Can \textit{de se} choice be \textit{ex ante} reasonable in games of imperfect recall? A complete analysis

Caspar Oesterheld and Vincent Conitzer

January 5, 2024

Abstract

In this paper, we study games of imperfect recall, such as the absent-minded driver or Sleeping Beauty. We can study such games from two perspectives. From the \textit{ex ante} perspective (a.k.a. the planning stage) we can assess entire policies from the perspective of the beginning of the scenario. For example, we can assess which policies are \textit{ex ante} optimal and which are Dutch books (i.e., lose money with certainty when it would be possible to walk away with a guaranteed non-negative reward). This perspective is conceptually unproblematic. The second is the \textit{de se} perspective (a.k.a. the action stage), which tries to assess individual choices from any given decision point in the scenario. How this is to be done is much more controversial. Multiple different theories have been proposed, both for how to form beliefs and how to choose based on these beliefs. To resolve such disagreements, multiple authors have shown results about whether particular \textit{de se} theories satisfy \textit{ex ante} standards of rational choice. For example, Piccione and Rubinstein (1997) show that the \textit{ex ante} optimal policy is always “modified multiself consistent”. In the terminology of the present paper (and others in this literature), they show that the \textit{ex ante} optimal policy is always compatible with choosing according causal decision theory and forming beliefs according to generalized thirding (a.k.a. the self-indication assumption). In this paper, we aim to give a complete picture of which of the proposed \textit{de se} theories match the \textit{ex ante} standards. Our first main novel result is that the \textit{ex ante} optimal policy is always compatible with choosing according to evidential decision theory and forming beliefs according to generalized double-halving (a.k.a. compartmentalized conditioning and the minimal-reference-class self-sampling assumption). Second, we show that assigning beliefs according to generalized single-halving (a.k.a. the non-minimal reference class self-sampling assumption) can avoid the Dutch book of Draper and Pust (2008). Nevertheless, we show that there are other Dutch books against agents who form beliefs according to generalized single-halving, regardless of whether they choose according to causal or evidential decision theory.

\textbf{Keywords:} causal decision theory, evidential decision theory, Newcomb’s problem, self-locating beliefs, imperfect recall, Sleeping Beauty, absent-minded driver, anthropics, doomsday argument, self-sampling assumption, self-indication assumption, observer selection effects
Figure 1: A graphical description of the absent-minded driver (Example 1) in our formalism. The agent makes the same observation in $s_0$ and $s_1$ – in game-theoretic parlance, $s_0$ and $s_1$ are in the same information set. For simplicity, we therefore omit the observation annotation.

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 always remember past observations. We illustrate this idea and some of its consequences using a well-known example: the absent-minded driver, which Piccione and Rubinstein (1997) describe as follows.

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, then 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 $\frac{2}{3}$.)

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 their 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; Sandholm, 2015; Čermák, Bošanský, and Lisy, 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 Rubin-
stein, 1997; Binmore, 1997; Detwarasiti and Shachter, 2005; Conitzer, 2019, Sect. “Imperfect Recall”; Emmons et al., 2022 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. Finding optimal Markov strategies in multi-model (a.k.a. concurrent) (fully observable) Markov decision processes (see, for example, Buchholz and Schefelowitsch, 2019; Steimle, Kaufman, and Denton, 2021; Su and Petrik, 2023) is also in part a problem of imperfect recall.

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 intelligent agents and another predicts a small universe with few intelligent agents, should our existence cause us to update toward the large-universe theory? 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 (Cirković, Sandberg, and Bostrom, 2010); the simulation argument (Bostrom, 2003).

Ex ante (Latin for “from before”) optimality, as illustrated above for the absent-minded driver, is the simplest normative notion of choice in scenarios of imperfect recall. 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 absent-minded 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 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 \( \pi \) as given and specify whether the individual choices in \( \pi \) are rational, given that the agent otherwise follows \( \pi \) (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 \( \frac{1}{2} \) in the absent-minded driver setting) and some observation (e.g., seeing an exit on the highway), what is the probability that I am in a particular state? For example, in the absent-minded driver, if I follow the (ex ante suboptimal) policy of continuing with probability \( \frac{1}{2} \) and that

\footnote{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 teammates (i.e., other agents that it has not been trained with), including humans (Hu et al., 2020; Treutlein et al., 2021).}
I see an exit on the highway, what probability should I assign to being at the second exit? People disagree about how to answer such questions in games of imperfect recall. In particular, in the above example, some would argue that the answer should be $\frac{1}{4}$, because the probability of seeing both exits is $\frac{1}{2}$ and of the two decision points in the two-exit trajectories, one, and thus $\frac{1}{2}$ in relative terms, is at the second exit. Others would argue that the answer should be $\frac{1}{3}$, because of the $1.5$ expected number of decision points, $0.5$, and thus $\frac{1}{3}$ in relative terms, are at the second exit. 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) and the self-indication assumption (Bostrom, 2010)), generalized single-halfing (GSH) (a.k.a. the (non-minimal-reference-class) self-sampling assumption (Bostrom, 2010)), and generalized double-halfing (GDH) (a.k.a. Z-consistency (Piccione and Rubinstein, 1997), compartmentalized conditionalization (Meacham, 2008), the minimal-reference-class self-sampling assumption (Bostrom, 2010), or simply “the Halfer Rule” (Briggs, 2010)). GT gives the $\frac{1}{3}$ answer in the above example, while both GDH and GSH give the $\frac{1}{4}$ answer in the above example. The names we use are based on the probabilities they assign in the 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 in our context, take some policy for the absent-minded driver and imagine that based on this policy, the agent assigns some probabilities to being at the two different exits – for example, $\frac{3}{4}$ to being at the first and $\frac{1}{4}$ to being at the second exit. Then CDT would assign an expected utility of $\frac{3}{4} \cdot 0 + \frac{1}{4} \cdot 4 = 1$ to exiting. In contrast, EDT will generally revise its self-locating probabilities based on its actions. For example, it assesses (the policy of deterministically) exiting to yield an expected utility of $0$ with certainty, reasoning: if I exit now, then this means that I exit (and would 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. We refer to them as CDT+GT, CDT+GSH, and so on. 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 more detail below.

Contributions As mentioned above, each of the six theories X specifies, for each given policy $\pi$, whether the agent acts in agreement with $X$ in all decision situations (i.e., upon all observations). When this is the case, we call $\pi$ compatible with theory


<table>
<thead>
<tr>
<th>Theory</th>
<th>CDT</th>
<th>EDT</th>
</tr>
</thead>
<tbody>
<tr>
<td>GT</td>
<td>✓ <em>ex ante</em> optimal (Piccione and Rubinstein, 1997)</td>
<td>✓ Dutch book (Briggs, 2010)</td>
</tr>
<tr>
<td>GDH</td>
<td>✓ Dutch book (Hitchcock, 2004)</td>
<td>✓ <em>ex ante</em> optimal (Corollary 7)</td>
</tr>
<tr>
<td>GSH</td>
<td>✓ Dutch book (Draper and Pust, 2008 Sect. 5)</td>
<td>✓ no compatible policy (Appendix E.5)</td>
</tr>
<tr>
<td>GSH*</td>
<td>✓ no DB if actions don’t affect observations (Corollary 12)</td>
<td>✓ Dutch book (Section 5.4.2)</td>
</tr>
</tbody>
</table>

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. The ✓ symbol indicates a positive result, the ✗ symbol indicates a negative result, the ✓ symbol in parentheses indicates a more limited positive result. The “*ex ante* optimal” entries indicate that in all scenarios the *ex ante* optimal policy is compatible with the respective theory. The “Dutch book” entries indicate that there is a scenario in which all theories compatible with the respective theories are Dutch book policies.

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 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 in the gaps that were left in the table by the existing literature. Specifically, our contributions are the following.

- The *ex ante* optimal policy is always compatible with evidential decision theory + generalized double-halfing (Corollary 7). 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 probabili-
ties to random events more consistently. Note that the unique CDT+GSH*-and EDT+GSH*-compatible policy in Draper and Pust’s scenario is still \textit{ex ante} suboptimal.

- We show that in each scenario in which history length is independent of the agent’s choices (including Draper and Pust’s), 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 the only CDT+GSH*-compatible policy is a Dutch book (Section 5.4.2).
- 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. That is, even if the agent is constrained to, for example, fully deterministic policies, EDT+GDH is still compatible under this restriction with the \textit{ex ante} optimal deterministic policy. For example, in the absent-minded driver the optimal deterministic policy is to always continue. This is compatible with EDT+GDH restricted to deterministic policies, but not compatible with any natural version of CDT. 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 and how our version of EDT avoids it. Again, we draw parallels to ideas from the literature on Newcomb-like decision problems.
- In general, any of the six \textit{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, we pose the question of whether an agent can avoid Dutch books, or perhaps even reliably choose the \textit{ex ante} optimal policy, while relying purely on \textit{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 theories of assigning self-locating beliefs that we consider (GDH, GSH(*), GT) define an expected utility for each combination of a policy and an observation $o$ that is observed with positive probability. If, say, the expected utility under $\pi_1$ is higher than the expected utility under $\pi_2$ conditional on every possible observation, we might expect the agent to never follow policy $\pi_2$. 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, Sect. 5) and Korzukhin (2020).

Finally, in Section 7, we conclude with a summary of the higher-level takeaways from our work and directions for future research.
2 Preliminaries

2.1 Single-player games of imperfect recall

A (single-player) game of imperfect recall, or scenario, is a tuple \((S, S_T, P_0, O, \omega, \{A_o\}_{o \in O}, T, u)\), 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 \(\omega : S - S_T \rightarrow O\) mapping non-terminal states onto the observations made by the agent in that state; for each observation \(o\), a finite set \(A_o\) of possible actions that the agent can choose from; a probabilistic, conditional transition mapping \(T\) that provides, for any current non-terminal state \(s\) with observation \(\omega(s) = o\) and any action \(a \in A_o\) taken by the agent, a probability distribution \(T(\cdot | s, a)\) over successor states; and a utility function \(u : S_T \rightarrow \mathbb{R}\) that maps terminal states onto real numbers.

As an example of a scenario, see the formalization of the absent-minded driver problem in Figure 1. Note that in scenarios with \(|O| = 1\), such as the absent-minded driver, we generally leave the element of \(O\) nameless and omit notation as to what the observation is in each state (i.e., we omit \(\omega\)). Throughout this paper, we will provide many more examples of scenarios of imperfect recall.

A (memoryless) policy is a probabilistic mapping \(\pi : O \rightarrow \Delta(A)\) that determines, for each observation \(o \in O\), a probability distribution \(\pi(\cdot | o)\) over actions. A history of the scenario is a finite sequence \(s_0...s_n\) of states where \(s_n \in S_T\) is a terminal state, and \(s_0, ..., s_{n-1} \notin S_T\) are non-terminal. Given a policy, the probability of a history is given by \(P(s_0...s_n \pi) = P_0(s_0) \prod_{i=1}^{n} (\sum_{a \in A} \pi(a | \omega(s_{i-1})) T(s_i | s_{i-1}, a))\).

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. To avoid these problems, we assume away infinite histories. Specifically, we assume throughout this paper that the game terminates with probability 1 when following any policy \(\pi\) from any state \(s \in S\). That is, we assume that for all states \(s\) and all policies \(\pi\), we have that \(\sum_{\pi=1}^{\infty} \sum_{s_1...s_n} P(s_0s_1s_2...s_n | s_0 = s, \pi) = 1\), where the inner sum is over all possible histories. For simplicity, this assumption is a little stronger than needed. Interestingly, to avoid infinite histories in expected utility calculations evidential decision theory requires weaker assumptions than causal decision theory, see Appendix A.

