0) = 1$, where the sum is over all histories of length m. For each $s_0 \in \mathrm{supp}(P_0)$, define $\mathrm{len}(s_0)$ to be that number m.*

In words, history length is choice independent in $\mathscr{E}$ if the initial state uniquely determines the length of the history independently of the agent's choices.

Theorem 14. *Let $\mathscr{E}$ be a scenario that randomizes only in the beginning and where history length is choice-independent. Let $\hat{\mathscr{E}}$ be the scenario that is equal to $\mathscr{E}$, except that $\hat{P}_0(s_0) \sim P_0(s_0)/\mathrm{len}(s_0)$. Then any (potentially mixed) policy is CDT+GSH-compatible in $\mathscr{E}$ if and only if it is CDT+GT-compatible in $\hat{\mathscr{E}}$.*

As an immediate consequence of Theorem 14 and Theorem 6, we can characterize CDT+GSH compatibility in such scenarios in terms of the policy derivatives of $Q_\pi(\hat{P}_0)$. Moreover, Theorem 14 is the key to a Dutch book avoidance result for CDT+GSH*.

Corollary 15. *Let $\mathscr{E}$ be a scenario that randomizes only in the beginning and where history length is choice-independent. Then there exists a CDT+GSH-compatible non-Dutch-book policy for $\mathscr{E}$.*

CDT+GSH* and CDT+GT are equivalent via an analogous transformation if we restrict attention to deterministic policies. Let $\mathscr{E}$ be a scenario that randomizes only in the beginning and let π be a deterministic policy. Then notice that each initial state s_0 of $\mathscr{E}$ uniquely and deterministically determines what history will be played. Let $\mathrm{len}_\pi(s_0)$ denote the length of that history.

[12]Similar equivalences will be familiar to computer scientists. For example, the equivalence between the non-determinism definition and the certificate verification definition of the complexity class NP is similarly a result about moving (a different type of) non-determinism to the beginning (e.g., Arora and Barak, 2009, Theorem 2.6).

Proposition 16. *Let $\mathscr{E}$ be a scenario that randomizes only in the beginning and let π be a deterministic policy. Let $\hat{\mathscr{E}}$ be the scenario that is equal to $\mathscr{E}$, except that $\hat{P}_0(s_0) \sim P_0(s_0)/\mathrm{len}_\pi(s_0)$. Then π is CDT+GSH-compatible in $\mathscr{E}$ if and only if π is CDT+GT-compatible in $\hat{\mathscr{E}}$.*

Note that one could equivalently state Theorem 14 and Proposition 16 in terms of dividing the *utilities* rather than the priors by the length of the history. Armstrong (2011) observes a similar connection between GSH and "copy-altruistic average utilitarianism".

5.4 The failure of CDT+GSH* when random choices affect history length

We now ask what happens if we allow the agent's choices to affect the length of the history (as is the case in the absent-minded driver, for instance), and when randomization is needed to solve the scenario. We will argue that CDT+GSH* can be Dutch-booked in those scenarios.

5.4.1 Random choices pose the same problem as random state transitions

As noted earlier, one might suspect that if the agent's choices can affect the history length, policy randomization can cause the same problems as exposed by the Dutch book of Draper and Pust (2008) (Section 5.1). Roughly, the problem of CDT+GSH (as per Definition 9) in Draper and Pust's case is that the environment flips a coin midway through the scenario, the coin flips determines how many many observations the agent will make and CDT+GSH thus assigns different probabilities to the coin flip's outcomes before versus after the coin is flipped. (We have fixed this by assuming that in some sense the coin is flipped at the very beginning such that all decisions are made from the perspective of the coin already having been flipped.) Could the same problem occur in a scenario with deterministic state transitions if the *agent* flips a coin half way through and the outcome determines the length of the history?

The suspicion is true; the answer to the question is yes. That is:

Proposition 17. *There is a scenario $(S, S_T, P_0, O, \omega, A, T, u)$ with the following properties.*

- *The state transitions are deterministic, i.e., for all $a \in A$, $s, s' \in S$, $T(s' \mid s, a)$ is either* 0 *or* 1.
- *All CDT+GSH-compatible policies are Dutch books.*

Because this result is unsurprising and the example needed for proving it is relatively complicated, we only prove this result in Appendix F.2. We will then also show in Appendix F.3 how moving the agent's policy randomization to the beginning of the scenario solves the example given in Appendix F.2.

5.4.2 Viewing random choices as predetermined fails

A natural conclusion might be: CDT+GSH* should not only view all randomization in the environment to happen in the beginning. When its choices affect the length of the

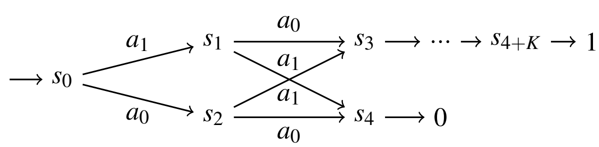

Figure 7: A graphical illustration of Example 6 in our formalism.

history, it should also view its own random choices as determined at the very beginning of the scenario. Put in another way, if the agent intends to use ten coin flips to determine her choices, perhaps she should view the outcome of these ten coin flips as determined at the very beginning of the scenario, and her choices as merely accessing the results of these coin flips. (We will give examples below of how to construct this; for a case in which this construction actually works, see Appendix F.3.)

Unfortunately, this approach does not seem to work in general. In some scenarios, no policy is compatible with CDT+GSH*. In other scenarios, the only policy compatible with CDT+GSH* is a Dutch book. In the following we give examples demonstrating these two claims, respectively. We want to emphasize that we think this is not a failure of our particular formal approach (of moving all randomization to the beginning) but of the very idea behind CDT plus generalized single-halfing. While we will use our formalism (and while we think this formalism is useful), we think that the following counterexample could also be given and discussed without such a formalism.

5.4.2.1 A scenario without a CDT+GSH*-compatible policy We here give a scenario in which no policy is CDT+GSH* compatible if we take CDT+GSH* to move policy randomization to the beginning of the scenario. It turns out that even Absent-Minded Driver is such a scenario. We here give a simpler example in which it is easier to see why CDT+GSH* fails.

Example 6. *First the agent faces a choice between a_0 and a_1 twice. She cannot distinguish between these two situations, retains no memory of whether she has already faced the choice or of what her choice was (if any). Her reward is 1 if she plays a_0 and a_1 exactly once each, and 0 otherwise. If a_1 was chosen exactly once, for a reward of 1, then the agent makes K further observations. ($K = 1$ works in this case, but it is useful to imagine that K is very large.) The agent's choices in these K situations do not affect her final reward – her reward remains 1. Figure 7 illustrates this scenario in our formalism.*

We first offer an intuition for why this scenario spells trouble for CDT+GSH*. Afterward, we will make the argument more formal.

Imagine for now that the agent follows an *ex ante* optimal policy π of mixing uniformly upon observing o_1 (and behaving arbitrarily upon observing o_2). (The argument applies in similar form to non-uniformly mixing upon o_1 as well. We will give the formal argument below for arbitrary mixing.) We will argue that π is not compatible with CDT+GSH*.

First, we consider GSH*'s beliefs given the policy π. Upon seeing o_1, GSH* should believe that it is failing to "anti-coordinate" with itself. That is, according to GSH* the agent should assign high probability (probability approaching 1 as K approaches infinity) that the agent plays either a_0 twice or a_1 twice if it follows the policy π under consideration. After all, in case of success only a small fraction of the agent's observations are o_1, while in case of success the agent only observes o_1. Between these two possibilities (playing a_0 twice and playing a_1 twice), GSH* distributes probability mass equally.

Now CDT enters the picture. For π to be CDT+GSH* compatible, CDT cannot upon observing o_1 favor playing a_0 or a_1 over following π. But with near 50% probability (approaching 50% as $K \to \infty$), playing, for example, a_1 increases utility from 0 to 1, namely in the case where following π leads to failure by playing a_0 twice. With near 50% probability (approaching 50% as $K \to \infty$), the agent would have played a_1 anyway. The probability that playing a_1 makes things worse, meanwhile, is very small (approaching zero as $K \to 0$). Hence, CDT+GSH* recommends a_1 over following π and so π is not CDT+GSH* compatible.

We now make this argument about CDT+GSH* formal. Let π be a policy that plays a_1 with probability p upon observing o_1. For CDT+GSH*, we have to construct a new version of the model of Figure 7 in which all randomization happens at the beginning of the scenario. In this case, this means moving the agent's randomization to the beginning of the scenario. We give this new model in Figure 8. The model has a new action a_π which represents following π but in a way that accesses the result of randomization conducted at the beginning of the scenario. Thus, the new model has four initial states that encode what choices will result from playing a_π. For example, the initial state $s_{0,1,1}$ is the state where playing a_π results in a_1 being played on both observations of o_1. Since p is the agent's probability of playing a_1, $P_0(s_{0,1,1}) = p^2$. Similarly, the initial state $s_{0,1,0}$ encodes the fact that playing a_π will result in a_1 on the first observation of o_1 and in a_0 on the second observation of o_1. Thus, $P_0(s_{0,1,0}) = p(1-p)$.

The second states encode the action that was in fact played in the first state (either by playing a_π or by directly playing a_0 or a_1). They also encode what happens when a_π is played, which is carried over from the initial state. For example, $s_{1,0,1}$ is the state in which a_0 was played in the first state (potentially via playing a_π in $s_{0,0,1}$) and where playing a_π will result in playing a_1.

We now calculate the GSH probabilities in this new model under the assumption that the agent always plays a_π. Leaving out the normalizing constants, the probabilities are

$$
\begin{aligned}
P_{\mathrm{GSH}}(s_{0,1,1} \mid o_1, \text{always } a_\pi) = P_{\mathrm{GSH}}(s_{1,1,1} \mid o_1, \text{always } a_\pi) &\sim \frac{p^2}{2}\\
P_{\mathrm{GSH}}(s_{0,0,0} \mid o_1, \text{always } a_\pi) = P_{\mathrm{GSH}}(s_{1,0,0} \mid o_1, \text{always } a_\pi) &\sim \frac{(1-p)^2}{2}\\
P_{\mathrm{GSH}}(s_{0,0,1} \mid o_1, \text{always } a_\pi) = P_{\mathrm{GSH}}(s_{1,0,1} \mid o_1, \text{always } a_\pi) &\sim \frac{(1-p)p}{2+K}\\
P_{\mathrm{GSH}}(s_{0,1,0} \mid o_1, \text{always } a_\pi) = P_{\mathrm{GSH}}(s_{1,1,0} \mid o_1, \text{always } a_\pi) &\sim \frac{p(1-p)}{2+K}.
\end{aligned}
$$

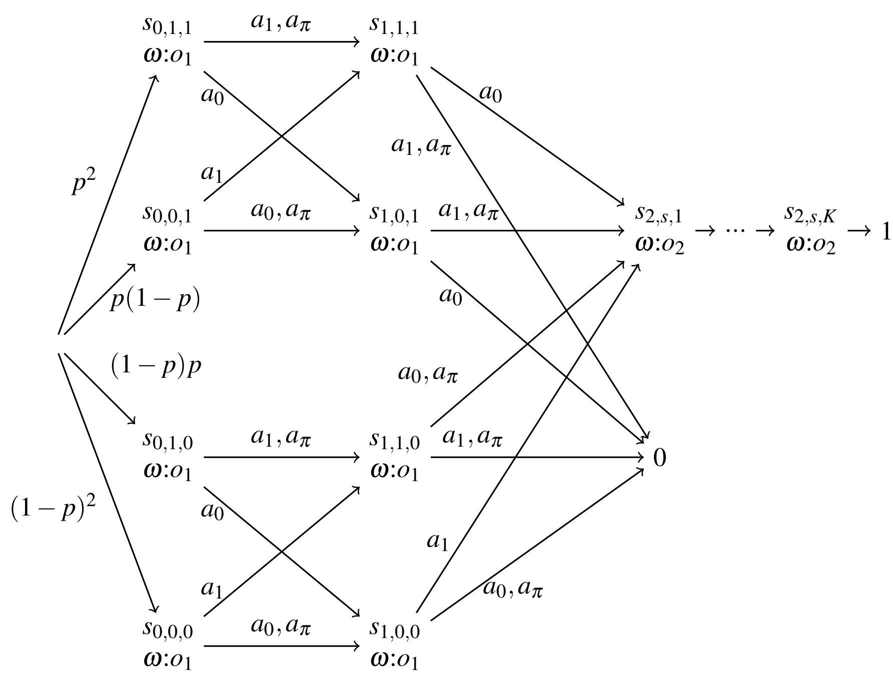

Figure 8: A alternative graphical illustration of Example 6 in our formalism. In this model, the probabilistic policy π (of choosing a_1 with probability p in o_1) can be followed by taking the action a_π deterministically.

Normalizing the probabilities, it is easy to see that

$$P_{\text{GSH}}(\{s_{0,1,1}, s_{1,1,1}, s_{0,0,0}, s_{1,0,0}\} \mid o_1, \text{always } a_\pi) \to 1 \text{ as } K \to \infty.$$

This is a more formal version of our earlier claim that GSH* assigns high probability to failure.