For any policy \(\pi\), state \(s\), and action \(a\), define \(Q(\pi)(s,a) := \sum_{\pi=1}^{\infty} \sum_{s_1...s_n} P(s_1 | s,a) P(s_2...s_n | s_1,\pi) u(s_n)\) to be the expected utility of being in state \(s\), choosing action \(a\) and then following policy \(\pi\). Define \(Q(\pi)(s) := \sum_{a \in A} \pi(a | \omega(s)) Q(\pi)(s,a)\) to analogously be the expected utility given that the current state is \(s\) and policy \(\pi\) is used. Finally, we use \(Q(P_0) := \sum_{s \in S} P_0(s) Q(\pi)(s)\) for the expected utility if an initial state is sampled from \(P_0\) and the agent follows \(\pi\).

There are many alternative ways to specify games of imperfect recall. The present one resembles 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 (sometimes also called stationary) policies, i.e., policies that only depend on the current observation, as opposed to the history of observations (cf. Littman, 1994; Li, Yin, and Xi, 2011). A more common represen-
tation of imperfect recall games is the tree representation of extensive-form games in game theory. In this setting, information sets take the role of the observations of our setting. That is, to express that the agent cannot distinguish between state \( s \) or state \( s' \), we assign \( s \) and \( s' \) the same observation, while in the extensive-form representation, \( s \) and \( s' \) would be in the same information set. There is also 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) (cf. Piccione and Rubinstein, 1997; Binmore, 1997; Detwarasiti and Shachter, 2005; Conitzer, 2019, Sect. “Imperfect Recall”; Emmons et al., 2022).

2.2 Two ex ante standards of rational choice

2.2.1 Ex ante optimal policies

Let \( \Pi \subseteq \Delta(A)^O \) be a set of policies. We call a policy \( \pi^* \in \Pi \) ex ante optimal in \( \Pi \) if \( \pi^* \in \arg \max \pi \in \Pi Q_{\pi}(P_0) \). The two most important cases of \( \Pi \) are \( \Pi = \Delta(A)^O \), which allows all mixed policies; and the set of the set of deterministic policies, which for each observation choose some action with probability 1, denoted by \( \Pi = A^O \). 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 \( \Delta(A)^O \to \mathbb{R} : \pi \mapsto Q_{\pi}(P_0) \).

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 mixed policy. For any \( p \), let \( \pi_p \) be the policy that continues with probability \( p \) and exits with probability \( 1-p \). The ex ante expected utility of \( \pi_p \) is \( Q_{\pi_p}(s_0) = P(s_0|s_1 \mid \pi_p) + 4P(s_0|s_1 4) = p^2 + 4p(1-p) \). The unique local and global optimum of this polynomial is \( p = 2/3 \). Hence, the unique globally and locally optimal policy is \( \pi_{2/3} \).

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 \( \pi^* \). In line with a number of previous works (Piccione and Rubinstein, 1997; Hitchcock, 2004; Draper and Pust, 2008; Briggs, 2010; Armstrong, 2011), the question we ask in this paper is whether ex ante optimality is consistent with what we will call de se theories of choice – where 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

A weaker standard of rationality is to avoid so-called Dutch books. Throughout this paper, a Dutch book policy is a policy that receives a negative payoff with probability 1 when there exists a different policy that guarantees a non-negative reward with probability 1.

\(^2\)Armstrong (2011) explicitly argues that even de se, agents should choose by finding and following an ex ante optimal policy.
Definition 1. Let $\mathcal{E}$ 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 $\pi$ for $\mathcal{E}$ a Dutch-book policy if it has a negative payoff with probability 1, i.e., if for all histories $s_0...s_n$ of $\mathcal{E}$, $P(s_0...s_n | \pi) > 0 \Rightarrow u(s_n) < 0$.

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 the 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 Beauty is told in the afternoon which day it is) was introduced by Lewis (2001).

Example 2 (Sleeping Beauty). 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, 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, due to imperfect recall, Beauty cannot tell whether it is Monday or Tuesday. Later she is told what day it is (but not how the coin came up). Two questions arise:
1. Upon waking up, what probability should Beauty assign to the coin having come up Heads?
2. Upon being told that it is Monday, what probability should Beauty assign to the coin having come up Heads?

A graphical description of this problem in our formalism is given in Figure 2.

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

On the first question, some – so-called halfers believe the answer is $1/2$ – and others believe the answer is $1/3$ – thirders. The simplest argument for the halfer position is that 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, i.e., people who answer the first question with $1/3$, tend to give the $1/2$ answer. After all, 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

9
the two positions. Halfers who give the $\frac{1}{2}$ answer for a second time are called double-halfers; we call halfers who give the $\frac{2}{3}$ answer single-halfers. The simplest argument for double-halfing is the same as the above argument for halving: 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 as the hypothesis that the coin came up Tails. It would therefore seem that a Bayesian reasoner should update toward the Heads hypothesis.

In the rest of this section, we provide procedures that generalize double-halfing, single-halfing, and thirding, respectively, to assign self-locating probabilities in any given scenario 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, Sect. 5). It is referred as compartmentalized conditionalization by Meacham (2008), as the minimal-reference class self-sampling assumption by Bostrom (2010), or simply “The Halfer Rule” by, 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 $s_0...s_n$.” To assign such probabilities, we in general (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 $P_{GDH}(i$-th in $s_0...s_n \mid \pi, o)$. Alternatively, we can think of GDH as assigning probabilities $P_{GDH}(s_0...s_n \mid \pi, o)$, since in any history $s_0...s_n$, we will take GDH to (uncontroversially) split probability mass equally among all time steps in $s_0, ..., s_n$, in which $o$ is observed (cf. Briggs, 2010, Sect. 2.3, and references therein). The crux of GDH is that $P_{GDH}(s_0...s_n \mid \pi, o)$ simply equals $P(s_0...s_n \mid \pi, o)$, which we define to be the non-self-locating probability of $s_0...s_n$ conditional on the fact that $o$ is observed at least once.

For the formal definition, define $\#(o, s_0...s_n) := \sum_{i=0}^{n-1} 2$ [w($s_i$) = o] to be the number of times $o$ is observed in $s_0...s_n$. Further, note that for any history $s_0, ..., s_n$, if w($s_i$) = o for some $i \in \{0, ..., n-1\}$, then

$$P(s_0...s_n \mid \pi, o) = \frac{P(s_0...s_n \mid \pi)}{\sum_{s'_0...s'_k} \sum_{i : w(s'_i) = o} P(s'_0...s'_k \mid \pi)},$$

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

**Definition 2.** Let $o$ be observed with positive probability under policy $\pi$. Then for all $s_0...s_n$ and $i = 0, 1, ..., n-1$, define $P_{GDH}(i$-th in $s_0...s_n \mid \pi, o) = 0$ if $w(s_i) \neq o$, and $P_{GDH}(i$-th in $s_0...s_n \mid \pi, o) = P(s_0...s_n \mid \pi, o)/\#(o, s_0...s_n)$ otherwise.
2.3.2 Generalized single-halfing

We move on to generalized single-halfing (GSH). GSH is also known as the (non-minimal reference class) self-sampling assumption (Bostrom, 2010). (Our version of GSH is, in our formalism, a maximum reference class version of the self-sampling assumption.) 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.

Like GDH, GSH is most naturally described as assigning probabilities to statements of the form, \(i\)-th in \(s_0\ldots s_n\). A natural interpretation of GSH is that it uses a prior \(P(i\text{-th in } s_0\ldots s_n \mid \pi) = \frac{1}{n} P(s_0\ldots s_n \mid \pi)\), which resembles the GDH probabilities. But then GSH performs a Bayes-like update. Recall that Bayes’ theorem states that

\[
P(x \mid y, z) = \frac{P(x \mid z) P(y \mid x, z) / (\sum_{x'} P(x' \mid z) P(y \mid x', z))}{P(y \mid z)}.
\]

Replacing \(x\) with the hypothesis \(i\)-th in \(s_0\ldots s_n\), \(y\) with the fact that I am observing \(o\), and \(z\) with the fact that the agent follows \(\pi\), we obtain the following definition of GSH.

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

\[
P_{\text{GSH}}(i\text{-th in } s_0\ldots s_n \mid \pi, o) = \frac{1}{\sum_{\omega(s_j) = o} P(s_0\ldots s_n \mid \pi)}.
\]

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

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

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

2.3.3 Generalized thirding

Finally, we describe generalized thirding (GT). GT was first given by Piccione and Rubinstein (1997 Sect. 5) as “consistency”; Bostrom (2010) calls it the self-indication assumption. To calculate \(P_{\text{GT}}(s \mid \pi, o)\), the agent asks: what fraction of the times that I observe \(o\) is \(s\) the current state? 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 \(#(s, s_0\ldots s_n) := \sum_{i=0}^{n} \mathbb{1}_{s_i = s}\) to be the number of occurrences of \(s\) in the history \(s_0\ldots s_n\), and \(C_\pi(s) := \sum_{s_0\ldots s_n} P(s_0\ldots s_n \mid \pi) \#(s, s_0\ldots s_n)\) to be the frequency of \(s\) under policy \(\pi\), i.e., the expected number of times \(s\) occurs under policy \(\pi\). Then, \(C_\pi(o) := \sum_{s \in S \setminus S_T : \omega(s) = o} C_\pi(s)\) is defined as the expected number of times that \(o\) is observed.

**Definition 4.** Let \(o\) be observed with positive probability under policy \(\pi\). Then \(P_{\text{GT}}(s \mid \pi, o) := 0\) if \(s \in S_T\) or \(\omega(s) \neq o\), and \(P_{\text{GT}}(s \mid \pi, o) := C_\pi(s)/C_\pi(o)\) otherwise.
Contrary to our definitions of GSH and GDH, the above definition does not assign probabilities \( P_{GT}(i \text{-th in } s_0...s_n | \pi, o) \). Such probabilities can easily be defined. However, we will not need them throughout the rest of this paper. Conversely, we could define GDH probabilities \( P_{GDH}(s | \pi, o) \) 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 recall depends on how she assigns probabilities (i.e., on 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 whatever distribution over actions she chooses now, she will also choose (and will have chosen in the past) in response to observing \( o \) at other decision points? This can be illustrated with 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 Beauty is told what day it is \( s'_{HMo}, s'_{TMo}, s'_{TTu} \). 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% probability. But what happens if the agent reasons not ex ante but de se, i.e., 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 \( s'_{STu,a} \) or \( s'_{STu,r} \)). 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:

- **(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 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.

- **(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 at one of the two decision points in the Tails branch. If I accept the bet, then at the other decision point I will also accept. Hence, in the Tails branch, accepting earns me \(4/3\). Since

---

\[3\]As usual, \( P_{GT}(i \text{-th in } s_0...s_n | \pi, o) = 0 \) if \( \omega(s_i) \neq o \). Otherwise, \( P_{GT}(i \text{-th in } s_0...s_n | \pi, o) = P(s_0...s_n | \pi) / C_\pi(o) \).
Figure 3: A graphical description of Example 3 in our formalism.

a 50% probability gain of $\frac{4}{3}$ outweighs a 50% probability loss of 1, I should accept the bet.

Piccione and Rubinstein (1997) discuss these two different styles of reasoning in their seminal work. Arntzenius (2002) is to our knowledge the first to make the connection to CDT versus EDT. But a few papers since have implicitly assumed CDT or EDT. For example, Draper and Pust (2008, Sect. 4) and Briggs (2010, Sect. 3.2) point out that Hitchcock’s (2004) analysis implicitly assumes CDT.

Having to choose between CDT and EDT, and between GT, GSH, and GDH, we have six theories for choice under imperfect recall to consider. However, we omit definitions and discussions of EDT+GT and CDT+GDH. Earlier work has shown, conclusively in our view, that these two combinations are vulnerable to Dutch books (Hitchcock, 2004, Sect. 6; Briggs, 2010, Sect. 3.3). Under the present agenda, we therefore have little more to say about CDT+GDH and EDT+GT.

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

To define EDT+GDH and EDT+GSH, we first note that our methods of assigning self-locating beliefs immediately allow us to assign expected utilities conditional on an observation and a policy.

**Definition 5.** For any policy $\pi$, we define the GDH/GSH expected utility, conditional on $o$ being observed, as $EU_{GDH/GSH}(\pi, o) := \sum_{i=0}^{n-1} \sum_{s_0, ..., s_n} P_{GDH/GSH}(i-th in s_0...s_n | \pi, o) u(s_n)$.

The crux of EDT relative to CDT is that it evaluates a distribution $\alpha \in \Delta(A)$ by the expected utility $EU_{GDH/GSH}(\pi_{o \rightarrow \alpha}, o)$ for some policy $\pi_{o \rightarrow \alpha}$ that chooses $\alpha$ upon observing $o$. After all, choosing $\alpha$ upon the current observation of $o$ is (definitive) evidence that the agent chooses $\alpha$ whenever she observes $o$.

But what happens if our scenario has multiple distinct observations that each occur with positive probability? Then, to evaluate a distribution $\alpha$ to play upon observing $o$, we need to calculate some expected utility $EU_{GDH/GSH}(\pi_{o \rightarrow \alpha}, o)$ where...
\( \pi_{a \rightarrow a}(\cdot \mid o) = a \) - but what should be the rest of \( \pi_{o \rightarrow a} \)? In principle, when evaluating a distribution \( \alpha \) upon \( o \), 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 variant 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 a general procedure here (cf. Sections 6.2 and 6.3).

To avoid the circularity, we take inspiration from the concept of Nash equilibrium in game theory and ratificationism in decision theory (Jeffrey [1965] 1983 Sect. 1.7; Weirich, 2016 Sect. 3.5; Bell et al. [2021]); an analogous method has also been used by Piccione and Rubinstein (1997) to formalize CDT in games of imperfect recall. We take any policy \( \pi \) and then merely ask for each observation (that occurs with positive probability given \( \pi \)): assuming the agent follows \( \pi \) for all observations other than \( o \), is it optimal as judged by EDT+GDH/GSH to play \( \pi(\cdot \mid o) \) upon observing \( o \)? This gives us a necessary condition for a policy \( \pi \) to be knowingly followed by an EDT+GDH/GSH agent.

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

**Definition 6.** We say that a policy \( \pi \in \Pi \) is compatible with EDT+GDH/EDT+GSH as restricted to \( \Pi \) if for all \( o \in O \) that are observed with positive probability under \( \pi \), \( \pi(\cdot \mid o) \in \arg \max_{\alpha \in \Delta(A) : \pi_{o \rightarrow a} \in P} \EU_{GDH/GSH}(\pi_{o \rightarrow a}, o) \).

Our calculations for Example 3 provide a first example of EDT+GDH/GSH reasoning. As a second example, we calculate the EDT+GDH/GSH-compatible policies in the absent-minded driver (Example 1). By Proposition 1 GSH and GDH give the same results in the absent-minded driver, so we only consider GDH. The calculation provides intuition for the proofs of our positive results in Section 4. Because there is only one possible observation in the absent-minded driver, for notational simplicity we drop the observation as an argument of \( \EU_{GDH} \) and \( P_{GDH} \). Letting \( \pi_p \) again be the policy that continues with probability \( p \), EDT+GDH requires that the agent continues with a probability from \( \arg \max_{p \in [0,1]} \EU_{GDH}(\pi_p) \). By definition, \( \EU_{GDH}(\pi_p) \) equals

\[
P_{GDH}(0\text{-th in } s_0|0 \mid \pi_p) \cdot 0 + P_{GDH}(0\text{-th in } s_0 s_1 4 \mid \pi_p) \cdot 4 + P_{GDH}(1\text{st in } s_0 s_1 4 \mid \pi_p) \cdot 4
\]
\[
+ P_{GDH}(0\text{-th in } s_0 s_1 1 \mid \pi_p) \cdot 1 + P_{GDH}(1\text{st in } s_0 s_1 1 \mid \pi_p) \cdot 1
\]
\[
= P(s_0 s_1 4 \mid \pi_p) \cdot 4 + P(s_0 s_1 1 \mid \pi_p) \cdot 1.
\]

Clearly, this is exactly the *ex ante* expected utility of \( \pi_p \). Thus, the *ex ante* optimal policy \( \pi_{2/3} \) (see Section 2.2.1) 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. We now give a simple scenario in which there are multiple compatible policies, one of which is a Dutch book policy (and thus also *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 costing $100,000 to repair. 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 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, then by accepting the offer she earns an extra $10; and if on Tuesday Alice believes that on Monday she accepted, then she better defuse the bomb.\footnote{There is also a third EDT+GDH/GSH-compatible policy: accepting on Monday with probability \( \frac{1}{5,000} \), and defusing on Tuesday with probability \( 1 - \frac{1}{10,000} \). When following this policy, the EDT+GDH/GSH agent is indifferent among 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.} We thus conclude the following.

Proposition 2. 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., \textit{ex ante} optimal or at least non-Dutch-book) compatible 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 + generalized thirding/single-halving

We now define causal decision theory (CDT) + GT (as per Piccione and Rubinstein [1997], who refer to it as “modified multiselves consistency”) and CDT+GSH. Many of the ideas here are analogous to those in the previous section. First we will define $\text{EU}_{\text{GT/GSH}}(\pi, o, a)$ to be the expected utility under GT/GSH probabilities of observing $o$, choosing $a$ now and otherwise following $\pi$, 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, $\pi(a \mid o)$ is positive only for actions $a$ that maximize $\text{EU}_{\text{GT/GSH}}(\pi, o, a)$.

**Definition 7.** For any policy $\pi$, any observation $o \in O$ observed with positive probability, and any $a \in A$, define $\text{EU}_{\text{GT/GSH}}(\pi, o, a) := \sum_{s \in S} \pi(s \mid \pi, o) \cdot Q_{\pi}(s, a)$.

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

Note that, in contrast to our definition in the case 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 $\text{EU}_{\text{GT/GSH}}(\pi, o, \alpha) = \sum_{a \in A} \alpha(a) \cdot \text{EU}_{\text{GT/GSH}}(\pi, o, a)$ However, it is then easy to see that choosing a distribution $\alpha$ maximizes $\text{EU}_{\text{GT/GSH}}(\pi, o, \alpha)$ if and only if all $a \in A$ with $\alpha(a) > 0$ maximize $\text{EU}_{\text{GT/GSH}}(\pi, o, a)$. We will not define CDT under a restricted set of policies (cf. Section 6.1).

Like EDT, CDT allows for multiple compatible policies, e.g., in Example 4.

**Proposition 3.** 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 $\bar{\pi}$ 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 8, 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 + generalized thirding

In this section, we review Piccione and Rubinstein’s (1997) Proposition 3) result that in every scenario the ex ante optimal policy is compatible with causal decision theory + generalized thirding. While Piccione and Rubinstein give a monolithic proof of this result, we first give a characterization of CDT+GT-compatible policies (Theorem 4). From the characterization, Piccione and Rubinstein’s result then follows (Corollary 5).

---

3 The analogous equality does not hold for EDT, i.e., in general $\text{EU}_{\text{GT/GSH}}(\pi_{a \rightarrow o}, o)$ is not necessarily equal to $\sum_{a \in A} \alpha(a) \cdot \text{EU}_{\text{GT/GSH}}(\pi_{a \rightarrow o}, o)$. The absent-minded driver serves as an example.
We prove these results in Appendix C. 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. $\pi(o \mid a)$. We define these in a way that takes into account that (infinitesimally) increasing $\pi(o \mid a)$ 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 thus define the derivative as follows.

**Definition 8.** Let $\pi$ be a policy, $a$ be an action and $o$ be an observation. Then for all $\varepsilon > 0$ define $\pi_{\varepsilon, a, o}(a' \mid o') = \pi(a' \mid o')$ if $o' \neq o$; $\pi_{\varepsilon, a, o}(a' \mid o') = (1 - \varepsilon)\pi(a' \mid o')$ if $o' = o$ and $a' \neq a$; and $\pi_{\varepsilon, a, o}(a' \mid o') = (1 - \varepsilon)\pi(a' \mid o') + \varepsilon$ if $o' = o$ and $a' = a$. Then define $\frac{d}{d\pi(o \mid a)} Q_\pi(P_0) \coloneqq \lim_{\varepsilon \to 0} (Q_{\pi_{\varepsilon, a, o}}(P_0) - Q_\pi(P_0))/\varepsilon$.

It turns out that the derivatives of $Q_\pi(P_0)$ are closely related to the causal expected utilities. For the following, define $\text{EU}_{\text{GT}}(\pi, o) \coloneqq \sum_{a \in A} \pi(a \mid o)\text{EU}_{\text{GT}}(\pi, o, a)$ to be the CDT+GT expected utility of following $\pi$ upon observing $o$.

It turns out that we can characterize CDT+GT in terms of the derivatives.

**Theorem 4.** A policy $\pi \in \Pi = \Delta(A)^O$ is CDT+GT compatible if and only if for all $o \in O$ and $a \in A$, $\frac{d}{d\pi(o \mid a)} Q_\pi(P_0) \leq 0$.

Theorem 4 can be viewed as a version of the policy gradient theorem in the theory of reinforcement learning (Jaakkola, Singh, and Jordan, 1994; Sutton et al., 1999; Theorem 1).

This characterization implies the following important corollary.

**Corollary 5** (Piccione and Rubinstein, 1997). Let $\pi$ be a globally ex ante optimal strategy from $\Pi = \Delta(A)^O$. Then $\pi$ is CDT+GT compatible.

Briggs (2010 Sect. 3.4 and 3.5) and Conitzer (2015b Sect. 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 + generalized double-halving

In this section, we prove novel positive results for EDT+GDH. Again, we first give a characterization of EDT+GDH-compatible policies (Theorem 5). 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.

**Theorem 6.** Let $\Pi \subseteq \Delta(A)^O$. A policy $\pi \in \Pi$ is EDT+GDH compatible in $\Pi$ if and only if for all $o \in O, \alpha \in \Delta(A)$ s.t. $\pi_{\alpha \rightarrow o} \in \Pi$ we have that $Q_{\pi}(P_0) \geq Q_{\pi_{\alpha \rightarrow o}}(P_0)$.
As a consequence of Theorem 6, \textit{ex ante} optimal policies are EDT+GDH compatible.

\textbf{Corollary 7.} Let $\Pi$ be any set of policies for $E$ and let $\pi$ be \textit{ex ante} optimal in $\Pi$. Then $\pi$ is compatible with EDT+GDH restricted to $\Pi$.

Briggs (2010, Sect. 3) gives a related result. It 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 6 and the equivalence between GDH and GSH on single-observation scenarios (Proposition 1) also directly imply the following.

\textbf{Corollary 8.} Let $E$ be a scenario that has only one observation (i.e., $|O| = 1$). Let $\Pi \subseteq A^O$. Then a policy $\pi \in \Pi$ is \textit{ex-ante} optimal if and only if $\pi$ is compatible with EDT+GDH/GSH restricted to $\Pi$.

Note that the same could not be said of CDT+GT, as shown by, e.g., Aumann, Hart, and Perry (1997, Sect. 5), Conitzer’s (2015) “Three Awakenings” (cf. our Example 8), and Korzukhin (2020).

With the help of Theorems 4 and 6, we can also obtain the following result.

\textbf{Corollary 9.} If a policy is EDT+GDH compatible (without any policy restriction), it is CDT+GT compatible.

Since Corollary 8 does not hold true for CDT, the converse of Corollary 9 also does not hold.