Using this probability distribution for the model in Figure 8, we can now show that the policy π is not CDT+GSH* compatible by showing that the deterministic policy of always playing a_π is not CDT+GSH compatible in the new model. Assume without loss of generality that $p \leq 1/2$, i.e., that a_1 is chosen with probability at most $1/2$ upon observing o_1. (The other case can be handled analogously.) Omitting normalizing constants, the CDT+GSH expected utility of playing a_1 upon observing o_1 is

$$\sum_{s \in S} P_{\text{GSH}}(s \mid o_1, \text{always } a_\pi) Q_{a_\pi}(s, a_1) \sim 2\overbrace{\frac{(1-p)^2}{2}}^{\geq 1/4} - 2\overbrace{\frac{(1-p)p}{2+K}}^{\leq 1} \geq \frac{1}{4} - \frac{2}{2+K}.$$

Meanwhile, the expected utility of going along with a_π under omission of the same normalizing constant is

$$\sum_{s \in S} P_{\text{GSH}}(s \mid o_1, \text{always } a_\pi) Q_{a_\pi}(s, a_\pi) \sim 4\overbrace{\frac{(1-p)p}{2+K}}^{\leq 1} \leq \frac{4}{2+K}.$$

For large enough K (specifically, $K \geq 11$), $1/2 - 2/2+K > 4/2+K$. That is, if we make K large enough, then upon observing a_π, CDT+GSH strictly prefers (at least) one of the available actions (a_1 if $p \leq 1/2$, a_2 if $p \geq 1/2$) over playing a_π. Thus, no policy is CDT+GSH* compatible.

5.4.2.2 A Dutch book against CDT+GSH* We now give a scenario in which the only policy compatible with CDT+GSH* is a Dutch book policy. We give an informal exposition here. We give a more formal analysis in Appendix F.4.

Example 7. *First the agent faces a choice between a_0 and a_1 three times. She cannot distinguish between these three situations, retains no memory of how often she has already faced the choice or of what her choices were. Her rewards are determined from the number of times she chooses a_1 in these situations according to the following table.*

Number of times a_1 is played	*Reward*
0	0
1	1
2	-1
3	$-\varepsilon$

Here, ε is some small but positive number, e.g., $\varepsilon = 1/100$. If a_1 was chosen exactly once, for a reward of 1*, then the agent faces the same decision problem between a_0 and a_1 another K times (for some large K). Importantly, the agent's choices in these K situations do not affect her final reward – her reward remains* 1*.*

Note first that in this scenario, the agent's choices affect the length of the scenario. Hence, this is, of course, not a counterexample to Corollary 15.

We now argue that Example 7 is a Dutch Book against CDT+GSH*. Clearly, the policy of always playing a_1 is CDT+GSH* compatible. It is left to argue that no other policy is CDT+GSH* compatible. To do so, we will argue that regardless of what policy π the agent follows, CDT+GSH* recommends deviating to play a_1. In short, the CDT+GSH* agent believes, regardless of its policy, that conditional on the choice between a_1 and a_π mattering at all, she is unlikely (specifically with probability approaching 0 as K goes to ∞) to be on track to play a_1 exactly once if she plays a_π. Instead, conditional on her choices mattering, following the policy (playing a_π) will likely lead to a_1 being played either 0 or 2 times. This is because for all policies π, the (*ex ante*, non-self-locating) probability of playing a_1 exactly once is never much bigger than the probability of playing a_1 zero or two times. Since the former branch contains many additional inconsequential decision situations, the CDT+GSH* agent believes she is very likely (with probability approaching 1 as $K \to \infty$) to be in a branch of the latter type, conditional on her choice mattering at all. Given this belief, CDT recommends playing a_1.

5.5 A new Dutch book against *evidential* decision theory plus GSH*

As noted at the beginning of Section 5.2, *evidential* decision theory plus GSH, too, avoids Draper and Pust's Dutch book if we let the randomization take place at the beginning of the scenario. Perhaps some version of EDT+GSH* can avoid Dutch books more generally than *C*DT+GSH*? It turns out that it is easy to construct Dutch book scenarios in which the strategy that partially saves CDT+GSH cannot even partially save EDT+GSH. In particular, there are Dutch book scenarios where actions do not affect the agent's future observations.

Proposition 18. *There is a scenario that only randomizes in the beginning, where the length of histories is choice independent and where the only policy compatible with EDT+GSH is a (deterministic) Dutch book policy.*

Example 8. *Let* $\epsilon > 0$*. On Sunday, Beauty is offered a bet that wins* $\$1-\epsilon$ *if the coin comes up Heads and loses* $\$1-2\epsilon$ *if the coin comes up Tails. The researchers then put Beauty to sleep and flip a fair coin. If the coin comes up Tails, then Beauty is awoken once on Monday and again on Tuesday (without being told what day it is). If the coin comes up Heads, then Beauty is awoken once on Monday (without being told what day it is). However, she is awoken again on Tuesday and told that it is Tuesday and that the coin has come up Heads. In Tails–Monday, Tails–Tuesday and Heads–Monday, Beauty is offered a choice between accepting and rejecting a bet. In the Tails branch, Beauty has to accept twice in order to reap the reward of the bet. This second bet loses* $\$1$ *if the coin comes up Heads and wins* $\$1-3\epsilon$ *if the coin comes up Tails. For a graphical description of this scenario in our formalism, see Figure 9.*

Since there are now equally many observations regardless of the agent's choices or the outcome of the coin flip, it seems clear that on Sunday Beauty should believe

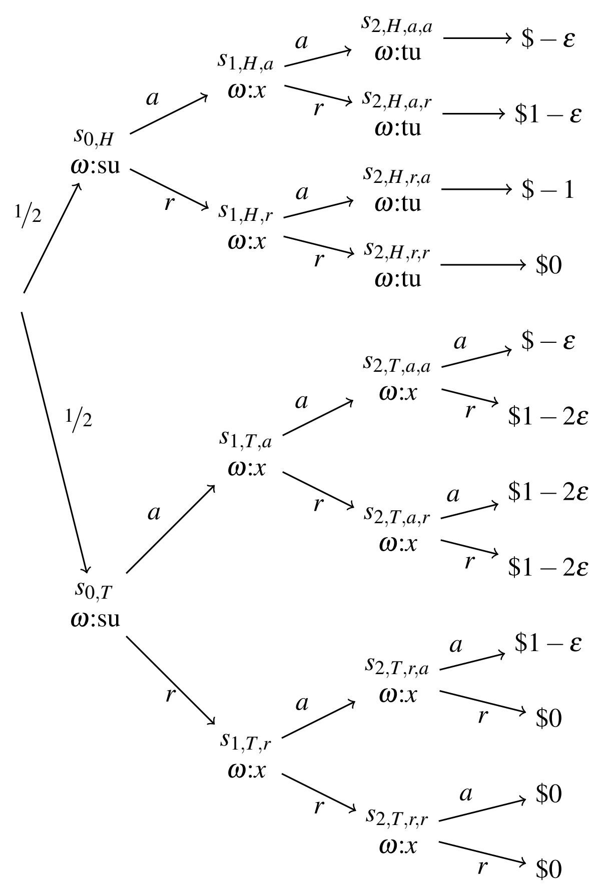

Figure 9: A graphical formalization of Example 8.

the coin flip outcome to be Heads/Tails with probability 50%. Thus, EDT+GSH requires accepting the bet on Sunday (in agreement with the *ex ante* view and all other combinations of EDT and CDT with GT, GDH and GSH). Upon waking up and being offered the second bet, GSH assigns equal probabilities to the states Tails–Monday, Tails–Tuesday and Heads–Monday and thus a probability of $2/3$ to Tails. If Beauty also uses EDT, then accepting with probability p increases her GSH-expected utility by $2/3 p^2(\$1-3\varepsilon) - 1/3 p(\$1-2\varepsilon)$ relative to not accepting. For small enough (but still positive) ε, the only global maximum of this function is $p=1$. Hence, EDT+GSH requires accepting the second bet also. But of course, accepting both bets yields a certain payoff of $-\varepsilon$, while always rejecting yields a certain payoff of 0.

It is worth noting that in the single-observation case, GSH is equivalent to GDH and therefore, by Corollary 10, the EDT+GSH-compatible policies are exactly the *ex ante* optimal policies.

Proposition 19. *Consider any scenario with only one observation (i.e., $|O|=1$) and* *any set of policies $\Pi \subseteq \Delta(A)$. Then policy $\pi \in \Pi$ is ex ante optimal in Π if and only if π is compatible with EDT+GSH restricted to Π.*

6 Discussion

6.1 CDT's need for randomization

Our main positive compatibility results for CDT+GT and CDT+GSH(*) (i.e., Corollary 7 (Piccione and Rubinstein, 1997) and Corollary 15) require unrestricted randomization (i.e., require that $\Pi = \Delta(A)^O$), in contrast to our positive result for EDT+GDH (specifically Corollary 10). Indeed, this ability to randomize is necessary for CDT. There are scenarios with imperfect recall (e.g., the absent-minded driver) in which no deterministic policy is compatible with any natural version of CDT+GT. Furthermore, there are scenarios with imperfect recall in which the only deterministic policy consistent with CDT+GT or CDT+GSH* is a Dutch book policy:

Proposition 20. *There is a scenario in which history length is choice independent and the only deterministic CDT+GT/CDT+GSH-compatible policy is a Dutch book policy.*

We prove Proposition 20 with the following variant of Conitzer's (2015) "Three Awakenings":

Example 9. *The agent faces a choice between a_0 and a_1 three times. She cannot distinguish between these three situations, retains no memory of how often she has already faced the choice or of what her choices were. Each choice of a_1 decreases her reward by 1. However, if she chooses a_1 exactly once or all three times, her reward is increased by 2. Thus, if she never chooses a_1, her reward is 0; if she chooses a_1 exactly once, her reward is 1; if she chooses a_1 exactly twice, her reward is -2; and if she chooses a_1 all three times, her reward is -1.*

The *ex ante* optimal randomized policy is to play a_1 with probability $1/2 - 1/(2\sqrt{2}) \approx 0.15$, which (by Corollaries 7 and 10) is compatible with CDT+GT and EDT+GDH. The optimal deterministic policy is to always play a_0. This is compatible with EDT+GDH restricted to deterministic policies (by Corollary 10). However, it is clearly not consistent with CDT+GT; given that the agent otherwise uses the policy of playing a_0 with probability 1, it would be better to play a_1 (once). However, the other deterministic policy of always playing a_1 *is* CDT+GT compatible; given that the agent follows this policy, choosing a_2 once decreases the reward from -2 to -1. But this policy voluntarily accepts a negative reward, when another deterministic policy achieves a reward of 0.

How CDT and related theories require and deal with randomization has been discussed in other contexts (e.g. Richter, 1984; Harper, 1986; Skyrms, 1986; Levinstein and Soares, 2020; Oesterheld and Conitzer, 2021, Section IV.1).[13] It is outside the

[13]Instead of relying on true randomization, some authors have proposed approaches involving subjective uncertainty about one's own choice (Arntzenius, 2008). These are meant to specify how a CDT agent should deal with cases in which true randomization is unavailable and no policy is compatible with CDT. However, agents implementing such approaches will typically still be subject to voluntary loss of money (e.g., in Example 9 or the Adversarial Offer case of Oesterheld and Conitzer, 2021).

scope of this paper to judge whether the assumption of being able to randomize is reasonable or what conclusions can be drawn from CDT's failure in the absence of the ability to randomize.

6.2 Conitzer's Dutch book against evidential decision theorists

Conitzer (2015a) claims to provide a Dutch book involving imperfect recall against EDT. Specifically, he provides a scenario in which he claims, translated to our terminology,

- there is a deterministic policy π_0 with a certain payoff of 0;
- only one policy $\tilde{\pi}$ is compatible with EDT (regardless of self-locating beliefs) and $\tilde{\pi}$ has a negative payoff with certainty.

This seems to contradict our Corollary 10. What is going on?

The reason why our analyses differ is that Conitzer considers a version of EDT that differs subtly from the one we define in Section 2.4.1. We first describe the difference abstractly and then illustrate it using an example. First, recall the following from our definition. We have some given policy π. Now the agent observes o. When evaluating the value of some distribution $\alpha \in \Delta(A)$ over actions to play in response to o, the EDT(+GDH) agent considers that on all other instances of observing o (including past ones), she also will choose and has chosen α. We imagine that when assessing $o \to \alpha$, she imagines that $o \to \alpha$ gives no evidence about her choice in any *other* situation. Conitzer gives a case in which it is very plausible that $o_1 \to \alpha$ also implies that $o_2 \to \alpha$ for two different observations o_1, o_2, because o_1 and o_2 are symmetric in the game.

We now illustrate this using an example. Because Conitzer's case is somewhat complicated, we give a simpler example that illustrates the same mechanism. The cost of the simplification is that in our example Conitzer's interpretation of EDT+GDH uniquely selects a policy that is merely *ex ante* suboptimal, as opposed to being a Dutch book policy.

Example 10. *We use the same three equiprobable branches as in Example 11. Again, at each observation of x or y, the agent is offered a bet on whether branch Z is realized. However, contrary to the previous version, the two offers in branch Z are now made and accepted independently. That is, if the agent accepts twice in branch Z, then she wins the bet twice; and if she accepts once, she wins the bet once. Also, the bet is now at somewhat worse than even odds. Specifically, it pays* $2/3$ *in branch Z and pays* -1 *in branch X/Y. The game is represented graphically in Figure 10.*

Clearly, the unique *ex ante* optimal policy for this scenario is to always reject for a certain payoff of 0.

Conitzer claims, translated to our example, that EDT+GDH recommends accepting the bet. The key idea is that an EDT agent should take her choice upon observing x as definitive evidence about her choice upon observing y (and *vice versa*). This is due to the combination of two reasons.

1. The observations x and y and the bets made in them are symmetric.
2. The bets are resolved independently, the payoffs are additive between the bets. Thus, if, say, the agent were hard-wired to make a particular choice in y, it seems that a choice in x can be made without knowing what the choice would be in y.

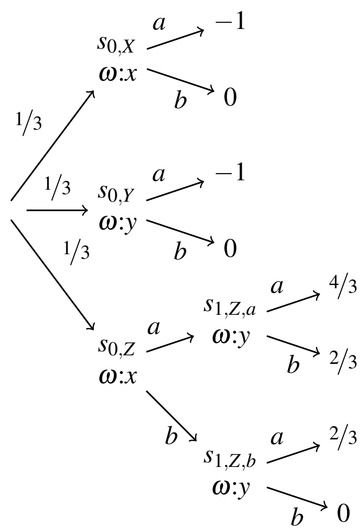

Figure 10: A graphical formalization of Example 10.

Although both also apply to Conitzer's original case, note that Conitzer only makes the symmetry point. However, symmetry between two observations alone arguably does not imply that an agent needs to choose the same for both observations, see Appendix C.

If she takes her choice in x as conclusive evidence of her choice in y, then her expected utility calculation changes. Again, branch X and Z are equally probable upon observing x. By accepting, the agent decreases her reward in branch X relative to not accepting by 1. But in branch Z, she now increases her utility by $4/3$ (not just $2/3$), because if she is in branch Z and she accepts, she accepts twice for a reward of $4/3$. Whereas, if she is in branch Z and she rejects, she rejects twice for a reward of 0. A gain of $4/3$ outweighs an equally probable loss of 1. Thus, the EDT agent accepts the bet in x and by the same argument also in y.

The general problem of the EDT agent as considered in Conitzer's argument is a discrepancy between entities that the agent can "evidentially control", and entities about which the agent thinks, "I might be this entity". In Example 10, upon observing x, the agent (as considered in Conitzer's argument) believes that she has evidential control over her choice for y in branch Z, but she does not think that she might currently be observing y (in branch Z). Interestingly, this discrepancy problem is common in Newcomb-like problems (without imperfect recall) and results in both CDT and EDT choosing *ex ante*-suboptimal policies. Newcomb's problem (Nozick, 1969) itself is an example in which CDT's choice (two-boxing) is an *ex ante* suboptimal policy. Now imagine that the way that the predictor in Newcomb's problem arrives at its predictions by creating a precise copy of the agent and having the copy make a choice between one- and two-boxing. Arguably, the CDT agent should then assign equal probability to being the copy versus the original. If (as assumed in this paper) the agent's goal (maximizing payoff) is independent of whether he is the original or the copy, CDT recommends one-boxing.[14] Note that the formalism of Section 2.1 can only model the latter version of Newcomb's problem. Similarly, one can give Newcomb-like cases in which EDT does not give the *ex ante* optimal policy[15], and we can similarly "save" EDT by having the agent identify with anything in the environment that can be used to predict the agent.[16]

6.3 The multiplicity of compatible policies

As we have seen (e.g., in Examples 4 and 9) and as other authors have also discussed (e.g. Aumann, Hart, and Perry, 1997; Korzukhin, 2020), there are scenarios in which multiple policies are compatible with the theories defined in Section 2.4. We and others have showed that the *ex ante* optimal policies are among the CDT+GT and EDT+GDH compatible policies. We have also showed much weaker and more nuanced but vaguely analogous results for CDT+GSH*, but for simplicity we will in the following focus on *ex ante* optimal policies and CDT+GT and EDT+GDH. We should draw some satisfaction from these results. For example, they show that we do not have to choose between *ex ante* optimality and *de se* rational choice.

However, we might also wonder whether *purely de se* reasoning on its own can – even in the face of a multiplicity of compatible policies – arrive at an *ex ante* optimal policy, without ever explicitly assuming an *ex ante* perspective. This paper has not discussed this question much so far; and we are not aware of much work that is trying to propose solutions to this problem.

Like a few other issues (the need for independent randomization as discussed in Section 6.1 and infinite histories as discussed in Appendix B), the multiplicity of compatible policies is a bigger problem for CDT than it is for EDT. First, by Corollary 11, in the single-observation case ($|O| = 1$) specifically, EDT+GDH faces no multiplicity problem at all. CDT+GT, on the other hand, faces such a problem even in the single-observation case (as demonstrated in Example 4 and also shown by, e.g., Aumann, Hart, and Perry, 1997, Section 5, and Korzukhin, 2020). This matters especially if we believe that the single-observation case is especially important. For example, in the literature on Newcomb-like problems, it has sometimes been argued that we should limit our expectations of CDT and EDT to individual decisions, and that we should not expect them to make good recommendations when applied to multiple different decision situations (see Oesterheld and Conitzer, 2021, Section IV.3 and references therein). A discussion of this issue is beyond the scope of this paper. Less importantly, by Corollary 12 the set of EDT+GSH compatible policies is a subset and in many cases a strict

[14] As far as we are aware, this argument for why CDT agents might one-box in Newcomb's problem has not been discussed in much detail in the literature. However, it is briefly mentioned by Neal (2006, p. 12f.); as well as various various blog posts (e.g., Aaronson, 2005; Taylor, 2016).

[15] To our knowledge, the oldest such case is a version of Newcomb's problem in which both boxes are transparent (first proposed, we believe, by Gibbard and Harper, 1981, Section 10; for further discussion, see Gauthier, 1989; Drescher, 2006, Section 6.2; Arntzenius, 2008, Section 7; Meacham, 2010, Section 3.2.2). Other such examples include Parfit's (1984) hitchhiker (Barnes, 1997), XOR Blackmail (Levinstein and Soares, 2020, Section 2) and Yankees vs. Red Sox (Arntzenius, 2008; Ahmed and Price, 2012, pp. 22-23).

[16] Again, we are not aware of any detailed discussion of this idea in the literature, but again the point has been made at least in a blog post (Treutlein, 2017).

subset of the set of CDT+GT compatible policies.

To close, we would like to join Aumann, Hart, and Perry, 1997 and Korzukhin (2020) in raising awareness for the problem of the multiplicity of compatible policies by giving a particular type of result. Roughly, at each decision perspective $o \in O$, an agent can not only evaluate her choices at that decision point; she can also evaluate entire policies π, most naturally by calculating expected utilities $\mathrm{EU}_{\mathrm{GSH/GDH/GT}}(\pi, o)$ and comparing them across policies π. The different decision perspectives might disagree in their preferences over entire policies. So in particular if in each o the agent played from the, say, CDT+GT-compatible policy that maximizes $\mathrm{EU}_{\mathrm{GT}}(\pi, o)$, then the agent will in general not follow a CDT+GT-compatible policy. (We give an example in Appendix H.) In general, it is unclear how rational agents resolve such disagreement across decision perspectives. This difficulty resembles the difficulty of equilibrium selection in game theory. However, we might expect that, for example, if all decision perspectives agree that some CDT+GT-compatible policy π is better than another CDT+GT-compatible policy policy π', then a CDT+GT agent would not follow π'. This resembles the use of Pareto efficiency as a goal in multi-agent interactions. The idea is more natural in single-player scenarios of imperfect recall, however, since we might expect the different decision perspectives of a single player to be better able to coordinate. For a success story of this approach, consider Example 4. In this example, both (accept, defuse) and (reject, not defuse) are consistent with CDT+GT and EDT+GDH. However, at both decision points, Alice prefers (reject, not defuse) over (accept, defuse) (regardless of whether she uses $\mathrm{EU}_{\mathrm{GDH}}$, $\mathrm{EU}_{\mathrm{GSH}}$ or $\mathrm{EU}_{\mathrm{GT}}$ to judge policies). Thus, without using any *ex ante* perspective, *de se* decision theories can avoid the certain loss in this scenario. We now show that such an approach can be used with GSH in general. Unfortunately, the following result shows that such reasoning can lead the agent badly astray, as the following result shows.

Theorem 21. *There is a scenario $\mathscr{E}$ with the following properties.*

- *$\mathscr{E}$ only randomizes in the beginning and the agent's choices do not affect her future observations. (In particular, history length is choice independent.)*
- *There is a CDT+GT-, CDT+GSH- and EDT+GDH-compatible deterministic Dutch book policy $\tilde{\pi}$.*
- *For all EDT+GDH/CDT+GT/CDT+GSH-compatible policies π other than $\tilde{\pi}$ and all observations o,* $\mathrm{EU}_{\mathrm{GDH/GT/GSH}}(\tilde{\pi}, o) > \mathrm{EU}_{\mathrm{GDH/GT/GSH}}(\pi, o)$.

Note that the first item means that this scenario is relatively unproblematic for CDT+GSH (see Theorem 14).

Since the *ex ante* optimal policy is (by Corollaries 7 and 10) EDT+GDH- and CDT+GT-compatible, $\mathrm{EU}_{\mathrm{GDH}}$ and $\mathrm{EU}_{\mathrm{GT}}$ prefer a Dutch book policy over the *ex ante* optimal policy from all decision perspective in this scenario.

Aumann, Hart, and Perry (1997, Section 5) provide a single-observation scenario with a similar property w.r.t. only CDT+GT. By removing the Sunday bet from Korzukhin's (2020) scenario, we obtain another single-observation scenario with a similar property w.r.t. CDT+GT. (Corollary 11 implies that multiple observations are necessary to obtain this kind of result for EDT+GDH.)

For simplicity, we first give a scenario that proves the theorem only for EDT+GDH and CDT+GT (and not for CDT+GSH*). In Appendix I, we then extend the scenario to also apply to CDT+GSH*.

Example 11. *At the beginning, the scenario randomizes uniformly between three possibilities:*

X) The agent observes x (once) and the scenario ends.

Y) The agent observes y (once) and the scenario ends.

Z) The agent observes x (once), then y (once), and then the scenario ends.

Upon observing x or y, the agent chooses from three actions: bet, pay, and pass. By choosing bet, they accept a bet on being in branch X or Y at slightly better than even odds, specifically, for each time they bet, they obtain 1 *if branch X or Y is realized and they lose* $2/3$ *if branch Z is realized. By choosing pay, they lose some small amount* $\varepsilon > 0$*. However, if branch Z is realized and the agent chooses to pay exactly once, they end up with a payoff of* -100*. Choosing to pass has no consequences in and of itself. A graphical description of this problem in our formalism is given in Figure 11.*

In this game, the two observations x and y are symmetric. This is done to keep our descriptions and arguments brief. It is inessential and all the same points apply if we introduce a minor asymmetry, e.g., if we increased the probability of branch X by 1% and correspondingly decreased the probability of branch Y by 1%. We mention this because symmetry of observations has been a central feature in (alleged) counterexamples given in previous work (see Section 6.2).

We now show that Example 11 has the properties claimed in Theorem 21. First notice that always passing ensures a non-negative reward. Our compatible Dutch book policy $\tilde{\pi}$ is the one that pays in both situations. Its expected utilities are $\mathrm{EU}_{\mathrm{GDH/GT}}(\tilde{\pi}, x) = \mathrm{EU}_{\mathrm{GDH/GT}}(\tilde{\pi}, y) = -3\varepsilon/2$. (Since no observation is ever made twice in any history, GDH and GT probabilities coincide in this problem.)

We now go through the list of other CDT+GT-compatible policies. (By Corollary 12, all EDT+GDH-compatible policies are CDT+GT-compatible. Hence, it is enough to consider the CDT+GT-compatible policies.) We start with the only other deterministic compatible policy π_{bet}, which is to bet in both x and y. This is also the *ex ante* optimal policy. However, upon observing x or y, the GDH/GT expected utility is $\mathrm{EU}_{\mathrm{GDH/GT}}(\pi_{\mathrm{bet}}, x) = \mathrm{EU}_{\mathrm{GDH/GT}}(\pi_{\mathrm{bet}}, y) = 1/2 \cdot 1 + 1/2 \cdot (-4/3) = -1/6$, which is less than $-3\varepsilon/2$ for small enough ε.