\section{On generalized single-halfing}

\subsection{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.\footnote{This version differs from Draper and Pust’s in two ways. First, we added small extra payoffs (+$\varepsilon$) to make sure that single-halfers \textit{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, as defined in this paper (Definition 3), 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. This is because adding an extra decision point on Sunday changes the GSH probabilities of Heads and Tails (at least as defined in Definition 3) on Monday from the usual $(2/5, 3/5)$ to $(3/5, 2/5)$. We think this is not an oversight on Draper and Pust’s part. Instead, we think that they assume a generalization in which the Sunday observation is in a different reference class than the Monday and Tuesday observations. Roughly, this means that the existence of the Sunday observation is viewed as relevant to updating on Monday and Tuesday. See Bostrom (2010) for a discussion of reference classes.}

We will specifically focus on CDT+GSH. We give a Dutch book against EDT+GSH in Section 5.5.

\textbf{Example 5} (adapted from Draper and Pust, 2008, Sect. 5). As a base scenario, take Example 2, 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

\footnotetext[3]{This version differs from Draper and Pust’s in two ways. First, we added small extra payoffs (+$\varepsilon$) to make sure that single-halfers \textit{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, as defined in this paper (Definition 3), 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. This is because adding an extra decision point on Sunday changes the GSH probabilities of Heads and Tails (at least as defined in Definition 3) on Monday from the usual $(2/5, 3/5)$ to $(3/5, 2/5)$. We think this is not an oversight on Draper and Pust’s part. Instead, we think that they assume a generalization in which the Sunday observation is in a different reference class than the Monday and Tuesday observations. Roughly, this means that the existence of the Sunday observation is viewed as relevant to updating on Monday and Tuesday. See Bostrom (2010) for a discussion of reference classes.}
Tails. On Monday, 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.

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 7. Consider the formalization 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 $\pi$, $Q_{\pi}(s_{Su}, \text{accept}) > Q_{\pi}(s_{Su}, \text{reject})$. Upon observing that it is Monday, since Beauty’s beliefs depend on $\pi$ and for mixed $\pi$ she might assign positive probability to multiple states. For simplicity assume that $\pi$ accepts the offer with probability 1 on both Monday and Sunday. Then omitting the normalizing constants, we obtain $P_{\text{GSH}}(s_{H}, Mo, a | mo, \pi_{\text{accept}}) = P_{\text{GSH}}(1\text{st in } s_{Su}s_{H}, Mo, a - 5 | mo, \pi_{\text{accept}}) \sim 1/2 \cdot 1/2 = 1/4$ and $P_{\text{GSH}}(s_{T}, Mo, a | mo, \pi_{\text{accept}}) = P_{\text{GSH}}(1\text{st in } s_{Su}s_{T}, Mo, a - 5 | mo, \pi_{\text{accept}}) \sim 1/2 \cdot 1/3 = 1/6$; 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 $P_{\text{GSH}}(s_{H}, Mo, a | mo, \pi_{\text{accept}}) = 3/5$ and $P_{\text{GSH}}(s_{T}, Mo, a | mo, \pi_{\text{accept}}) = 2/5$, respectively. Hence, the causal expected utility of accepting relative to rejecting is $3/5 \cdot (12 + \varepsilon) - 2/5 \cdot 18 = 3\varepsilon/5$. Hence, CDT+GSH accepts the second bet. With the help of Draper and Pust we have thus shown the following.

**Proposition 10** (Draper and Pust, 2008). There is a scenario in which the only CDT+GSH-compatible policy is a (deterministic) Dutch book policy.

Example 5 is not a Dutch book against evidential decision theory + GSH as defined in this paper. We show and discuss this in Appendix E.1.

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

We now defend single-halfing against Draper and Pust’s Dutch book. In particular, we argue that single-halfers should strictly prefer to reject the bet on Sunday! This may be surprising since accepting the Sunday bet seems unproblematic. For instance, accepting the Sunday bet is *ex ante* optimal. Our approach will therefore not address the concern of *ex ante* suboptimality: we will have the agent reject the Sunday bet and accept the Monday bet, which is *ex ante* suboptimal. In the following we first give the argument informally and then make it formal.

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
Figure 5: A graphical formalization of Example 5 (adapted from Draper and Pust, 2008 Sect. 5). In this formalization, the only CDT+GSH-compatible policy is accepting the bet on both Sunday and Monday, which is a Dutch book.
towards the Heads hypothesis. In particular, I should not bet at (close to) even odds that the coin will come up Tails.

We now want to make the argument formal. First, we can easily verify that GSH (as defined in Definition 3) indeed updates toward the Heads histories, even upon observing that it is Sunday. For simplicity, consider again the policy \( \pi \) 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

\[ P_{\text{GSH}}(0-\text{th in } s_{\text{Su}} | su, \pi_{\text{accept}}) = \frac{3}{5} \]

and

\[ P_{\text{GSH}}(0-\text{th in } s_{\text{T-Su}} | su, \pi_{\text{accept}}) = \frac{2}{5} \],

respectively.

So why does Draper and Pust’s Dutch book work against CDT+GSH? The problem lies in how CDT as defined in Definition 7 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 \( s_{\text{Su}} \), 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 \( \frac{3}{5} \) not \( \frac{1}{2} \).

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 (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 7.

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 that

\[ P_{\text{GSH}}(s_{\text{H-Su}} | su, \pi_{\text{accept}}) = \frac{3}{5} \] and \( P_{\text{GSH}}(s_{\text{T-Su}} | su, \pi_{\text{accept}}) = \frac{2}{5} \). 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*

Generalizing the insight of the previous section, we now describe a general theory CDT+GSH* that assumes that all 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.

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

---

7This line of argument would not work if we considered a version of the self-sampling assumption in which the Sunday observation is in its own reference class. As noted in footnote Footnote 6, this is plausibly what Draper and Pust had in mind.
Figure 6: An alternative formalization of Example 5 (adapted from Draper and Pust, 2008, Sect. 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.
Definition 9. We say that a scenario $E$ randomizes only in the beginning if $T(s' \mid s, a)$ is 0 or 1 for all $s, s' \in S$ and all $a \in A_{\pi(s)}$.

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. We do not do this here, because it is intuitive but formally cumbersome. 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.

How is CDT+GSH* supposed to deal with policy randomization? If random state transitions cause problems, do random policies cause the same problems? The answer is yes and we 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 10. Let $E$ be a scenario that randomizes only in the beginning. We say that history length is choice independent in $E$ if for each $s_0$ with $P_0(s_0) > 0$, there is a natural number $\text{len}(s_0)$ such that for all $\pi$, $\sum_{s_1 \ldots s_{\text{len}(s_0)}} P(s_0 s_1 \ldots s_{\text{len}(s_0)} \mid \pi, s_0) = 1$, where the sum is over all histories of length $\text{len}(s_0)$.

In words, history length is choice independent in $E$ if the initial state uniquely determines the length of the history independently of the agent’s choices.

The key realization now is that if history length is choice independent, (CDT+)GSH is like (CDT+)GT, except that gives lower weight to utilities achieved in longer histories.

Theorem 11. Let $E$ be a scenario that randomizes only in the beginning and where history length is choice-independent. Let $\hat{E}$ be the scenario that is equal to $E$, except that $\hat{P}_0(s_0) \sim P_0(s_0) / \text{len}(s_0)$. Then any (potentially mixed) policy is CDT+GSH-compatible in $E$ if and only if it is CDT+GT-compatible in $\hat{E}$.

Note that one could equivalently state Theorem 11 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”.

As an immediate consequence of Theorem 11 and Theorem 4, we can characterize CDT+GSH compatibility in such scenarios in terms of the policy derivatives of $Q_{\pi}(P_0)$. Moreover, Theorem 11 implies the following Dutch book avoidance result for CDT+GSH*.

Corollary 12. Let $E$ 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 $E$.

5.4 CDT+GSH* fails when choices affect history length

What happens if we the agent’s choices affect the length of the history, as in the absent-minded driver? We will argue that natural generalizations of the ideas from the previous
sections don’t save CDT+GSH* from Dutch books.

5.4.1 Random choices pose the same problem as random state transitions

It is natural to suspect that if the agent’s choices can affect the history length, policy randomization causes 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 7) in Draper and Pust’s case is that the environment flips a coin midway through the scenario and the coin flip determines the history length, then CDT+GSH assigns different probabilities to the coin flip’s outcome before versus after the coin is flipped. The following result shows that the same problem arises if the agent flips a coin midway through.

**Proposition 13.** There is a scenario that randomizes only in the beginning and in which 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 E.3. We will then also show in Appendix E.4 how moving the agent’s policy randomization to the beginning of the scenario solves the example given in Appendix E.3.

5.4.2 Viewing random choices as predetermined fails – A Dutch book against CDT+GSH*

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

To be more precise, we will imagine that for any scenario \( \mathcal{E} \) and any policy \( \pi \) for \( \mathcal{E} \), we construct some scenario \( \mathcal{E}^\pi \) which is like \( \mathcal{E} \) except that there is an action \( a_\pi \) that implements \( \pi \) by accessing some feature of the state that is determined at random. To determine whether \( \pi \) is CDT+GSH* compatible we can then ask: is the policy of always choosing \( a_\pi \) CDT+GT compatible in \( \mathcal{E}^\pi \)? In Appendix E.4, we show how this approach solves the example that we use to prove Proposition 13.

Unfortunately, this approach does not seem to work in general. In general, there need not even exist a policy that is CDT+GSH* compatible in this sense. We discuss this in Appendix E.5. More importantly, we here give a scenario in which the only CDT+GSH* compatible policy is a Dutch book. 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 + generalized single-halfing.

**Example 6.** 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 by the number of times she chooses \( a_1 \) in these situations as follows: \( 0 \rightarrow 0, 1 \rightarrow 1, 2 \rightarrow 24 \).
−1, 3 → −ε. Here, ε is some small but positive number, e.g., ε = 1/100. Afterward, if \( a_1 \) was chosen exactly once (for a reward of 1) the agent faces the same decision problem between \( a_0 \) and \( a_1 \) another \( K \) times (for some large \( K \)). The agent’s choices in these \( K \) situations do not affect her final reward – her reward remains 1.

We now argue informally that Example 6 is a Dutch Book against CDT+GSH*; see Appendix E.7 for a detailed, rigorous analysis. 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 \( \pi \) 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 \( a_\pi \) mattering at all, she is unlikely (specifically with probability approaching 0 as \( K \) goes to \( \infty \)) to be on track to play \( a_1 \) exactly once if she plays \( a_\pi \). Instead, conditional on her choices mattering, following the policy (playing \( a_\pi \)) will likely lead to \( a_1 \) being played either 0 or 2 times. This is because for all policies \( \pi \), 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. Specifically, as we show in Appendix E.7 the probability of playing \( a_1 \) exactly once is at most \( \frac{3}{2} \) times the probability of playing \( a_1 \) zero or two times. Since the branch in which \( a_1 \) is played exactly once contains many additional inconsequential decision situations, the CDT+GSH* agent believes it is very likely (probability approaching 1 as \( K \to \infty \)) that if her choice matters at all she is on track (by playing \( a_\pi \)) to play \( a_1 \) zero or two times. Given this belief, CDT recommends playing \( a_1 \) over playing \( a_\pi \) or \( a_0 \).