What mixed compatible policies are there? To answer this question, notice first that CDT+GT never recommends passing. Regardless of what policy the agent uses in the other decision situation, it is always better (as judged by CDT+GT) to bet than to pass. Hence, we are left to find a policy that randomizes between bet and pay in at least one of x and y. For a randomized policy to be CDT+GT-compatible, it must induce indifference between bet and pay. This means that the agent must randomize in *both* decision situations (x and y) (since without randomization in x/y the agent is not indifferent in y/x). By symmetry between x and y, the distribution needed for indifference is the same in x and y. Thus, the third and last CDT+GT-compatible policy is some π_p that bets with probability p and pays with probability $1-p$ in both x and y. It is easy to see that $\mathrm{EU}_{\mathrm{GDH/GT}}(\pi_p, x/y)$ can be written as a convex distribution of the analogous expected utilities for four different deterministic policies, namely the four different policies that map x and y to bet and pay.[17] It is easy to see that from all decision perspectives the utility of paying in both x and y is strictly greater than the utility of the three other deterministic policies. Thus, it is also greater than their convex combination.

References

Aaronson, Scott (2005). *Dude, it's like you read my mind*. URL: `https://www.scottaaronson.com/blog/?p=30`.

Ahmed, Arif and Huw Price (2012). "Arntzenius on 'Why ain'cha rich?'" In: *Erkenntnis* 77.1, pp. 15–30. DOI: `10.1007/s10670-011-9355-2`.

Armstrong, Stuart (2011). *Anthropic decision theory*. URL: `https://arxiv.org/abs/1110.6437`.

Arntzenius, Frank (2002). "Reflections on Sleeping Beauty". In: *Analysis* 62.1, pp. 53–62. DOI: `10.2307/3329069`.

— (2008). "No Regrets, or: Edith Piaf Revamps Decision Theory". In: *Erkenntnis* 68, pp. 277–297. DOI: `10.1007/s10670-007-9084-8`.

Arora, Sanjeev and Boaz Barak (2009). *Computational Complexity – A Modern Approach*. Cambridge University Press.

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

Barnes, R. Eric (1997). "Rationality, Dispositions, and the Newcomb Paradox". In: *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 88.1, pp. 1–28. URL: `https://www.jstor.org/stable/4320784`.

Binmore, Ken (1997). "A Note On Imperfect Recall". In: *Understanding Strategic Interaction*. Ed. by W. Albers et al. Springer. DOI: `10.1007/978-3-642-60495-9_5`.

Bostrom, Nick (2003). "Are We Living in a Computer Simulation?" In: *Philosophical Quarterly* 53.211, pp. 243–255. DOI: `10.1111/1467-9213.00309`.

— (2010). *Anthropic Bias: Observation Selection Effects in Science and Philosophy*. Ed. by Robert Nozick. Studies in Philosophy. Routledge.

[17] In this specific case, this can be formally verified as follows:

$$\begin{aligned}
\mathrm{EU_{GT}}(\pi_p, x) &= \frac{1}{2}(p + (1-p)(-\varepsilon)) \\
&\quad + \frac{1}{2}(p^2(-4/3) + 2p(1-p)(-100) + (1-p)^2(-2\varepsilon)) \\
&= \frac{1}{2}(p^2 + p(1-p) + p(1-p)(-\varepsilon) + (1-p)^2(-\varepsilon)) \\
&\quad + \frac{1}{2}(p^2(-4/3) + 2p(1-p)(-100) + (1-p)^2(-2\varepsilon)) \\
&= p^2(1/2 + 1/2(-4/3)) + p(1-p)(1/2 \cdot 1 + 1/2(-100)) \\
&\quad + (1-p)p(1/2(-\varepsilon) + 1/2(-100)) + (1-p)^2(1/2(-\varepsilon) + 1/2(-2\varepsilon)) \\
&= p^2\mathrm{EU_{GT}}(\pi_{\mathrm{bet}}, x) + p(1-p)\mathrm{EU_{GT}}((x \mapsto \text{bet}, y \mapsto \text{pay}), x) \\
&\quad + (1-p)p\mathrm{EU_{GT}}((x \mapsto \text{pay}, y \mapsto \text{bet}), x) + (1-p)^2\mathrm{EU_{GT}}(\tilde{\pi}, x).
\end{aligned}$$

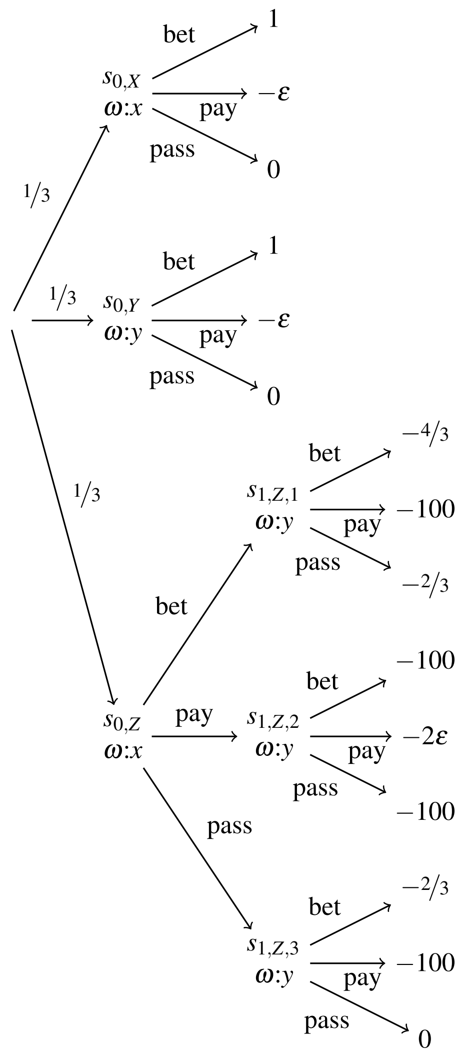

Figure 11: A graphical formalization of Example 11.

Briggs, Rachael (2010). “Putting a value on Beauty”. In: *Oxford Studies in Epistemology*. Vol. 3. Oxford University Press, pp. 3–24. URL: http://joelvelasco.net/teaching/3865/briggs10-puttingavalueonbeauty.pdf.

Carter, Brandon (1983). “The anthropic principle and its implications for biological evolution”. In: *Philosophical Transactions of the Royal Society A* 310.1512. DOI: 10.1098/rsta.1983.0096.

Čermák, Jiří, Branislav Bošanský, and Viliam Lisý (2017). “An Algorithm for Constructing and Solving Imperfect Recall Abstractions of Large Extensive-Form Games”. In: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)*, pp. 936–942. URL: https://www.ijcai.org/proceedings/2017/0130.pdf.

Cesa-Bianchi, Nicolò and Gábor Lugosi (2006). *Prediction, Learning, and Games*. Cambridge University Press.

Ćirković, Milan M., Anders Sandberg, and Nick Bostrom (2010). “Anthropic Shadow: Observation Selection Effects and Human Extinction Risks”. In: *Risk Analysis* 30.10, pp. 1495–1506. DOI: 10.1111/j.1539-6924.2010.01460.x.

Conitzer, Vincent (2015a). “A Dutch book against sleeping beauties who are evidential decision theorists”. In: *Synthese* 192.9, pp. 2887–2899. DOI: 10.1007/s11229-015-0691-7.

— (2015b). “Can rational choice guide us to correct *de se* beliefs?” In: *Synthese* 192, pp. 4107–4119. DOI: 10.1007/s11229-015-0737-x.

— (2019). “Designing Preferences, Beliefs, and Identities for Artificial Intelligence”. In: *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19) Senior Member / Blue Sky Track*.

Detwarasiti, Apiruk and Ross D. Shachter (2005). “Influence Diagrams for Team Decision Analysis”. In: *Decision Analysis* 2.4, pp. 183–244. DOI: 10.1287/deca.1050.0047.

Draper, Kai and Joel Pust (2008). “Diachronic Dutch Books and Sleeping Beauty”. In: *Synthese* 164, pp. 281–287. DOI: 10.1007/s11229-007-9226-1.

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

Elga, Adam (2000). “Self-locating belief and the Sleeping Beauty problem”. In: *Analysis* 60.2, pp. 143–147.

Emmons, Scott et al. (2021). “Symmetry, Equilibria, and Robustness in Common-Payoff Games”. URL: https://preflib.github.io/gaiw2021/papers/GAIW_2021_paper_32.pdf.

Friederich, Simon (2021). “Fine-Tuning”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2021. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/archives/win2021/entries/fine-tuning/.

Ganzfried, Sam and Tuomas Sandholm (2014). “Potential-aware imperfect-recall abstraction with earth mover’s distance in imperfect-information games”. In: *AAAI’14: Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, pp. 682–690.

Gauthier, David (1989). “In the Neighbourhood of the Newcomb-Predictor (Reflections on Rationality)”. In: *Proceedings of the Aristotelian Society, New Series, 1988–1989*. Vol. 89, pp. 179–194.

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.

Harper, William L. (1986). “Mixed Strategies and Ratifiability in Causal Decision Theory”. In: *Erkenntnis* 24.1, pp. 25–36. URL: https://www.jstor.org/stable/20006545.

Hitchcock, Christopher (2004). “Beauty and the Bets”. In: *Synthese* 139, pp. 405–420. DOI: 10.1023/B:SYNT.0000024889.29125.c0.

Hu, Hengyuan et al. (2020). ““Other-Play” for Zero-Shot Coordination”. In: *Proceedings of the 37th International Conference on Machine Learning (ICML’20)*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, pp. 4399–4410. URL: https://proceedings.mlr.press/v119/hu20a.html.

Isbell, J. R. (1957). “Finitary Games”. In: *Contributions to the Theory of Games, Volume 3*. Princeton University Press, pp. 79–96.

Jeffrey, Richard C. (1983). *The Logic of Decision*. 2nd ed. First published in 1965. University of Chicago Press.

Korzukhin, Theodore (2020). “A Dutch book for CDT thirders”. In: *Synthese*. DOI: 10.1007/s11229-020-02841-7.

Levinstein, Benjamin A. and Nate Soares (2020). “Cheating Death in Damascus”. In: *The Journal of Philosophy* 117.5, pp. 237–266. DOI: 10.5840/jphil2020117516.

Lewis, David (2001). “Sleeping Beauty: reply to Elga”. In: *Analysis* 61.3, pp. 171–176. URL: www.jstor.org/stable/3329230.

Li, Yanjie, Baoqun Yin, and Hongsheng Xi (2011). “Finding optimal memoryless policies of POMDPs under the expected average reward criterion”. In: *European Journal of Operational Research* 211, pp. 556–567. DOI: 10.1016/j.ejor.2010.12.014.

Meacham, Christopher J. G. (2008). “Sleeping beauty and the dynamics of de se beliefs”. In: *Philosophical Studies* 138, pp. 245–269. DOI: 10.1007/s11098-006-9036-1.

— (2010). “Binding and its consequences”. In: *Philosophical Studies* 149.1, pp. 49–71. DOI: 10.1007/s11098-010-9539-7.

Neal, Radford M. (2006). *Puzzles of Anthropic Reasoning Resolved Using Full Non-indexical Conditioning*. Tech. rep. 0607. Department of Statistics, University of Toronto. URL: https://arxiv.org/pdf/math/0608592v1.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 (2017). *Decision Theory and the Irrelevance of Impossible Outcomes*. URL: https://casparoesterheld.com/2017/01/17/decision-theory-and-the-irrelevance-of-impossible-outcomes/.

Oesterheld, Caspar and Vincent Conitzer (2021). "Extracting Money from Causal Decision Theorists". In: *The Philosophical Quarterly*. DOI: https://doi.org/10.1093/pq/pqaa086.

Parfit, Derek (1984). *Reasons and Persons*. Oxford 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. DOI: 10.1006/game.1997.0536.

Richter, Reed (1984). "Rationality revisited". In: *Australasian Journal of Philosophy* 62.4, pp. 392–403. DOI: 10.1080/00048408412341601.

Schwarz, Wolfgang (2015). "Lost memories and useless coins: revisiting the absent-minded driver". In: *Synthese* 192, pp. 3011–3036. DOI: 10.1007/s11229-015-0699-z.

Skyrms, Brian (1986). "Deliberational Equilibria". In: *Topoi* 5, pp. 59–67. DOI: 10.1007/BF00137830.

— (1990). *The Dynamics of Rational Deliberation*. Harvard University Press.

Solomon, Toby Charles Penhallurick (2021). "Causal decision theory's predetermination problem". In: *Synthese* 198, pp. 5623–5654. DOI: 10.1007/s11229-019-02425-0.

Spencer, Jack (2021). "An argument against causal decision theory". In: *Analysis* 81.1, pp. 52–61. DOI: 10.1093/analys/anaa037.

Taylor, Jessica (2016). *In memoryless Cartesian environments, every UDT policy is a CDT+SIA policy*. URL: https://www.alignmentforum.org/posts/5bd75cc58225bf06703751b2/in-memoryless-cartesian-environments-every-udt-policy-is-a.

Treutlein, Johannes (2017). *Anthropic uncertainty in the Evidential Blackmail*. URL: https://casparoesterheld.com/2017/05/12/anthropic-uncertainty-in-the-evidential-blackmail/.

Treutlein, Johannes et al. (2021). "A New Formalism, Method and Open Issues for Zero-Shot Coordination". In: *Proceedings of the 38th International Conference on Machine Learning (ICML'21)*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, pp. 10413–10423. URL: https://proceedings.mlr.press/v139/treutlein21a.html.

Waugh, Kevin et al. (2009). "A Practical Use of Imperfect Recall". In: *Proceedings of the Eighth Symposium on Abstraction, Reformulation, and Approximation (SARA2009)*. Lake Arrowhead, CA, USA, pp. 175–182.

Weirich, Paul (2016). "Causal Decision Theory". In: *The Stanford Encyclopedia of Philosophy*. Spring 2016.

A Proof of Proposition 1

Proposition 1. *Let $\mathscr{E}$ be a scenario and let $\Pi = \Delta(A)^O$ or $\Pi = A^O$. Then $\mathscr{E}$ has at least one policy that is ex ante optimal in Π, if $\Pi = \Delta(A)^O$ or $\Pi = A^O$.*

$$\begin{tikzcd} {} \arrow[r] & s_0 \arrow[loop above, "b"] \arrow[r, "a"] & 1 \end{tikzcd}$$

Figure 12: A scenario in which there exists a policy (namely always playing b) that does not lead to a terminal state.

Proof. The case $\Pi = A^O$ is trivial, because A^O is finite. $\Delta(A)^O$ is compact and $Q_\pi(P_0)$ is continuous in π. The proposition therefore follows from (a version of) the extreme value theorem. $\square$

B Why CDT requires additional assumptions about the scenario to be well-defined

To define *ex ante* expected utility and EDT, we need to assume that for all policies π, the probability that a terminal state is reached at some point is 1. More formally, we need to assume that

$$\sum_{s_0...s_n} P(s_0...s_n \mid \pi) = 1,$$

where the sum is over all histories that end in some terminal state s_n. Figure Figure 12 gives a minimal example of a scenario that we exclude by this assumption. In this scenario, it is unclear how one would assess (*ex ante* or otherwise) the policy of choosing b with probability 1.

As noted in the main text, CDT requires stronger assumptions to be well defined. To illustrate this requirement, consider the scenario of Figure 13. First notice that for each policy, a terminal state is reached with probability 1. (To see this, distinguish between the policy that takes b with probability 1 and all other policies.) Hence, we can without problem assign *ex ante* expected utilities $Q_\pi(s_0)$ to all policies. (In particular, the *ex ante* expected utility of a policy that plays a with probability p is simply p.) It is easy to verify that EDT can also be applied to this scenario without trouble. In fact, the proof of our results about EDT+GDH in Section 4 only require that the scenario ensures that $\sum_{s_0...s_n} P(s_0...s_n \mid \pi) = 1$ and not the stronger assumption given in the main text.

For CDT, on the other hand, the scenario of Figure 13 spells trouble. Consider the policy of always choosing b. To determine whether this policy is CDT compatible, we need to calculate a value $Q_{\text{always } b}(s_0, a)$, i.e., the expected utility of choosing a in s_0, assuming the agent will always play b otherwise. However, this expected utility is undefined: if the agent follows a in s_0 and then always plays b, the infinite history $s_0 s_1 s_1 s_1 ...$ will be realized and no terminal state will be reached. If the agent assigned, say, a utility of 0 to infinite histories, then always b would not be CDT compatible; in fact, the scenario would have no compatible policy.

Because of scenarios such as this one, we restrict attention in the main text to scenarios in which the Q values are well defined, i.e., to scenarios in which even if the agent deviates from any given policy once, a terminal state will be reached with probability 1. Note that, for simplicity, the main text makes the slightly stronger assumption

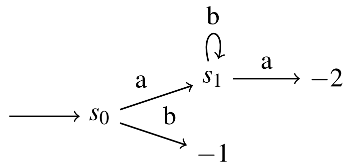

Figure 13: In this scenario, the *ex ante* expected utilities of all policies are well defined, but one of the causal expected utilities (namely, $Q_{\text{always b}}(s_0, a)$) is undefined.

that for every state s, policy π, and action a, choosing a in s and then following π reaches a terminal state with probability 1 (even if s is reached with probability 0 given π).

CDT is unable to pass judgment in situations that are unproblematic from an *ex ante* or EDT perspective. Is this an argument against CDT? This question is beyond the scope of this paper, but we note that this problem relates to a general critique of CDT: CDT gives weight to events and counterfactuals that the agent knows are impossible. For example, in Newcomb's problem with an infallible predictor, CDT considers (and gives weight to) what happens if the predictor predicts two-boxing and the agent one-boxes (see, e.g., Solomon, 2021).[18] Similarly, in the scenario of Figure 13, CDT gives weight to the event that it chooses b with probability 1 but chooses a.

C On the benefits of asymmetric choice in symmetric situations

In this short section, we argue that *de se* rational agents might make asymmetric choices in a pair of observations that are symmetric to one another. Consider the scenario in Figure 14. In this scenario, the agent chooses twice between a and b, once she makes this choice in o_1 and once in o_2. She receives a reward of 1 if she chooses a once and b once – regardless of whether she chooses a in o_1 and b in o_2 or *vice versa*. If she chooses a twice her payoff is -10, and if she chooses b twice her utility is 0. The two situations o_1 and o_2 are symmetric in the following sense: if we take a policy π and construct a new policy π' with $\pi'(\cdot \mid a) = \pi(\cdot \mid b)$ and $\pi'(\cdot \mid b) = \pi(\cdot \mid a)$, then $Q_{\pi'}(s_0) = Q_{\pi}(s_0)$. (In the graph of Figure 14, this symmetry between o_1 and o_2 is not so apparent. To make them appear more symmetric in the graph, we could first let the scenario decide at random whether o_1 or o_2 is the first observation.) Of course, the two optimal strategies break this symmetry and choose a in one of the two observations and b in the other. We find it plausible that (without having to commit to or otherwise select such a policy *ex ante*), a rational agent would be able to break this symmetry by having some general convention with herself. For example, alphabetical order suggests the strategy of playing a in o_1 and b in o_2.

[18]Also see a blog post by Oesterheld (2017).

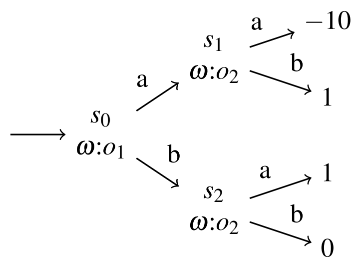

Figure 14: A scenario in which it is beneficial for an agent to follow an asymmetric policy, i.e., a policy that chooses differently in o_1 and o_2, despite the fact that o_1 and o_2 are symmetric.

D Proofs of Lemma 5, Theorem 6 and Corollary 7

Recall our definition of derivatives with respect to the policy:

Definition 10. *Let π be a policy, a be an action and o be an observation. Then for all $\varepsilon > 0$ define*

$$\pi_{\varepsilon,a,o}(a' \mid o') = \begin{cases} \pi(a' \mid o'), & \text{if } o' \neq o \\ (1-\varepsilon)\pi(a' \mid o), & \text{if } o' = o \text{ and } a' \neq a \\ (1-\varepsilon)\pi(a' \mid o) + \varepsilon, & \text{if } o' = o \text{ and } a' = a \end{cases}.$$

Then define

$$\frac{d}{d\pi(a \mid o)} Q_\pi(P_0) := \lim_{\varepsilon \downarrow 0} \frac{Q_{\pi_{\varepsilon,a,o}}(P_0) - Q_\pi(P_0)}{\varepsilon}.$$

Lemma 5. *For o observed with positive probability,*

$$\frac{d}{d\pi(a \mid o)} Q_\pi(P_0) = C_\pi(o)(\mathrm{EU}_{\mathrm{GT}}(\pi, o, a) - \mathrm{EU}_{\mathrm{GT}}(\pi, o)).$$

Proof. For this proof, define $T(s_{i+1} \mid s_i, \pi) := \sum_{a \in A} \pi(a \mid \omega(s_i)) T(s_{i+1} \mid s_i, a)$.

By definition, we need to consider $1/\varepsilon(Q_{\pi_{\varepsilon,a,o}}(P_0) - Q_\pi(P_0))$ as ε goes to 0 from above. We will focus on the minuend,

$$1/\varepsilon Q_{\pi_{\varepsilon,a,o}}(P_0) = 1/\varepsilon \sum_{s_0 \dots s_n} P_0(s_0) \left(\prod_{i=0}^{n-1} T(s_{i+1} \mid s_i, \pi_{\varepsilon,a,o}) \right) u(s_n)$$

In the left sum, for s_i with $\omega(s_i) = o$, $T(s_{i+1} \mid s_i, \pi_{\varepsilon,a,o}) = \varepsilon T(s_{i+1} \mid s_i, a) + (1 - \varepsilon) T(s_{i+1} \mid s_i, \pi)$. We can multiply the left side out. Writing and working with this sum would be quite complicated. So instead we describe it. Roughly, we can sort the summands by the order of ε (the exponent of ε), which intuitively is the number of times in the history that the ε-probability deviation from π occurs. So the order 0 term is simply

$$1/\varepsilon \sum_{s_0 \dots s_n} (1-\varepsilon)^{\#(o, s_0 \dots s_n)} P(s_0) \left(\prod_{i=0}^{n-1} T(s_{i+1} \mid s_i, \pi) \right) u(s_n)$$

For small ε, this makes up the vast majority of ${}^1\!/\!\varepsilon Q_{\pi_{\varepsilon,a,o}}(P_0)$. However, these terms will cancel out with the corresponding summands for $s_0...s_n$ in $Q_\pi(P_0)$.

The order 1 term is

$$\sum_{s_0...s_n,k:\ \omega(s_k)=o} (1-\varepsilon)^{\#(o,s_0...s_n)-1} P(s_0) T(s_{k+1} \mid a, s_k) \left(\prod_{i \neq k} T(s_{i+1} \mid \pi, s_i) \right) u(s_n).$$

Note that the ε probability of choosing a as opposed to choosing from π in s_k is canceled out by ${}^1\!/\!\varepsilon$. As $\varepsilon \to 0$, $(1-\varepsilon)^{\#(o,s_0...s_n)-1} \to 1$ for all $s_0,...,s_n$. Hence the order 1 term converges to

$$\begin{aligned}
& \sum_{s_0...s_n,k:\ \omega(s_k)=o} P(s_0) T(s_{k+1} \mid a, s_k) \left(\prod_{i \neq k} T(s_{i+1} \mid \pi, s_i) \right) u(s_n) \\
= & \sum_{s \in S:\ \omega(s)=o} \underbrace{\sum_{\text{prefix } s_0...s_k:\ s_k=s} \left(\prod_{i=0}^{k-1} T(s_{i+1} \mid \pi, s_i) \right)}_{=C_\pi(s)} \\
& \underbrace{\sum_{s_{k+1}...s_n} T(s_{k+1} \mid a, s_k) \left(\prod_{i=k+1}^{n} T(s_{i+1} \mid \pi, s_i) \right) u(s_n)}_{=Q_\pi(a,s)} \\
= & \sum_{s \in S:\ \omega(s)=o} C_\pi(s) Q_\pi(a,s).
\end{aligned}$$

In the higher order terms, ε occurs with an exponent of at least 2, or at least 1 after canceling out with the multiplication by ${}^1\!/\!\varepsilon$. Thus, these terms become arbitrarily small as $\varepsilon \to 0$.

We conclude that

$${}^1\!/\!\varepsilon(Q_{\pi_{\varepsilon,a,o}}(P_0) - Q_\pi(P_0)) \to \sum_{s \in S:\ \omega(s)=o} C_\pi(s)(Q_\pi(a,s) - Q_\pi(s))$$

Finally, notice that this sum is by Definition 5 equal to

$$\begin{aligned}
& \sum_{s \in S:\ \omega(s)=o} C_\pi(o) P_{\mathrm{GT}}(s \mid \pi, o) \left(Q_\pi(s,a) - Q_\pi(s) \right) \\
= & C_\pi(o)(\mathrm{EU}_{\mathrm{GT}}(\pi,o,a) - \mathrm{EU}_{\mathrm{GT}}(\pi,o))
\end{aligned}$$ □

Lemma 5 implies the following.

Theorem 6. *A policy $\pi \in \Pi = \Delta(A)^O$ is CDT+GT compatible if and only if for all $o \in O$ and $a \in A$, $\frac{d}{d\pi(a|o)} Q_\pi(P_0) \leq 0$.*

Theorem 6 in turn directly implies Corollary 7 – clearly the derivative at a global optimum must be non-positive in all directions.