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

EDT+GSH as discussed in this paper avoids Draper and Pust’s Dutch book, both in its basic form (see Appendix E.1) and in the variant that moves randomization to the beginning of the scenario (Figure 6). Nonetheless, there are simple Dutch book scenarios for EDT+GSH 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 14.** 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 7.** Let \( \varepsilon > 0 \). The researchers flip a fair coin. On Sunday, Beauty is offered a bet that wins \( S1 – \varepsilon \) if the coin came up Heads and loses \( S1 – 2\varepsilon \) if the coin came up Tails. The researchers then put Beauty to sleep. If the coin came up Tails, then Beauty is awoken once on Monday and again on Tuesday (without being told what day it is). If the coin came 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 for the bet to become into effect once. This second bet loses \( S1 \) if the coin comes up Heads and wins \( S1 – 3\varepsilon \) if the coin comes up Tails. For a graphical description of this scenario in our formalism, see Figure 7.
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 that the coin came up 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 $\frac{2}{3}$ to Tails. If Beauty also uses EDT, then accepting with probability $p$ increases her GSH-expected utility by $\frac{2}{3}p^2(\$1 - 3\varepsilon) - \frac{1}{3}p(\$1 - 2\varepsilon)$ relative to not accepting. For small enough (but still positive) $\varepsilon$, 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.
6 Discussion

6.1 CDT needs randomization

Our main positive result for EDT+GDH (Corollary 7) applies even if we restrict the agent to, for example, the set of deterministic policies. Our main positive compatibility results for CDT+GT and CDT+GSH(*) (i.e., Corollary 5 (Piccione and Rubinstein, 1997) and Corollary 12), on the other hand, only apply in the case of unrestricted randomization (i.e., they require that $\Pi = \Delta(A)^G$). 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 15.** 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 15 with the following variant of Conitzer’s (2015) “Three Awakenings”:

**Example 8.** 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 $\frac{1}{2} - 1/(2\sqrt{2}) \approx 0.15$, which (by Corollaries 5 and 7) 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 7). 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 −1 to −2. However, always choosing $a_1$ is a Dutch book policy.

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, Sect. IV.1). It is outside the 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 termi-
nology, that the only EDT compatible policy (regardless of self-locating beliefs) is a Dutch book. This seems to contradict our Corollary. 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 from our definition that conditional on a policy $\pi$ and an observation $o$, EDT evaluates the value of any distribution $\alpha \in \Delta(A)$ via the expected value of $\pi_{o \rightarrow \alpha}$. That is, the the EDT(+GDH) agent considers that if she chooses $\alpha$ upon $o$ now, then on all other instances of observing $o$ (including past ones), she also will choose and will have chosen $\alpha$. For all other $o' \neq o$, on the other hand, we imagine that choosing $\alpha$ upon $o$ gives no evidence about choice in $o'$. Conitzer gives a case in which it is very plausible that $o_1 \rightarrow \alpha$ also implies that $o_2 \rightarrow \alpha$ 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 9.** We use the same three equiprobable branches as in Example 10. 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 $\frac{2}{3}$ in branch Z and pays $-1$ in branch X/Y. The game is represented graphically in Figure 8.

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 $y$. 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 B.

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 $\frac{4}{3}$ (not just $\frac{2}{3}$), because if she is in branch Z and she accepts, she accepts twice for a reward of $\frac{4}{3}$. Whereas, if she is in branch Z and she rejects, she rejects twice for a reward of 0. A
Figure 8: A graphical formalization of Example 9.

The gain of $\frac{4}{3}$ 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 cause of failure of the EDT agent as considered in Conitzer’s argument is a discrepancy between decision points that the agent can “evidentially control”, and decision points about which the agent thinks, “I might be this decision point”. In Example 9 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 the present paper) the agent’s goal is independent of whether he is the original or the copy, CDT recommends one-boxing. 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, and we can similarly “save”

---

8 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).

9 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 Sect. 10; for further discussion, see Gauthier, 1989; Drescher, 2006 Sect. 6.2; Arntzenius, 2008 Sect. 7; Meacham, 2010 Sect. 3.2.2). Other such examples include Parfit’s (1984) hitchhiker (Barnes, 1997), XOR Blackmail (Levinstein and Soares, 2009).
EDT by having the agent identify with anything in the environment that can be used to predict the agent.10

6.3 The multiplicity of compatible policies
As we have seen (e.g., in Examples 4 and 8) 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 should draw some satisfaction from these results. However, we might also wonder whether 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 unaware of any work that proposes solutions to this problem.

Note that the multiplicity of compatible policies is a bigger problem for CDT than it is for EDT. First, by Corollary 8, in the single-observation case (|O| = 1), EDT+GDH and EDT+GSH face no multiplicity problem. 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 Sect. 5, and Korzukhin, 2020). This matters especially if we believe that the single-observation case is particularly 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 Sect. IV.3 and references therein). Second, by Corollary 9 the set of EDT+GDH compatible policies is a subset and in many cases a strict subset of the set of CDT+GT compatible policies. Nevertheless, both EDT and CDT face a problem of multiplicity when |O| > 1.

We would rather like to join Aumann, Hart, and Perry, 1997 and Korzukhin 2020 in raising awareness for the problem of the multiplicity of compatible policies. We do this by showing that an intuitively compelling approaches leads to bad policies. 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 \( \pi \), most naturally by calculating expected utilities \( \text{EU}_{\text{GDH/GDH/GT}}(\pi, o) \) and comparing them across policies \( \pi \). 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 \( \text{EU}_{\text{GT}}(\pi, o) \), then the agent will in general not follow a CDT+GT-compatible policy. (We give an example in Appendix G.) 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 \( \pi \) is better than another CDT+GT-compatible policy \( \pi' \), then a CDT+GT agent would not follow \( \pi' \). This resembles the use of Pareto optimality as a goal in multi-agent interactions. The idea is more natural in single-player scenarios.

10 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).
of imperfect recall, however, since we might expect the different decision perspectives of a single player to be better at coordinating. 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 EU_{GDH}, EU_{GSH} or EU_{GT} to judge policies). Thus, without using any ex ante perspective, de se decision theories can avoid the certain loss in this scenario. Unfortunately, the following result shows that such reasoning can lead the agent badly astray.

**Theorem 16.** There is a scenario $\mathcal{E}$ with the following properties.

- $\mathcal{E}$ 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 $\pi$ other than $\tilde{\pi}$ and all observations $o$, $EU_{GDH/GT/GSH}(\tilde{\pi}, o) > EU_{GDH/GT/GSH}(\pi, o)$.

Note that the first item means that this scenario is relatively unproblematic for CDT+GSH (see Theorem 11).

Aumann, Hart, and Perry (1997, Sect. 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 8 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 H, we then extend the scenario to also apply to CDT+GSH*.

**Example 10.** At the beginning, the scenario randomizes uniformly between three possibilities: X) The agent observes $x$ once. Y) The agent observes $y$ once. Z) The agent observes $x$ once and then $y$ once. 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 9.

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 10 has the properties claimed in Theorem 16. 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 
\[ \text{EU}_{\text{GDH/GT}}(\tilde{\pi}, x) = \text{EU}_{\text{GDH/GT}}(\tilde{\pi}, y) = -\frac{3\epsilon}{2}. \]
(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 9, 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 \( \pi_{\text{bet}} \), which is to bet in both \( x \) and \( y \). This is also the \textit{ex ante} optimal policy. However, upon observing \( x \) or \( y \), the GDH/GT expected utility is 
\[ \text{EU}_{\text{GDH/GT}}(\pi_{\text{bet}}, x) = \text{EU}_{\text{GDH/GT}}(\pi_{\text{bet}}, y) = \frac{1}{2} \cdot \frac{1}{2} \cdot (-\frac{4}{3}) = -\frac{1}{6}, \] which is less than \(-\frac{3\epsilon}{2}\) for small enough \( \epsilon \).

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 \((x \text{ and } y)\) (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 \( \pi_p \), that bets with probability \( p \) and pays with probability \( 1 - p \) in both \( x \) and \( y \). It is easy to see that 
\[ \text{EU}_{\text{GDH/GT}}(\pi_p, x/y) \] can be written as a convex combination of the analogous expected utilities for four different deterministic policies, namely the four different policies that map \( x \) and \( y \) to bet and pay.\footnote{In this specific case, this can be formally verified as follows:}

\[
\begin{align*}
\text{EU}_{\text{GDH/GT}}(\pi_p, x) &= \frac{1}{2} (p + (1 - p)(-\epsilon)) \\
&\quad + \frac{1}{2} (p^2 + 2p(1 - p)(-\epsilon)) \\
&\quad + \frac{1}{2} (p^2 + 2(1 - p)(-\epsilon)) \\
&\quad = \frac{1}{2} (p^2 - p + (1 - p)(-\epsilon)) \\
&\quad + \frac{1}{2} (p^2 - (1 - p)(-\epsilon)) \\
&\quad = \frac{1}{2} (p^2 + (1 - p)(-\epsilon)) + \frac{1}{2} (1 - p)(-\epsilon)
\end{align*}
\]

\[
\begin{align*}
\text{EU}_{\text{GDH/GT}}(\pi_p, y) &= \frac{1}{2} (p + (1 - p)(-\epsilon)) \\
&\quad + \frac{1}{2} (p^2 + 2p(1 - p)(-\epsilon)) \\
&\quad + \frac{1}{2} (p^2 + 2(1 - p)(-\epsilon)) \\
&\quad = \frac{1}{2} (p^2 - p + (1 - p)(-\epsilon)) \\
&\quad + \frac{1}{2} (p^2 - (1 - p)(-\epsilon)) \\
&\quad = \frac{1}{2} (p^2 + (1 - p)(-\epsilon)) + \frac{1}{2} (1 - p)(-\epsilon)
\end{align*}
\]

\[ \text{EU}_{\text{GDH/GT}}(\pi_p, x/y) = \frac{1}{2} \cdot \frac{1}{2} \cdot (-\frac{4}{3}) = -\frac{1}{6}, \] which is less than \(-\frac{3\epsilon}{2}\) for small enough \( \epsilon \).

7 Conclusion

Together with a body of existing work, the present paper shows which \textit{de se} methods of choice abide \textit{ex ante} standards of rational choice. We find that causal decision
Figure 9: A graphical formalization of Example 10.
theory works when combined with generalized thirding (a.k.a. consistency and the self-indication assumption), while evidential decision theory works when combined with generalized double-halfing (aka Z-consistency, the minimum-reference class self-sampling assumption). Other combinations are in general vulnerable to Dutch books. An important negative result is that generalized single halfing a.k.a. the self-sampling assumption is in general vulnerable to Dutch books, whether combined with CDT or EDT. This is especially important considering that single-halfing is intuitively appealing and has been used in a number of anthropic arguments, including the famous doomsday argument. The present work (in accord with Draper and Pust (2008)) suggests that we reject these arguments.

While our work aims to give a complete analysis of de se versus ex ante rational choice, it nonetheless opens many avenues for further work. Throughout this paper, we have found connections between foundational areas of self-locating beliefs (and anthropics) and the decision theory of Newcomblike problems: the main results show (in line with Briggs [2010]) suggests that positions on Newcomb’s problem (one- versus two-boxing) commit ourselves to positions on the Sleeping Beauty problem and vice versa; in Section 6.1 we point out that both in Newcomb-like problems and games of imperfect recall, CDT hinges on randomization in a way that EDT does not; in Section 6.2 we connected Conitzer’s alleged imperfect recall counterexample to EDT to alleged Newcomblike (perfect recall) counterexamples to CDT and EDT; in Appendix A we show how, like some Newcomblike problems, CDT is sensitive to impossible counterfactuals in a way that EDT isn’t. We believe that future work will benefit from understanding Newcomblike problems through games of imperfect recall and vice versa.

References

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


Li, Yanjie, Baoqun Yin, and Hongsheng Xi (2011). “Finding optimal memoryless policies of POMDPs under the expected average reward criterion”. In: European Jour-


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.


A 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 \( \pi \), the probability that a terminal state is reached at some point is 1. More formally, we need to assume that

\[
\sum_{s_0, \ldots, s_n} P(s_0, \ldots, s_n \mid \pi) = 1,
\]

where the sum is over all histories that end in some terminal state \( s_n \). Figure Figure 10 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 11. 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

---

Figure 10: A scenario in which there exists a policy (namely always playing \( b \)) that does not lead to a terminal state.


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.


problem assign \textit{ex ante} expected utilities $Q_{\pi}(s_0)$ to all policies. (In particular, the \textit{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 $\sum_{s_0 \ldots s_n} P(s_0 \ldots s_n | \pi) = 1$ and not the stronger assumption given in the main text.