Corollary 7 (Piccione and Rubinstein, 1997)**.** *Let π be a globally ex ante optimal strategy from $\Pi = \Delta(A)^O$. Then π is CDT+GT compatible.*

E Proofs of Theorem 8 and Corollary 12

Theorem 8. *Let $\Pi \subseteq \Delta(A)^O$. A policy $\pi \in \Pi$ is EDT+GDH compatible in Π if and only if for all $o \in O, \alpha \in \Delta(A)$ s.t. $\pi_{o\to\alpha} \in \Pi$, $Q_\pi(P_0) \geq Q_{\pi_{o\to\alpha}}(P_0)$.*

Proof. Let π be a policy. Note that neither side of the claimed equivalence puts any restrictions on what π does in observations that are made with probability 0; we only need to consider o that are observed with positive probability. EDT+GDH compatibility means that for all o observed with positive probability in π, it is

$$\pi(\cdot \mid o) \in \operatorname*{arg\,max}_{\alpha\in\Delta(A):\ \pi_{o\to\alpha}\in\Pi} \sum_{s_0\ldots s_n} \sum_{i=1}^{n-1} P_{\text{GDH}}(i\text{-th in } s_0\ldots s_n \mid \pi_{o\to\alpha}, o) u(s_n).$$

First note that if o is observed at least once in $s_0\ldots s_n$, it is

$$P(s_0\ldots s_n \mid \pi_{o\to\alpha}, o) = \frac{P(s_0\ldots s_n \mid \pi_{o\to\alpha})}{P(o \mid \pi_{o\to\alpha})},$$

where $P(o \mid \pi_{o\to\alpha})$ is the probability that o is observed at least once given that policy $\pi_{o\to\alpha}$ is used.

We thus get that

$$\begin{aligned} &\operatorname*{arg\,max}_{\alpha\in\Delta(A):\ \pi_{o\to\alpha}\in\Pi} \sum_{s_0\ldots s_n} \sum_{i=1}^{n-1} P_{\text{GDH}}(i\text{-th in } s_0\ldots s_n \mid \pi_{o\to\alpha}, o) u(s_n) \\ = \ &\operatorname*{arg\,max}_{\alpha\in\Delta(A):\ \pi_{o\to\alpha}\in\Pi} \sum_{s_0\ldots s_n \text{ with } o} \sum_{i:\ \omega(s_i)=o} \frac{P(s_0\ldots s_n \mid \pi_{o\to\alpha})}{\#(o, s_0\ldots s_n) P(o \mid \pi_{o\to\alpha})} u(s_n), \end{aligned}$$

where the first sum on the right-hand side is over all histories that give rise to observation o at some point. Dividing by the number of agents with observation o in a history and summing over all times at which o is observed cancel each other out, such that this equals

$$= \operatorname*{arg\,max}_{\alpha\in\Delta(A):\ \pi_{o\to\alpha}\in\Pi} \frac{1}{P(o \mid \pi_{o\to\alpha})} \sum_{s_0\ldots s_n \text{ with } o} P(s_0\ldots s_n \mid \pi_{o\to\alpha}) u(s_n).$$

Now note that $P(o \mid \pi_{o\to\alpha})$ is constant in α, i.e., the probability that you observe o at least once cannot depend on what you would do when you observe o. Thus, the argmax equals

$$\operatorname*{arg\,max}_{\alpha\in\Delta(A):\ \pi_{o\to\alpha}\in\Pi} \sum_{s_0\ldots s_n \text{ with } o} P(s_0\ldots s_n \mid \pi_{o\to\alpha}) u(s_n).$$

Finally, this argmax equals

$$\operatorname*{arg\,max}_{\alpha\in\Delta(A):\ \pi_{o\to\alpha}\in\Pi} \sum_{s_0\ldots s_n} P(s_0\ldots s_n \mid \pi_{o\to\alpha}) u(s_n).$$

This is because $\sum_{s_0\ldots s_n \text{ without } o} P(s_0\ldots s_n \mid \pi_{o\to\alpha}) u(s_n)$ is constant across α. Thus, we can add this term and this argmax remains the same. By definition, we have thus derived that π is EDT+GDH compatible if and only if for all o that are observed with positive probability,

$$\pi(\cdot \mid o) \in \underset{\alpha\in\Delta(A):\ \pi_{o\to\alpha}\in\Pi}{\arg\max} Q_{\pi_{o\to\alpha}}(P_0),$$

as claimed. $\square$

Corollary 12. *If a policy is EDT+GDH compatible (without any policy restriction), it is CDT+GT compatible.*

Proof. We prove the contrapositive, i.e., that every policy that is not compatible with CDT+GT is also not compatible with EDT+GDH. So let π be any policy that is not compatible with EDT+GDH. Then by Theorem 6, there is an $o \in O$ observed with positive probability when following π and an action $a \in A$ s.t. $d/d\pi(a \mid o)Q_\pi(P_0) > 0$. Hence, for sufficiently small ε, $Q_{\pi_{\varepsilon,a,o}}(P_0) > Q_\pi(P_0)$. By Theorem 8, π is not EDT+GDH compatible. $\square$

F Proofs on single-halfing

F.1 Proofs of Theorem 14, Corollary 15 and Proposition 16

For any prefix history $s_0...s_i$, i.e., any history that doesn't end in a terminal state, define

$$P_{\mathrm{GT}}(s_0...s_i \mid \pi, o) := \frac{P(s_0...s_i \mid \pi)}{C_\pi(o)}$$

to be the generalized thirder's probability of being in state s_i after the prefix history $s_0...s_{i-1}$ occurred.

We will also use the random variable H for the history of the scenario. For any history $s_0....s_n$ (that ends in a terminal state $s_n \in S_T$ as usual), we define $\mathrm{len}(s_0...s_n) = n$ to be the (observation) length of the history.

We now first prove a result about CDT+GSH. This result will establish the similarity between CDT+GSH and CDT+GT, without assuming that the scenario is first transformed to randomize only in the beginning.

Lemma 22. *Let π be a policy. Then π is CDT+GSH-compatible if and only if for all o that are observed with positive probability, $\pi(\cdot \mid o)$ assigns positive probability only to actions from*

$$\underset{a\in A}{\arg\max} \sum_{\text{prefix } s_0...s_i:\ \omega(s_i)=o} P_{\mathrm{GT}}(s_0...s_i \mid \pi, o)\mathbb{E}\left[\frac{1}{\mathrm{len}(H)} \mid s_0...s_i, \pi\right] Q_\pi(s_i, a).$$

Proof. By Definition 9, CDT+GSH requires that for all o that are observed with positive probability, the agent choose from

$$\underset{a\in A}{\arg\max} \sum_{s\in S} P_{\mathrm{GSH}}(s \mid o, \pi) Q_\pi(s, a).$$

Now we can fill in the definition for P_{GSH}, omitting the normalizing denominator, which is constant across s:

$$\arg\max_{a \in A} \sum_{i, s_0 \ldots s_n :\ \omega(s_i) = o} \frac{1}{n} P(s_0 \ldots s_n \mid \pi) Q_\pi(s_i, a).$$

Now notice that $P(s_0 \ldots s_n \mid \pi) = P_{\mathrm{GT}}(s_0 \ldots s_i \mid \pi, o) P(s_{i+1} \ldots s_n \mid \pi, s_i) C_\pi(o)$, where $P(s_{i+1} \ldots s_n \mid \pi, s_i)$ is the probability that the following states are $s_{i+1} \ldots s_n$ when the current state is s_i and the agent uses policy π. Since $C_\pi(o)$ is constant w.r.t. what the argmax and sum are over, we can omit it. Hence, the above argmax is equal to

$$\begin{aligned}
& \sum_{i, s_0 \ldots s_n :\ \omega(s_i) = o} \frac{1}{n} P_{\mathrm{GT}}(s_0 \ldots s_i \mid \pi, o) P(s_{i+1} \ldots s_n \mid \pi, s_i) Q_\pi(s, a) \\
= & \sum_{\text{prefix } s_0 \ldots s_i :\ \omega(s_i) = o} P_{\mathrm{GT}}(s_0 \ldots s_i \mid \pi, o) \left(\sum_{s_{i+1} \ldots s_n} \frac{1}{n} P(s_{i+1} \ldots s_n \mid \pi, s_i) \right) Q_\pi(s, a).
\end{aligned}$$

Clearly, this is equal to the desired argmax term. $\square$

While the above lemma talks about CDT+GSH in general, we can now apply the lemma to CDT+GSH* to obtain Theorem 14 and Proposition 16.

Proposition 16. *Let $\mathscr{E}$ be a scenario that randomizes only in the beginning and let π be a deterministic policy. Let $\hat{\mathscr{E}}$ be the scenario that is equal to $\mathscr{E}$, except that $\hat{P}_0(s_0) \sim P_0(s_0)/\mathrm{len}_\pi(s_0)$. Then π is CDT+GSH-compatible in $\mathscr{E}$ if and only if π is CDT+GT-compatible in $\hat{\mathscr{E}}$.*

Proof. In the following we distinguish between $P_{\mathrm{GT}}^{\mathscr{E}}$ and $P_{\mathrm{GT}}^{\hat{\mathscr{E}}}$, which are the generalized thirder's distributions over states in $\mathscr{E}$ and $\hat{\mathscr{E}}$, respectively. We use $P^{\mathscr{E}}$ and $P^{\hat{\mathscr{E}}}$, and $C_\pi^{\mathscr{E}}$ and $C_\pi^{\hat{\mathscr{E}}}$ analogously. By Lemma 22, π is CDT+GSH-compatible in $\mathscr{E}$ if and only if for all o, $\mathrm{supp}(\pi(\cdot \mid o))$ is a subset of

$$\begin{aligned}
& \arg\max_{a \in A} \sum_{\text{prefix } s_0 \ldots s_i :\ \omega(s_i) = o} P_{\mathrm{GT}}^{\mathscr{E}}(s_0 \ldots s_i \mid \pi, o) \mathbb{E}\left[\frac{1}{\mathrm{len}(H)} \mid s_0 \ldots s_i, \pi \right] Q_\pi(s_i, a) \\
= & \arg\max_{a \in A} \sum_{\text{prefix } s_0 \ldots s_i :\ \omega(s_i) = o} P_{\mathrm{GT}}^{\mathscr{E}}(s_0 \ldots s_i \mid \pi, o) / \mathrm{len}_\pi(s_0) Q_\pi(s_i, a)
\end{aligned}$$

Now notice that

$$\begin{aligned}
P_{\mathrm{GT}}^{\hat{\mathscr{E}}}(s_0 \ldots s_i \mid \pi, o) &= \frac{P^{\hat{\mathscr{E}}}(s_0 \ldots s_i \mid \pi)}{C_\pi^{\hat{\mathscr{E}}}(o)} \\
&= \frac{\hat{P}_0(s_0) T(s_1 \mid \pi, s_0) \ldots T(s_i \mid \pi, s_{i-1})}{C_\pi^{\hat{\mathscr{E}}}(o)} \\
&= \frac{P_0(s_0) T(s_1 \mid \pi, s_0) \ldots T(s_i \mid \pi, s_{i-1}) / \mathrm{len}_\pi(s_0)}{C_\pi^{\hat{\mathscr{E}}}(o) \sum_{s_0'} P_0(s_0') / \mathrm{len}_\pi(s_0')} \\
&= \frac{P^{\mathscr{E}}(s_0 \ldots s_i \mid \pi) / \mathrm{len}_\pi(s_0)}{C_\pi^{\hat{\mathscr{E}}}(o) \sum_{s_0'} P_0(s_0') / \mathrm{len}_\pi(s_0')}
\end{aligned}$$

The denominator is constant across $s_0...s_i$ and a. Moreover, $C_\pi^{\mathscr{E}}(o)$ is also constant across $s_0...s_i$ and a. Thus, we can rewrite the argmax as follows:

$$\begin{aligned}
& \arg\max_{a\in A} \sum_{\text{prefix } s_0...s_i:\ \omega(s_i)=o} P_{\mathrm{GT}}^{\mathscr{E}}(s_0...s_i \mid \pi, o)/\mathrm{len}_\pi(s_0) Q_\pi(s_i, a) \\
= \ & \arg\max_{a\in A} \sum_{\text{prefix } s_0...s_i:\ \omega(s_i)=o} P_{\mathrm{GT}}^{\hat{\mathscr{E}}}(s_0...s_i \mid \pi, o) Q_\pi(s_i, a) \\
= \ & \arg\max_{a\in A} \sum_{s:\ \omega(s)=o} P_{\mathrm{GT}}^{\hat{\mathscr{E}}}(s \mid \pi, o) Q_\pi(s, a).
\end{aligned}$$

Overall we have no shown that π is CDT+GSH-compatible in $\mathscr{E}$ if and only if for all o, $\mathrm{supp}(\pi(\cdot \mid o))$ is a subset of

$$\arg\max_{a\in A} \sum_{s:\ \omega(s)=o} P_{\mathrm{GT}}^{\hat{\mathscr{E}}}(s \mid \pi, o) Q_\pi(s, a),$$

i.e., if and only if π is CDT+GT compatible in $\hat{\mathscr{E}}$. □

Theorem 14. *Let $\mathscr{E}$ be a scenario that randomizes only in the beginning and where history length is choice-independent. Let $\hat{\mathscr{E}}$ be the scenario that is equal to $\mathscr{E}$, except that $\hat{P}_0(s_0) \sim P_0(s_0)/\mathrm{len}(s_0)$. Then any (potentially mixed) policy is CDT+GSH-compatible in $\mathscr{E}$ if and only if it is CDT+GT-compatible in $\hat{\mathscr{E}}$.*

Proof. This is proved in exactly the same way as Proposition 16. □

Corollary 15. *Let $\mathscr{E}$ be a scenario that randomizes only in the beginning and where history length is choice-independent. Then there exists a CDT+GSH-compatible non-Dutch-book policy for $\mathscr{E}$.*

Proof. Consider the *ex ante* optimal policy π^* for $\mathscr{E}$ as defined in Theorem 14. By Corollary 7, π^* is CDT+GT compatible in $\hat{\mathscr{E}}$. By Theorem 14, π^* is thus CDT+GSH compatible in $\mathscr{E}$. It is easy to see that π^* cannot be a Dutch Book. □

F.2 Why CDT plus GSH needs to view policy randomization as predetermined

In Section 5, we have eliminated CDT+GSH's vulnerability to Dutch books by having it imagine that all randomization occurs at the beginning of the scenario. We illustrated the need for this by using Draper and Pust's (2008) Dutch book. However, strictly speaking, based on their scenario we can only show that CDT+GSH needs to imagine that the *scenario* only randomizes in the beginning. But in our definition of CDT+GSH*-compatibility as applied to some mixed policy π, we imagine that all results of π's randomization are determined in the very beginning of the scenario and that the agent (by playing some action a_π) merely accesses these actions that were sampled at the very beginning of the scenario. In this section, we show why this is necessary. In particular, we show a scenario in which the scenario is completely deterministic and in which the only CDT+GSH-compatible policy is mixed and loses money with certainty.