For CDT, on the other hand, the scenario of Figure 11 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 $s_0 s_1 s_1 s_1 \ldots$ 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 that for every state $s$, policy $\pi$, and action $a$, choosing $a$ in $s$ and then following $\pi$ reaches a terminal state with probability 1 (even if $s$ is reached with probability 0 given $\pi$).

CDT is unable to pass judgment in situations that are unproblematic from an \textit{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). Similarly, in the scenario of Figure 11 CDT gives weight to the event that it chooses $b$ with probability 1 but chooses $a$.

\footnote{Also see a blog post by Oesterheld (2017).}
Figure 12: 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.

B 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 12. 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 $\pi$ and construct a new policy $\pi'$ with $\pi'(\cdot|a) = \pi(\cdot|b)$ and $\pi'(\cdot|b) = \pi(\cdot|a)$, then $Q_{\pi'}(s_0) = Q_{\pi}(s_0)$. (In the graph of Figure 12, 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 $o_2$.

C Proofs of Lemma 17, Theorem 4 and Corollary 5

Recall our definition of derivatives with respect to the policy:

**Definition 8.** Let $\pi$ be a policy, $a$ be an action and $o$ be an observation. Then for all $\varepsilon > 0$ define $\pi_{\varepsilon,a,o}(a' | o') = \pi(a'| o')$ if $o' \neq o$; $\pi_{\varepsilon,a,o}(a' | o') = (1-\varepsilon)\pi(a'| o')$ if $o' = o$ and $a' \neq a$; and $\pi_{\varepsilon,a,o}(a' | o') = (1-\varepsilon)\pi(a'| o') + \varepsilon$ if $o' = o$ and $a' = a$. Then define \( \frac{d}{d\pi(a|o)} Q_\pi(P_0) := \lim_{\varepsilon \downarrow 0} (Q_{\pi_{\varepsilon,a,o}}(P_0) - Q_\pi(P_0))/\varepsilon \).

**Lemma 17.** For any $o$ observed with positive probability, $\frac{d}{d\pi(a|o)} Q_\pi(P_0) = C_\pi(o)(\text{EU}_{GT}(\pi,o,a) - \text{EU}_{GT}(\pi,o))$. 

40
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 \( \varepsilon \)? It is helpful to focus on the case where \( \pi(a \mid o) = 0 \), i.e., where \( \pi \) would otherwise never take action \( a \) when observing \( o \). The crucial idea is that as \( \varepsilon \to 0 \), the probability that \( \pi_{\varepsilon,a,o} \) plays \( a \) multiple times in \( o \) diminishes at a rate on the order of \( \varepsilon^2 \). In contrast, the probability that \( \pi_{\varepsilon,a,o} \) plays \( a \) just once diminishes at a rate on the order of \( \varepsilon \). Therefore, 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 \( \pi \), as compared to always following \( \pi \). 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 \( \pi \), as \( \varepsilon \to 0 \). Thus, the expected effect of selecting \( a \) once in \( o \) is \( \text{EU}_{\text{GT}}(\pi, o, a) - \text{EU}_{\text{GT}}(\pi, o) \). The factor of \( C_{\pi}(s) \) reflects the fact that for any given \( \varepsilon \) 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 \( \pi \).

From Lemma [17] we directly obtain the following result.

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

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

\[
1/\varepsilon Q_{\pi_{\varepsilon,a,o}}(P_0) = 1/\varepsilon \sum_{x_0 \ldots x_n} P_0(x_0) \left( \prod_{i=0}^{n-1} T(s_{i+1} \mid s_i, \pi_{\varepsilon,a,o}) \right) u(s_n)
\]

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 \( \varepsilon \) (the exponent of \( \varepsilon \)), which intuitively is the number of times in the history that the \( \varepsilon \)-probability deviation from \( \pi \) occurs. So the order 0 term is simply

\[
1/\varepsilon \sum_{x_0 \ldots x_n} (1 - \varepsilon)^{\#(o,x_0 \ldots x_n)} P_0(x_0) \left( \prod_{i=0}^{n-1} T(s_{i+1} \mid s_i, \pi) \right) u(s_n)
\]

For small \( \varepsilon \), 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 \( x_0 \ldots x_n \) in \( Q_{\pi}(P_0) \).

The order 1 term is

\[
\sum_{x_0 \ldots x_k \cdot \omega(x_k) = o} (1 - \varepsilon)^{\#(o,x_0 \ldots x_n) - 1} P_0(x_0) T(s_{k+1} \mid a, s_k) \left( \prod_{i \neq k} T(s_{i+1} \mid \pi, s_i) \right) u(s_n).
\]

Note that the \( \varepsilon \) probability of choosing \( a \) as opposed to choosing from \( \pi \) in \( s_k \) is canceled out by \( 1/\varepsilon \). As \( \varepsilon \to 0 \), \( (1 - \varepsilon)^{\#(o,x_0 \ldots x_n) - 1} \to 1 \) for all \( x_0, \ldots, x_n \). Hence the order 1
term converges to

\[
\sum_{s_0 \ldots s_n: \omega(s_k) = o} P(s_0) T(s_{k+1} | a, s_k) \left( \prod_{i \neq k} T(s_{i+1} | \pi, s_i) \right) u(s_n)
\]

\[
= \sum_{s \in S: \omega(s) = o} \sum_{\text{prefix } s_0 \ldots s_k: s_k = s} \prod_{i=0}^{k-1} T(s_{i+1} | \pi, s_i)
\]

\[
= C_\pi(s) \prod_{i=k+1}^{n} T(s_{i+1} | \pi, s_i) u(s_n)
\]

\[
= \sum_{s \in S: \omega(s) = o} C_\pi(s) Q_\pi(a, s).
\]

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

We conclude that

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

Finally, notice that this sum is by Definition [4] equal to

\[
\sum_{s \in S: \omega(s) = o} C_\pi(o) P_{GT}(s | \pi, o) (Q_\pi(s) - Q_\pi(s))
\]

\[
= C_\pi(o) (EU_{GT}(s, o, a) - EU_{GT}(\pi, o)) \quad \square
\]

Lemma [17] implies the following.

**Theorem 4.** A policy \( \pi \in \Pi = \Delta(A)^O \) is CDT+GT compatible if and only if for all \( o \in O \) and \( a \in A \), \( \frac{d}{d\pi(o|o)} Q_\pi(P_0) \leq 0 \).

Theorem 4 in turn directly implies Corollary 5 – clearly the derivative at a global optimum must be non-positive in all directions.

**Corollary 5** (Piccione and Rubinstein, 1997). Let \( \pi \) be a globally ex ante optimal strategy from \( \Pi = \Delta(A)^O \). Then \( \pi \) is CDT+GT compatible.

**D Proofs of Theorem 6 and Corollary 9**

**Theorem 6.** Let \( \Pi \subseteq \Delta(A)^O \). A policy \( \pi \in \Pi \) is EDT+GDH compatible in \( \Pi \) if and only if for all \( o \in O, a \in A \) s.t. \( \pi_o \to a \in \Pi \) we have that \( Q_\pi(P_0) \geq Q_{\pi_o \to a}(P_0) \).
Proof. Let $\pi$ be a policy. Note that neither side of the claimed equivalence puts any restrictions on what $\pi$ 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 $\pi$, it is

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

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

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

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

We thus get that

$$\arg\max_{\alpha \in \Delta(A)} \sum_{\alpha \in \Pi} \sum_{i=1}^{n-1} P_{\text{GDH}}(i\text{-th in } s_0 \ldots s_n \mid \pi_{\alpha \rightarrow o}, o) u(s_n) = \arg\max_{\alpha \in \Delta(A)} \sum_{\alpha \in \Pi} \sum_{i, o(s_i) = o} P(s_0 \ldots s_n \mid \pi_{o \rightarrow \alpha}) \#(o, s_0 \ldots s_n) P(o \mid \pi_{o \rightarrow \alpha}) u(s_n),$$

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

$$= \arg\max_{\alpha \in \Delta(A)} \sum_{\alpha \in \Pi} \frac{1}{P(o \mid \pi_{o \rightarrow \alpha})} \sum_{s_0 \ldots s_n \text{ with } o} P(s_0 \ldots s_n \mid \pi_{o \rightarrow \alpha}) u(s_n).$$

Now note that $P(o \mid \pi_{o \rightarrow \alpha})$ is constant in $\alpha$, 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

$$\arg\max_{\alpha \in \Delta(A)} \sum_{\alpha \in \Pi} P(s_0 \ldots s_n \mid \pi_{o \rightarrow \alpha}) u(s_n).$$

Finally, this argmax equals

$$\arg\max_{\alpha \in \Delta(A)} \sum_{\alpha \in \Pi} P(s_0 \ldots s_n \mid \pi_{o \rightarrow \alpha}) u(s_n).$$

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

$$\pi(\cdot \mid o) \in \arg\max_{\alpha \in \Delta(A)} Q_{\pi_{o \rightarrow \alpha}}(P_0),$$

as claimed. \qed
Corollary 9. 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 \( \pi \) be any policy that is not compatible with EDT+GDH. Then by Theorem 4, there is an \( o \in O \) observed with positive probability when following \( \pi \) and an action \( a \in A \) such that \( d/\pi(a \mid o)Q_\pi(P_0) > 0 \). Hence, for sufficiently small \( \epsilon \), \( Q_\pi(\epsilon, a, o)(P_0) > Q_\pi(P_0) \). By Theorem 6, \( \pi \) is not EDT+GDH compatible.

E  Proofs on single-halfing

E.1 EDT+GSH avoids the Dutch book in Example 5

We here show that EDT+GSH avoids the Dutch book in Example 5, even if we use the formalization of Figure 5.

For now, let \( \pi \) be a policy that accepts the second bet.

\[
EU_{GSH}(a \mid su, \pi) = P_{GSH}(0-th in s_{Su}, Mo, a \mid s_{su} \rightarrow a)(-15) + P_{GSH}(0-th in s_{Su}, Mo, a \mid s_{H, Tu, a, r} \mid s_{su} \rightarrow a)(15 + \epsilon)
\]

Now again \( P_{GSH}(0-th in s_{Su}, Mo, a \mid s_{su} \rightarrow a) = 3/5 \) and \( P_{GSH}(0-th in s_{Su}, Mo, a \mid s_{H, Tu, a, r} \mid s_{su} \rightarrow a) = 2/5 \). Thus, \( EU_{GSH}(a \mid su, \pi) < 0 \). Clearly, \( EU_{GSH}(r \mid su, \pi) = 0 \). Thus, if \( \pi \) rejects the bet on Monday/Tuesday, EDT+GSH prefers rejecting the bet on Sunday. It can easily be shown that more.

This result is again contrary to Draper and Pust’s analysis. The underlying difference is that we use a version of GSH with a single reference class, cf. Footnote 6.

E.2 Proofs of Theorem 11 and Corollary 12 and Proposition 19

For any prefix history \( s_0 \ldots s_i \), i.e., any history that doesn’t end in a terminal state, define

\[
P_{GT}(s_0 \ldots s_i \mid \pi, o) := \frac{P(s_0 \ldots s_i \mid \pi)}{C_\pi(o)}
\]

to be the generalized thirder’s probability of being in state \( s_i \) after the prefix history \( s_0 \ldots s_{i-1} \) occurred.

We will also use the random variable \( H \) for the history of the scenario. For any history \( s_0 \ldots s_n \) (that ends in a terminal state \( s_n \in S_T \) as usual), we define \( \text{len}(s_0 \ldots s_n) = n \) 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 18. Let \( \pi \) be a policy. Then \( \pi \) 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
Let \( \pi \) be a deterministic policy. Let \( E \) be a scenario that randomizes only in the beginning and let \( \hat{\pi} \) be a deterministic policy. Then notice that each initial state \( P \) and \( \hat{\pi} \) in the beginning and let \( E \) restrict attention to deterministic policies. Let \( E \) be a scenario that randomizes only in the beginning and let \( \hat{\pi} \) be a deterministic policy. Then notice that each initial state \( \pi \) in the beginning and let \( E \) be a deterministic policy. Then notice that each initial state \( \pi \) in the beginning and let \( E \) be a deterministic policy. Then notice that each initial state.