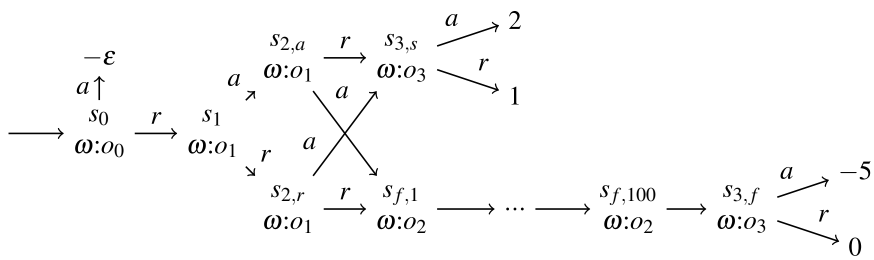

Figure 15: A graphical formalization of Example 12.

Proposition 17. *There is a scenario $(S, S_T, P_0, O, \omega, A, T, u)$ with the following properties.*

- *The state transitions are deterministic, i.e., for all $a \in A$, $s, s' \in S$, $T(s' \mid s, a)$ is either* 0 *or* 1.
- *All CDT+GSH-compatible policies are Dutch books.*

Example 12. *The scenario proceeds in three parts.*

1. *At the very beginning, the agent is offered to end the scenario for a price of ε.*
2. *The agent then plays the following coordination game against herself. She faces the same choice twice. If she chooses differently in the two situations, she receives a reward of* 1. *She cannot distinguish between these two situations, retains no memory of whether she has already faced the choice or of what her choice was (if any). If the agent fails to coordinate in part 2, she faces an additional N situations without having to make a (relevant) decision.*
3. *The agent is offered a bet that pays -5 if she failed to coordinate in part two and pays* 1 *if she succeeded in coordinating.*

The scenario is visualized in our formalism in Figure 15.

We now argue that Example 12 proves Proposition 17. The first two points are easy. Clearly the scenario's state transitions are deterministic – in fact, even the initial distribution is deterministic. Furthermore, the deterministic policy of always rejecting achieves a reward of 0 with certainty.[19]

We now argue that all CDT+GSH-compatible policies accept the offer in s_0 with probability 1 and thus lose money with certainty. To do so, assume for contradiction that a policy π is CDT+GSH-compatible but rejects in s_0 with positive probability. Then o_1 and o_3 are observed with positive probability. It is easy to verify that regardless of what π does in other observations, CDT+GSH compatibility then requires that the agent randomizes uniformly in o_1. Consequently, $s_{3,s}$ and $s_{3,f}$ occur with equal probability. However, because $s_{3,f}$ only occurs in very long histories, a GSH agent believes conditional on observing o_3 that it is in $s_{3,s}$ with overwhelming probability, specifically (using the fact that π randomizes uniformly in o_1) $P_{\mathrm{GSH}}(s_{3,s} \mid \pi, o_3) = 26/27$. Thus, to be CDT+GSH compatible, π has to accept upon observing o_3. Now in s_0, the expected

[19]Note that the *ex ante* optimal policy is to reject in part 1, mix uniformly upon observation o_1, i.e., in the coordination game in part 2, and to reject the bet upon observation o_3 in part 3, for an expected payoff of $1/2$.

value of rejecting is $1/2 \cdot 2 + 1/2 \cdot (-5) = -3/2$. Since this is less than -1, the agent strictly prefers accepting in s_0, contradicting the assumption that the agent rejects with positive probability in s_0.

F.3 How CDT+GSH* solves Example 12

We now show how CDT+GSH* avoids the Dutch book of Example 12. Specifically, we show that the policy π of rejecting in o_0, mixing uniformly in o_1 and accepting in o_3 is CDT+GSH*-compatible. Note that this is not the optimal policy – the optimal policy rejects in o_3. In fact, while this policy is not a Dutch book, its *ex ante* expected utility is actually worse than the Dutch book of paying the price of ε in s_0.

First, what does the modified scenario look like? First, we add an action a_π that corresponds to following the policy π described above in any given situation. Thus, in o_0, a_π is equivalent to rejecting the offer, and in o_3, a_π is equivalent to accepting the offer. Upon observing o_2, i.e., when playing the coordination stage of the game, π randomizes. As always, CDT+GSH* works by moving this randomization to the beginning of the scenario. Thus, there are now four different initial states. The initial state encodes what choices will result from playing a_π. For example, if the initial state $s_{0,r,a}$ is selected, then following choosing a_π will result in playing r on the first observation of c and will result in playing a on the second observation of o_1. Because π samples uniformly, the initial state is also selected uniformly by the scenario. Note that after the agent has made a choice upon a first observation of o_1, the scenario only remembers the actual choice made, not the one that would have been made, had the agent played a_π. For instance, the state $s_{2,r,a}$ indicates that the agent has played r upon her first observation of c, and that playing a_π now (i.e., upon her second observation of o_1) will result in playing a. Consequently, this state can not only be reached by playing a_π in $s_{1,r,a}$, but also by playing r in $s_{1,r,a}$ or $s_{1,a,a}$. The complete formal model is given in Figure 16.

It is easy to verify that in this new model, the policy of always playing a_π is CDT+GSH-compatible. We omit a detailed calculation and only provide some notes here. First, the expected utility calculation upon observing o_3 is essentially the same as in the original scenario. Second, GSH's belief in short histories (and thus in its ability to successfully coordinate) now also affects the GSH probabilities over states conditional on o_1 and on the policy of always playing a_π. In particular, conditional on o_1, GSH assigns most probability to the states $s_{1/2,a,r}$ and $s_{1/2,r,a}$. Between these four states, the GSH probabilities are uniform. Of course, this consistency (in its belief in short histories) is the whole point of CDT+GSH*. However, it has the odd consequence that CDT+GSH *strictly* prefers a_π over both a and r. The strictness of this preference is harmless for the present scenario, because it is a strict preference in the right direction – we want a_π to be CDT+GSH compatible. However, it illustrates a mechanism that we will see is CDT+GSH*'s downfall in Section 5.4.2. Finally, CDT+GSH*'s most obvious divergence from CDT+GSH occurs when observing o_0. CDT+GSH*, again, has high confidence in short histories and thus successful coordination throughout the scenario, even upon observing o_0. In particular, upon observing o_0 GSH is confident in this new model that it is in either $s_{0,a,r}$ or $s_{0,r,a}$. Since the agent's reward in these two states is 2 under following a_π, CDT+GSH prefers rejecting the offer.

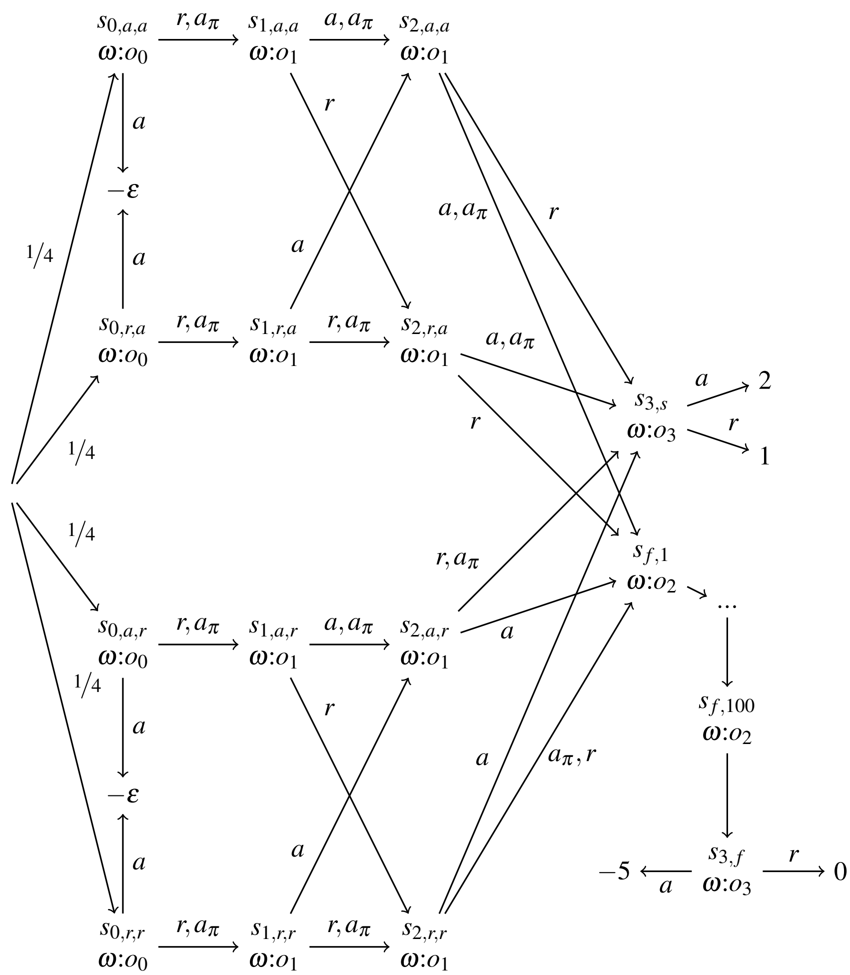

Figure 16: An alternative model of Example 12 in which the policy π of rejecting in s_1, mixing uniformly in c and accepting in b can be followed by deterministically playing a_π.

F.4 A more formal analysis of our Dutch book against CDT+GSH* (Example 7)

In this section, we give a more formal analysis of Example 7. We first recall the example here.

Example 7. *First the agent faces a choice between a_0 and a_1 three times. She cannot distinguish between these three situations, retains no memory of how often she has already faced the choice or of what her choices were. Her rewards are determined from the number of times she chooses a_1 in these situations according to the following table.*

Number of times a_1 is played	*Reward*
0	0
1	1
2	-1
3	$-\varepsilon$

Here, ε is some small but positive number, e.g., $\varepsilon = 1/100$. If a_1 was chosen exactly once, for a reward of 1, then the agent faces the same decision problem between a_0 and a_1 another K times (for some large K). Importantly, the agent's choices in these K situations do not affect her final reward – her reward remains 1.

First, we analyze this problem from the *ex ante* perspective, as well as the perspective of CDT+GT. Obviously, the agent can guarantee a non-negative payoff for herself by never playing a_1. It is also easy to see that the globally optimal strategy is to play a_1 with some small positive probability aimed at obtaining the maximum reward of 1 with a much higher chance than obtaining the reward of -1. Specifically, the expected utility of the policy π_p that chooses a_1 with probability p is

$$Q_{\pi_p}(s_0) = 3p(1-p)^2 - 3p^2(1-p) - \varepsilon p^3.$$

This function is plotted for $\varepsilon = 1/10$ in Figure 17. For $\varepsilon = 1/10$, the *ex ante* optimal policy is to play a_1 with probability

$$p = 1/59\left(30 - \sqrt{310}\right) \approx 0.210054.$$

By Corollary 7, this policy is also CDT+GT-compatible; and by Corollary 11 it is the only EDT+GDH-compatible policy. It is easy to see that another CDT+GT-compatible strategy is to play a_1 with probability 1. A third CDT+GT-compatible policy can be found at the second zero of the derivative of $Q_{\pi_p}(s_0)$, which is also the global minimum of $Q_{\pi_p}(s_0)$. Specifically, this point is at

$$p = 1/59\left(30 + \sqrt{310}\right) \approx 0.806895.$$

We now show that the only CDT+GSH*-compatible strategy in Example 7 is to play a_1 with probability 1, which loses money with probability 1. First, it is easy to see

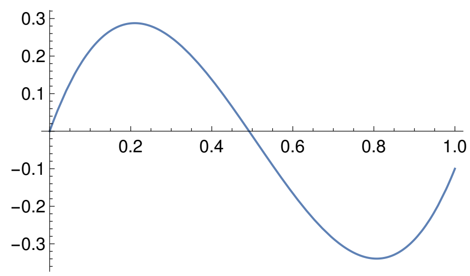

Figure 17: The *ex ante* expected utility in Example 7 plotted against the probability of taking action a_1.

that the policy of playing a_1 with probability 1 is CDT+GSH* compatible. It is left to show that no other policy is CDT+GSH*-compatible.