**Proof.** By Definition \[ \text{CDT+GSH requires that for all } o \text{ that are observed with positive probability, the agent choose from} \]

\[
\arg \max_{a \in A} \sum_{\pi \in \mathcal{P}} P_{\text{GT}}(s_0 \ldots s_i | \pi, o) \underset{\mathcal{H}}{E} \frac{1}{\text{len}(H)} s_0 \ldots s_i, \pi \right) Q_{\pi}(s_i, a).
\]

Now notice that \( P(s_0 \ldots s_n | \pi) = P_{\text{GT}}(s_0 \ldots s_i | \pi, o) P(s_{i+1} \ldots s_n | \pi, s_i) C_{\pi}(o) \), where \( P(s_{i+1} \ldots s_n | \pi, s_i) \) 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 \( \pi \). 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

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

Now notice that \( P(s_0 \ldots s_n | \pi) = P_{\text{GT}}(s_0 \ldots s_i | \pi, o) P(s_{i+1} \ldots s_n | \pi, s_i) C_{\pi}(o) \), where \( P(s_{i+1} \ldots s_n | \pi, s_i) \) 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 \( \pi \). 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

\[
\sum_{i, s_0 \ldots s_n: \omega(s_i) = o} \frac{1}{n} \sum_{s_0 \ldots s_i: \omega(s_i) = o} P_{\text{GT}}(s_0 \ldots s_i | \pi, o) P(s_{i+1} \ldots s_n | \pi, s_i) Q_{\pi}(s, a) \]

Clearly, this is equal to the desired argmax term.

While the above lemma talks about CDT+GSH in general, we can now apply the lemma to CDT+GSH* to obtain Theorem \[ \text{CDT+GSH* and CDT+GT are equivalent via an analogous transformation if we restrict attention to deterministic policies. Let } \mathcal{E} \text{ be a scenario that randomizes only in the beginning and let } \pi \text{ be a deterministic policy. Then notice that each initial state } s_0 \text{ of } \mathcal{E} \text{ uniquely and deterministically determines what history will be played. Let } \text{len}_{\pi}(s_0) \text{ denote the length of that history.} \]

**Proposition 19.** Let \( \mathcal{E} \) be a scenario that randomizes only in the beginning and let \( \pi \) be a deterministic policy. Let \( \hat{\mathcal{E}} \) be the scenario that is equal to \( \mathcal{E} \), except that \( \hat{P}_0(s_0) \sim P_0(s_0) / \text{len}_{\pi}(s_0) \). Then \( \pi \) is CDT+GSH-compatible in \( \mathcal{E} \) if and only if \( \hat{\pi} \) is CDT+GT-compatible in \( \hat{\mathcal{E}} \).

**Proof.** In the following we distinguish between \( P_{\text{GT}}^{\mathcal{E}} \) and \( P_{\text{GT}}^{\hat{\mathcal{E}}} \), which are the generalized thirders’ distributions over states in \( \mathcal{E} \) and \( \hat{\mathcal{E}} \), respectively. We use \( P_{\text{GT}}^{\mathcal{E}} \) and \( P_{\text{GT}}^{\hat{\mathcal{E}}} \) and \( C_{\pi}^{\mathcal{E}} \) and \( C_{\pi}^{\hat{\mathcal{E}}} \) analogously. By Lemma \[ \pi \text{ is CDT+GSH-compatible in } \mathcal{E} \text{ if and only if for} \]

\[45\]
all $o$, $\text{supp}(\pi(\cdot \mid o))$ is a subset of

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

$$= \arg\max_{a \in A} \sum_{\text{prefix } s_0...s_i: \omega(s_i) = o} P_{GT}^\phi(s_0...s_i \mid \pi, o) / \text{len}_{\pi}(s_0) Q_{\pi}(s_i, a)$$

Now notice that

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

$$= \frac{\hat{P}_0(s_0)T(s_1 \mid \pi, s_0)...T(s_i \mid \pi, s_{i-1})}{C_{\pi}^\phi(o)}$$

$$= \frac{P_0(s_0)T(s_1 \mid \pi, s_0)...T(s_i \mid \pi, s_{i-1}) / \text{len}_{\pi}(s_0)}{C_{\pi}^\phi(o) \sum_{o'} P_0(s'_0) / \text{len}_{\pi}(s'_0)}$$

$$= \frac{P^\phi(s_0...s_i \mid \pi) / \text{len}_{\pi}(s_0)}{C_{\pi}^\phi(o) \sum_{o'} P_0(s'_0) / \text{len}_{\pi}(s'_0)}$$

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

$$\arg\max_{a \in A} \sum_{\text{prefix } s_0...s_i: \omega(s_i) = o} P_{GT}^\phi(s_0...s_i \mid \pi, o) / \text{len}_{\pi}(s_0) Q_{\pi}(s_i, a)$$

$$= \arg\max_{a \in A} \sum_{\text{prefix } s_0...s_i: \omega(s_i) = o} P_{GT}^\phi(s_0...s_i \mid \pi, o) Q_{\pi}(s_i, a)$$

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

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

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

i.e., if and only if $\pi$ is CDT+GT compatible in $\mathcal{E}$. 

\[\square\]

**Theorem 11.** Let $\mathcal{E}$ be a scenario that randomizes only in the beginning and where history length is choice-independent. Let $\hat{\mathcal{E}}$ be the scenario that is equal to $\mathcal{E}$, except that $\hat{P}_0(s_0) \sim P_0(s_0) / \text{len}(s_0)$. Then any (potentially mixed) policy is CDT+GSH-compatible in $\mathcal{E}$ if and only if it is CDT+GT-compatible in $\hat{\mathcal{E}}$.

**Proof.** This is proved in exactly the same way as Proposition[19] \[\square\]

**Corollary 12.** Let $\mathcal{E}$ 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 $\mathcal{E}$.

46
Proof. Consider the ex ante optimal policy \( \pi^* \) for \( \mathcal{E} \) as defined in Theorem 11. By Corollary 5, \( \pi^* \) is CDT+GT compatible in \( \hat{\mathcal{E}} \). By Theorem 11, \( \pi^* \) is thus CDT+GSH-compatible in \( \mathcal{E} \). It is easy to see that \( \pi^* \) cannot be a Dutch Book. \qed

E.3 Why CDT+GSH needs to view policy randomization as pre-determined

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 \( \pi \), we imagine that all results of \( \pi \)’s randomization are determined in the very beginning of the scenario and that the agent (by playing some action \( a_\pi \)) 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.

Proposition 13. There is a scenario that randomizes only in the beginning and in which all CDT+GSH-compatible policies are Dutch books.

Example 11. The scenario proceeds in three parts.

1. At the very beginning, the agent is offered to end the scenario for a price of \( \varepsilon \).
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 13.

We now argue that Example 11 proves Proposition 13. 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.\footnote{Note that the \textit{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 $1/2$.}

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 $\pi$ 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 $\pi$ does in other observations, CDT+GSH compatibility then requires that the agent randomizes uniformly in $o_1$. Consequently, $s_{3, a}$ 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 $o_1$ that it is in $s_{3, a}$ with overwhelming probability, specifically (using the fact that $\pi$ randomizes uniformly in $o_1$) $P_{GSH}(s_{3, a} \mid \pi, o_3) = 26/27$. Thus, to be CDT+GSH compatible, $\pi$ has to accept upon observing $o_3$. Now in $s_0$, the expected 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 $s_0$.

E.4 How CDT+GSH* solves Example\footnote{Note that the \textit{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 $1/2$.}

We now show how CDT+GSH* avoids the Dutch book of Example\footnote{Note that the \textit{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 $1/2$.} Specifically, we show that the policy $\pi$ 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 $o_1$. In fact, while this policy is not a Dutch book policy, its \textit{ex ante} expected utility is actually worse than the Dutch book policy of paying the price of $\varepsilon$ in $s_0$.

First, what does the modified scenario look like? First, we add an action $a_\pi$ that corresponds to following the policy $\pi$ described above in any given situation. Thus, in $o_0$, $a_\pi$ is equivalent to rejecting the offer, and in $o_3$, $a_\pi$ is equivalent to accepting the offer. Upon observing $o_2$, i.e., when playing the coordination stage of the game, $\pi$ 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 $a_\pi$. For example, if the initial state $s_{0, r, a}$ is selected, then following choosing $a_\pi$ 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 $\pi$ 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 $a_\pi$. For instance, the state $s_{2, r, a}$ indicates that the agent has played $r$ upon her first observation of $c$, and that playing $a_\pi$ 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 $a_\pi$ 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\footnote{Note that the \textit{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 $1/2$.}.

It is easy to verify that in this new model, the policy of always playing $a_\pi$ 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 $a_\pi$. In particular, conditional on $o_1$, GSH assigns most probability to the states $s_1/a/r$ and $s_1/r,a$. 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 $a_\pi$ 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 $a_\pi$ 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 $a_\pi$, CDT+GSH prefers rejecting the offer.

E.5 Is there always a CDT+GSH* compatible policy?

In the main text, we show that CDT+GSH* can be Dutch-booked, i.e., there exists a scenario in which the only policy compatible with CDT+GSH* is a Dutch book. Here we ask another question: Do CDT+GSH*-compatible policies always exist? The answer to this question is complicated. CDT+GSH* as defined in the main text does not always have a compatible policy.

We here give a scenario in which there is 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 the 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 12. 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 15 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 $\pi$ 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 $\pi$ is not compatible with CDT+GSH*.

First, we consider GSH*’s beliefs given the policy $\pi$. Upon seeing $o_1$, GSH* should believe that it is failing to “anti-coordinate” with itself. That is, according to GSH* the
Figure 14: An alternative model of Example 11 in which the policy $\pi$ of rejecting in $s_1$, mixing uniformly in $c$ and accepting in $b$ can be followed by deterministically playing $a_\pi$.

Figure 15: A graphical illustration of Example 12 in our formalism.
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 \( \pi \) 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 \( \pi \) to be CDT+GSH* compatible, CDT cannot upon observing \( o_1 \) favor playing \( a_0 \) or \( a_1 \) over following \( \pi \). 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 \( \pi \) 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 \( \pi \) and so \( \pi \) is not CDT+GSH* compatible.

We now make this argument about CDT+GSH* formal. Let \( \pi \) 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 [15] 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 [16]. The model has a new action \( a_\pi \) which represents following \( \pi \) 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 \( a_\pi \). For example, the initial state \( s_{0,1,1} \) is the state where playing \( a_\pi \) 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 \( a_\pi \) 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 \( a_\pi \) or by directly playing \( a_0 \) or \( a_1 \)). They also encode what happens when \( a_\pi \) 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 \( a_\pi \) in \( s_{0,0,1} \)) and where playing \( a_\pi \) will result in playing \( a_1 \).