Recall that to evaluate a policy π that plays a_1 with probability $p < 1$, we construct deterministic scenario $\hat{\mathscr{E}}_\pi^d$, in which the actions of π is rolled out beginning and encoded in the initial state. We then ask whether always playing a_π – which takes the actions encoded in the state – is compatible with CDT+GSH. We will show that if $p < 1$, the agent prefers a_1 over playing a_π, conditional on playing a_π in all other states.

Consider the following two kinds of states that the agent may be in:

1. States with the following property:
    - Always playing a_π leads to playing a_1 once.
    - Playing a_1 (as opposed to a_π) in this state, while otherwise following a_π, leads to a_1 being played twice.
2. States with the following property:
    - Always playing a_π leads to playing a_1 either 0 or 2 times.
    - Playing a_1 (as opposed to a_π) in this state, while otherwise following a_π, leads to a_1 being played 1 or 3 times.

Under the first type of state, playing a_1 substantially (by 2) *decreases* utility relative to playing a_π. Under the second type of state, playing a_1 substantially (by 1 or $1-\varepsilon$) increases utility relative to playing a_π. In all other states, it makes no difference whether the agent plays a_1 or a_π.

Finally, to see that a_π is not CDT+GSH*-compatible, we argue that (for large K), GSH assigns much higher probability to the second type than the first, for all π. To do so, notice first that the ratio of the GT probabilities of the first and second kind of states is bounded above – in other words, the GT probability of being in the first kind of state can only be made to be larger than than the GT probability of the second type of state by at most some fixed factor, regardless of the agent's policy. This is intuitive, but semi-formally this is because the probability of zero or two (utility-relevant) a_1s being played is $(1-p)^3 + 3p^2(1-p)$, while the probability of exactly one utility-relevant a_1

being played is $3p(1-p)^2$. We can than see that the ratio is bounded as follows:

$$\frac{3p(1-p)^2}{(1-p)^3+3p^2(1-p)} = \frac{3p(1-p)}{(1-p)^2+3p^2} \le \frac{3\,\overbrace{p(1-p)}^{\le 1/4}}{\underbrace{(1-p)^2+p^2}_{\ge 1/2}} \le \frac{3}{2}.$$

However, if we use GSH probabilites instead of GT probabilities, then the first type of states receive a penalty of $3/K+3$ relative to the second kind of states. Thus, as $K \to \infty$, GSH assigns arbitrarily low probability to being in the first type of state. It follows that for large enough K, CDT+GSH* recommends playing a_1 regardless of the policy.

G A *de se* criterion for Dutch books

The following result provides a *de se* criterion for whether a given policy π is a Dutch book.

Proposition 23. *Let X be any one of GDH, GSH, GT. A policy π is a Dutch book if and only if*

- *For all o observed with positive probability under π, $P_{\mathrm{GDH/GSH/GT}}(s_0...s_n \mid o,\pi) > 0 \implies u(s_n) < 0$; and*
- *there is a policy π_0 s.t. for all observations o observed with positive probability under π_0, $P_{\mathrm{GDH/GSH/GT}}(s_0...s_n \mid o,\pi_0) > 0 \implies u(s_n) \ge 0$.*

Proposition 23 follows directly from the following lemma.

Lemma 24. *For any history $s_0...s_n$ and policy π, the following two statements are equivalent:*

- *$P(s_0...s_n \mid \pi) > 0$.*
- *There exists $o \in O$ that is observed with positive probability s.t. $P_{\mathrm{GDH/GSH/GT}}(s_0...s_n \mid \pi, o) > 0$.*

Proof. By definition, a policy π is a Dutch book policy if and only if it yields negative reward with probability 1 and there is another policy π_0 that yields non-negative reward with probability 1. It is easy to see from the definitions of GDH, GSH and GT that $P_{\mathrm{GDH/GSH/GT}}(s_0...s_n \mid o,\pi) > 0$ implies that $P(s_0...s_n \mid \pi) > 0$. That is, all three of our methods for assigning self-locating beliefs assign positive probability only to histories that are in fact possible under the given policy. Furthermore, it is easy to see that if $P(s_0...s_n \mid \pi) > 0$, then $P_{\mathrm{GDH/GSH/GT}}(s_0...s_n \mid o,\pi) > 0$. □

While Proposition 23 provides a purely *de se* criterion for avoiding Dutch books, its relevance to expected utility maximizers in particular is unclear, since it is a criterion about possible outcomes and not about expected utilities.

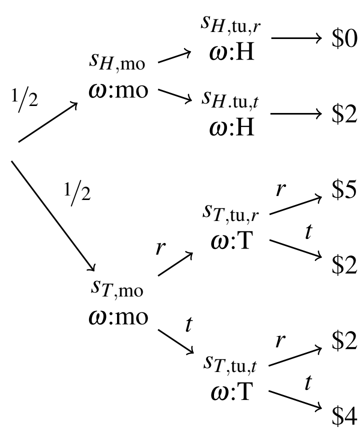

Figure 18:

H A disagreement in preferences over policies between different decision points

In this section, we describe a decision scenario in which the two decision perspectives disagree about which compatible policy is best.

Proposition 25. *There exists a decision scenario with two CDT+GT-, CDT+GSH- and EDT+GDH-compatible policies* π_1, π_2 *and observations* o_1, o_2 *observed with positive probability under both* π_1, π_2 *s.t.*

$$\begin{aligned} \mathrm{EU}_{\mathrm{GT}}(\pi_1, o_1) > \mathrm{EU}_{\mathrm{GT}}(\pi_2, o_1) \quad & \textit{and} \quad \mathrm{EU}_{\mathrm{GSH}}(\pi_1, o_1) > \mathrm{EU}_{\mathrm{GSH}}(\pi_2, o_1) \\ & \textit{and} \quad \mathrm{EU}_{\mathrm{GDH}}(\pi_1, o_1) > \mathrm{EU}_{\mathrm{GDH}}(\pi_2, o_1) \end{aligned}$$

but

$$\begin{aligned} \mathrm{EU}_{\mathrm{GT}}(\pi_1, o_2) < \mathrm{EU}_{\mathrm{GT}}(\pi_2, o_2) \quad & \textit{and} \quad \mathrm{EU}_{\mathrm{GSH}}(\pi_1, o_2) < \mathrm{EU}_{\mathrm{GSH}}(\pi_2, o_2) \\ & \textit{and} \quad \mathrm{EU}_{\mathrm{GDH}}(\pi_1, o_2) < \mathrm{EU}_{\mathrm{GDH}}(\pi_2, o_2). \end{aligned}$$

Example 13. *On Monday, Alice can choose to take* \$2 *or refrain. With 50% Alice is offered the same choice again on Tuesday. With the remaining 50%, Alice is woken up on Tuesday without facing a choice. On Tuesday, Alice does not remember her Monday choice, but she always knows whether it is Monday or Tuesday. If Alice is indeed offered the* \$2 *twice and she refrains on both occasions from taking the* \$2*, she receives* \$5*. We formalize this in our framework in Figure 18.*

We will only consider the following two deterministic policies: always refrain, and always take. Clearly, always taking is *ex ante* optimal and *ex ante* strictly better than always refraining. Moreover, both policies are CDT+GT-, CDT+GSH- and EDT+GDH-compatible. It is easy to verify that conditional on observing mo, all our theories assign equal probability to $s_{H,\mathrm{mo}}$ and $s_{T,\mathrm{mo}}$. Hence, $EU_{\mathrm{GT/GSH/GDH}}(\text{always take}, \mathrm{mo}) = 3$, while $EU_{\mathrm{GT/GSH/GDH}}(\text{always refrain}, \mathrm{mo}) = 2.5$. Upon facing a choice on Tuesday (i.e., upon observing T), it is easy to see that regardless of theory for self-locating beliefs, the agent assigns probability 1 to being in $s_{T,\mathrm{tu},r}$ if she always refrains and probability 1 to $s_{T,\mathrm{tu},t}$ if she always takes. Hence, $EU_{\mathrm{GT/GSH/GDH}}(\text{always take}, T) = 4$, while $EU_{\mathrm{GT/GSH/GDH}}(\text{always refrain}, T) = 5$.

I Extending Example 11 to cover CDT+GSH

Theorem 21. *There is a scenario $\mathscr{E}$ with the following properties.*

- *$\mathscr{E}$ only randomizes in the beginning and the agent's choices do not affect her future observations. (In particular, history length is choice independent.)*
- *There is a CDT+GT-, CDT+GSH- and EDT+GDH-compatible deterministic Dutch book policy $\tilde{\pi}$.*
- *For all EDT+GDH/CDT+GT/CDT+GSH-compatible policies π other than $\tilde{\pi}$ and all observations o,* $\mathrm{EU}_{\mathrm{GDH/GT/GSH}}(\tilde{\pi}, o) > \mathrm{EU}_{\mathrm{GDH/GT/GSH}}(\pi, o)$.

Example 11. *At the beginning, the scenario randomizes uniformly between three possibilities:*

X) The agent observes x (once) and the scenario ends.
Y) The agent observes y (once) and the scenario ends.
Z) The agent observes x (once), then y (once), and then the scenario ends.

Upon observing x or y, the agent chooses from three actions: bet, pay, and pass. By choosing bet, they accept a bet on being in branch X or Y at slightly better than even odds, specifically, for each time they bet, they obtain 1 *if branch X or Y is realized and they lose* $2/3$ *if branch Z is realized. By choosing pay, they lose some small amount* $\varepsilon > 0$*. However, if branch Z is realized and the agent chooses to pay exactly once, they end up with a payoff of* -100*. Choosing to pass has no consequences in and of itself. A graphical description of this problem in our formalism is given in Figure 11.*

As promised, we now extend Example 11 to apply to CDT+GSH(*) as well. (Note again that the present scenario is relatively unproblematic for CDT+GSH (as per Theorem 14), and that on this scenario CDT+GSH and CDT+GSH* as discussed in Section 5 are equivalent.) It is easy to see that in Example 11, the set of CDT+GSH-compatible policies also consists of π_{bet}, π_{pay}, and some mixed policy π_p. The problem is that $\mathrm{EU}_{\mathrm{GSH}}(\pi_{\mathrm{bet}}, x/y) = 2/3 \cdot 1 + 1/3 \cdot (-4/3) = 2/9$. Thus, in Example 11, CDT+GSH* actually prefers the *ex ante* optimal policy over the Dutch book.

Nonetheless, we can use a very similar scenario for CDT+GSH*. We only need to adjust the odds of the bet to account for CDT+GSH* assigning lower probability to branch Z. Specifically, if the bet paid, e.g., 1 in branch X/Y and $-4/3$ in Z, then accepting in x and y is still CDT+GSH*-compatible (and accepting the bet is still always preferred by CDT+GSH* to passing). At the same time, it is then $\mathrm{EU}_{\mathrm{GSH}}(\pi_{\mathrm{bet}}, x/y) = 2/3 \cdot 1 + 1/3(-8/3) = -2/9 < -4\varepsilon/3 = 2/3(-\varepsilon) + 1/3(-2\varepsilon) = \mathrm{EU}_{GSH}(\tilde{\pi}, x/y)$ (for small enough ε), as desired. It's easy to see that the rest of the argument works out as it does for CDT+GT and EDT+GDH in Example 11.

Of course, Theorem 21 claims that there is a scenario that works for all three theories simultaneously. If we do modify the odds of the bet as suggested in the previous paragraph, then π_{bet} ceases to be CDT+GT- or EDT+GDH-compatible. A simple solution to this is that we offer the bet of the previous paragraph as an *alternative* to the bet in Example 11, while increasing the stakes to make CDT+GSH* prefer taking the new bet over taking the old bet, despite the odds of the new one being worse. Formally, we add an action betGSH that provides a reward of 3 in branch X and Y and a reward of -4 in Z. Then the expected reward of the new bet, as judged by CDT+GSH, is $2/3 \cdot 3 - 1/3 \cdot 4 = 2 - 4/3 = 2/3$, while the expected reward of the old bet is $2/3 \cdot 1 - 1/3 \cdot 2/3 = 4/9$.