We now calculate the GSH probabilities in this new model under the assumption that the agent always plays \( a_\pi \). Leaving out the normalizing constants, the probabilities are

\[
\begin{align*}
P_{\text{GSH}}(s_{0,1,1} \mid o_1, \text{always } a_\pi) &= P_{\text{GSH}}(s_{1,1,1} \mid o_1, \text{always } a_\pi) \sim \frac{p^2}{2} \\
P_{\text{GSH}}(s_{0,0,0} \mid o_1, \text{always } a_\pi) &= P_{\text{GSH}}(s_{1,0,0} \mid o_1, \text{always } a_\pi) \sim \frac{(1-p)^2}{2} \\
P_{\text{GSH}}(s_{0,0,1} \mid o_1, \text{always } a_\pi) &= P_{\text{GSH}}(s_{1,0,1} \mid o_1, \text{always } a_\pi) \sim \frac{(1-p)p}{2+K} \\
P_{\text{GSH}}(s_{0,1,0} \mid o_1, \text{always } a_\pi) &= P_{\text{GSH}}(s_{1,1,0} \mid o_1, \text{always } a_\pi) \sim \frac{p(1-p)}{2+K}.
\end{align*}
\]
Figure 16: A alternative graphical illustration of Example [12] in our formalism. In this model, the probabilistic policy $\pi$ (of choosing $a_1$ with probability $p$ in $o_1$) can be followed by taking the action $a_\pi$ deterministically.
Normalizing the probabilities, it is easy to see that

\[ P_{GSH}(\{s_0,1,1,s_1,1,1,s_0,0,0,s_1,0,0\} \mid o_1, \text{always } a_\pi) \rightarrow 1 \text{ as } K \rightarrow \infty. \]

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 16 we can now show that the policy \( \pi \) is not CDT+GSH* compatible by showing that the deterministic policy of always playing \( a_\pi \) 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

\[
\sum_{s \in S} P_{GSH}(s \mid o_1, \text{always } a_\pi)Q_{a_\pi}(s,a_1) \sim \frac{2}{2+K} \geq \frac{1}{4} - \frac{2}{2+K}.
\]

Meanwhile, the expected utility of going along with \( a_\pi \) under omission of the same normalizing constant is

\[
\sum_{s \in S} P_{GSH}(s \mid o_1, \text{always } a_\pi)Q_{a_\pi}(s,a_\pi) \sim \frac{4}{2+K} \leq \frac{4}{2+K}.
\]

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 \( a_\pi \), 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 \( a_\pi \). Thus, no policy is CDT+GSH* compatible.

**E.6 An odd fix to ensure the existence of CDT+GSH*-compatible policies (and avoid Dutch books?)**

The above counterexample hinges on the idea that for a policy \( \pi \) to be CDT+GSH* compatible, \( a_\pi \) has to be CDT+GSH compatible in the scenario where \( a_\pi \) is a deterministic implementation of \( \pi \). We now give an alternative definition under which, we will argue, CDT+GSH*-compatible policies do always exist.

First, for every policy \( \pi \), let \( \delta_\pi \) be the environment in which the agent can choose \( a_\pi \) to follow \( \pi \) deterministically. Further, let \( EU_{GSH}^{a_\pi} \) be the GSH expected utility in \( \delta_\pi \). We could now define \( \pi \) to be CDT+GSH* compatible if for all \( o \), \( \pi \) assigns positive probability only to actions from

\[
\arg \max_{a \in A} EU_{GSH}^{a_\pi}(a_\pi,o,a).
\]

This definition solves Example 12. The policy of mixing uniformly is CDT+GSH compatible in this sense.

In fact, CDT+GSH*-compatible policies in this new sense always exist. This can be established by a standard argument (used by Nash [1950] to prove the existence of
Nash equilibria). First notice that $E_{\text{GSH}}^\pi(a, o, a)$ is continuous in $\pi$. Now consider the set-valued function $f$ that maps each $\pi$ to the set of policies $\pi'$ that for each $o$ only choose from the above argmax. From the continuity of $E_{\text{GSH}}^\pi(a, o, a)$, it follows that $f$ is a so-called Kakutani function. From Kakutani’s fixed-point theorem, it follows that there is policy $\pi$ such that $\pi \in f(\pi)$. By definition, this policy $\pi$ is CDT+GSH* compatible in the above sense.

We do not know whether this method avoids Dutch book arguments in general. In any case, we do not find this alternative version of CDT+GSH* as conceptually appealing as the original version. If the agent prefers some action $a$ over following the policy $a_\pi$, then it seems there is no sense in which it is rational to follow $a_\pi$. Therefore, we view this idea as more of a technical curiosity.

**E.7 A more formal analysis of our Dutch book against CDT+GSH* (Example 6)***

In this section, we give a more formal analysis of Example 6. We first recall the example here.

**Example 6.** 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 by the number of times she chooses $a_1$ in these situations as follows: $0 \mapsto 0, 1 \mapsto 1, 2 \mapsto -1, 3 \mapsto -\varepsilon$. Here, $\varepsilon$ is some small but positive number, e.g., $\varepsilon = \frac{1}{100}$. Afterward, if $a_1$ was chosen exactly once (for a reward of 1) the agent faces the same decision problem between $a_0$ and $a_1$ another $K$ times (for some large $K$). 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 $\pi_p$ that chooses $a_1$ with probability $p$ is

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

For $\varepsilon = \frac{1}{10}$, the *ex ante* optimal policy is to play $a_1$ with probability

$$p = \frac{1}{99} \left( 30 - \sqrt{310} \right) \approx 0.210054.$$  

By Corollary 5 this policy is also CDT+GT-compatible; and by Corollary 8 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

$$p = \frac{1}{99} \left( 30 + \sqrt{310} \right) \approx 0.806895.$$
We now show that the only CDT+GSH*-compatible strategy in Example 6 is to play \( a_1 \) with probability 1, which loses money with probability 1. First, it is easy to see 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 \( \pi \) that plays \( a_1 \) with probability \( p < 1 \), we construct deterministic scenario \( \tilde{E}_d^\pi \), in which the actions of \( \pi \) is rolled out beginning and encoded in the initial state. We then ask whether always playing \( a_\pi \) – 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 \( a_\pi \), conditional on playing \( a_\pi \) 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 \( a_\pi \) leads to playing \( a_1 \) once.
   - Playing \( a_1 \) (as opposed to \( a_\pi \)) in this state, while otherwise following \( a_\pi \), leads to \( a_1 \) being played twice.

2. States with the following property:
   - Always playing \( a_\pi \) leads to playing \( a_1 \) either 0 or 2 times.
   - Playing \( a_1 \) (as opposed to \( a_\pi \)) in this state, while otherwise following \( a_\pi \), 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 \( a_\pi \). Under the second type of state, playing \( a_1 \) substantially (by 1 or \( 1 - \varepsilon \)) increases utility relative to playing \( a_\pi \). In all other states, it makes no difference whether the agent plays \( a_1 \) or \( a_\pi \).

Finally, to see that \( a_\pi \) 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 \( \pi \). 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) \( a_1 \)s being played is \( (1 - p)^2 + 3p^2(1 - p) \), 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:

\[
\frac{3p(1 - p)^2}{(1 - p)^3 + 3p^2(1 - p)} = \frac{3p(1 - p)}{(1 - p)^2 + 3p^2} \leq \frac{3p(1 - p)}{(1 - p)^2 + p} \leq \frac{3}{2}.
\]

However, if we use GSH probabilities instead of GT probabilities, then the first type of states receive a penalty of \( \frac{3}{k+3} \) 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.
F  A de se criterion for Dutch books

The following result provides a de se criterion for whether a given policy \( \pi \) is a Dutch book.

**Proposition 20.** Let \( X \) be any one of GDH, GSH, GT. A policy \( \pi \) is a Dutch book if and only if

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

Proposition 20 follows directly from the following lemma.

**Lemma 21.** For any history \( s_0\ldots s_n \) and policy \( \pi \), the following two statements are equivalent:

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

**Proof.** By definition, a policy \( \pi \) is a Dutch book policy if and only if it yields negative reward with probability 1 and there is another policy \( \pi_0 \) that yields non-negative reward with probability 1. It is easy to see from the definitions of GDH, GSH and GT that \( P_{\text{GDH/GSH/GT}}(s_0\ldots s_n \mid o, \pi) > 0 \) implies that \( P(s_0\ldots s_n \mid \pi) > 0 \). 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 \( P(s_0\ldots s_n \mid \pi) > 0 \), then \( P_{\text{GDH/GSH/GT}}(s_0\ldots s_n \mid o, \pi) > 0 \).

While Proposition 20 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.

G  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 22.** There exists a decision scenario with two CDT+GT-, CDT+GSH- and EDT+GDH-compatible policies \( \pi_1, \pi_2 \) and observations \( o_1, o_2 \) observed with positive probability under both \( \pi_1, \pi_2 \) s.t.

\[
\text{EU}_{\text{GT}}(\pi_1, o_1) > \text{EU}_{\text{GT}}(\pi_2, o_1) \quad \text{and} \quad \text{EU}_{\text{GSH}}(\pi_1, o_1) > \text{EU}_{\text{GSH}}(\pi_2, o_1) \\
\text{and} \quad \text{EU}_{\text{GDH}}(\pi_1, o_1) > \text{EU}_{\text{GDH}}(\pi_2, o_1)
\]

but

\[
\text{EU}_{\text{GT}}(\pi_1, o_2) < \text{EU}_{\text{GT}}(\pi_2, o_2) \quad \text{and} \quad \text{EU}_{\text{GSH}}(\pi_1, o_2) < \text{EU}_{\text{GSH}}(\pi_2, o_2) \\
\text{and} \quad \text{EU}_{\text{GDH}}(\pi_1, o_2) < \text{EU}_{\text{GDH}}(\pi_2, o_2).
\]
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 17.

We will only consider the following two deterministic policies: always refrain, and always take. Clearly, always taking is \textit{ex ante} optimal and \textit{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 $\omega$, all our theories assign equal probability to $s_{H,\omega \land r}$, $s_{H,\omega \land t}$ and $s_{T,\omega \land r}$, $s_{T,\omega \land t}$. Hence, \( \text{EU}_{GT/GSH/GDH}(\text{always take}, \omega) = 3 \), while \( \text{EU}_{GT/GSH/GDH}(\text{always refrain}, \omega) = 2.5 \). 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 $s_{T,\omega \land r}$ if she always refrains and probability 1 to $s_{T,\omega \land t}$ if she always takes. Hence, \( \text{EU}_{GT/GSH/GDH}(\text{always take}, T) = 4 \), while \( \text{EU}_{GT/GSH/GDH}(\text{always refrain}, T) = 5 \).

H Extending Example 10 to cover CDT+GSH

Theorem 16. There is a scenario $\mathcal{E}$ with the following properties.

- $\mathcal{E}$ 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 $\pi$ other than $\tilde{\pi}$ and all observations $\omega$, \( \text{EU}_{GDH/GSH}(\tilde{\pi}, \omega) > \text{EU}_{GDH/GSH}(\pi, \omega) \).
Example 10. At the beginning, the scenario randomizes uniformly between three possibilities: X) The agent observes x once. Y) The agent observes y once. Z) The agent observes x once and then y once. 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 \( \frac{1}{3} \) if branch X or Y is realized and they lose \( \frac{2}{3} \) 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 9.

As promised, we now extend Example 10 to apply to CDT+GSH(*) as well. (Note again that the present scenario is relatively unproblematic for CDT+GSH (as per Theorem 11), 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 10, the set of CDT+GSH-compatible policies also consists of \( \pi_{\text{bet}}, \pi_{\text{pay}}, \) and some mixed policy \( \pi_p \). The problem is that \( \text{EU}_{GSH}(\pi_{\text{bet}}, x/y) = \frac{2}{3} \cdot 1 + \frac{1}{3} \cdot (-\frac{4}{3}) = \frac{2}{9} \). Thus, in Example 10, 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 \( -\frac{4}{3} \) 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 \( \text{EU}_{GSH}(\pi_{\text{bet}}, x/y) = \frac{2}{3} \cdot 1 + \frac{1}{3}(-\frac{8}{3}) = -\frac{2}{9} < -\frac{4\varepsilon}{3} = \frac{2}{3}(\varepsilon) + \frac{1}{3}(-2\varepsilon) = \text{EU}_{GSH}(\tilde{\pi}, x/y) \) (for small enough \( \varepsilon \)), 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 10.

Of course, Theorem 16 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 \( \pi_{\text{bet}} \) 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 10 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 \( \frac{2}{3} \cdot 3 - \frac{1}{3} \cdot 4 = 2 - \frac{4}{3} = \frac{2}{3} \), while the expected reward of the old bet is \( \frac{2}{3} \cdot 1 - \frac{1}{3} \cdot 2 = \frac{4}{9} \).