Theory and Decision
https://doi.org/10.1007/s11238-025-10080-w

Tickles, iteration, and habits

Kenny Easwaran[1] · Benjamin A. Levinstein[2] · Ted Shear[3]

Accepted: 27 July 2025

Abstract
At first pass, Evidential Decision Theory (EDT) recommends one-boxing in Newcomb's Problem and Causal Decision Theory (CDT) recommends two-boxing. However, it has been acknowledged that concrete instances of the problem have messy features complicating their analyses. Recently, a third competitor, Functional Decision Theory (FDT) has emerged recommending one-boxing in some versions and two-boxing in others. This paper explores the verdicts of these competing theories in a few variations of the problem involving iteration and habit. We argue that this motivates the re-evaluation of case intuitions in decision theory and greater consideration of formal and structural relationships between CDT and FDT.

Keywords Causal decision theory · Functional decision theory · Newcomb's problem · Habits

1 Introduction

Newcomb's Problem, introduced to the philosophical literature by Nozick (1969), raises questions about rational choice in the presence of act-state dependent evidential probabilities. While Nozick's discussion focused on the conflict between the divergent recommendations of dominance reasoning and expected utility maximization, it has since become a cornerstone in the debate between proponents of

✉ Ted Shear
shear.ted@tedshear.org

Kenny Easwaran
k.easwaran@uci.edu

Benjamin A. Levinstein
benlevin@illinois.edu

[1] Department of Logic and Philosophy of Science, University of California, Irvine, Irvine, CA, USA

[2] Department of Philosophy, University of Illinois, Urbana-Champaign, Urbana-Champaign, IL, USA

[3] Department of Philosophy, University of Colorado Boulder, Boulder, CO, USA

Published online: 19 August 2025

Evidential Decision Theory (EDT) and Causal Decision Theory (CDT). At first pass, EDT seems to recommend one-boxing, while CDT seems to recommend two-boxing. However, Eells (1981, 1982) famously resisted this treatment. He proposed the so-called "tickle defense", suggesting that EDT would recommend two-boxing, in plausible realizations of Newcomb's Problem, due to the presence of a "tickle". If such "tickles" exist in cases that intuitions track, then the two-boxer's intuition does not weigh in favor of CDT over EDT. Debate surrounding the tickle defense was long and tumultuous. (Mercifully, its details will not occupy present interest and we take no stand on whether it is successful.)

Another response has involved proposing a new decision theory, Functional Decision Theory (FDT).[1] One secondary goal of this paper is to introduce this theory to philosophers who are not familiar with it and to show that it gives verdicts that agree with EDT in some cases and with CDT in others.

However, following Eells, it has often been observed that subtle under-specified features of the causal story in various presentations of Newcomb's Problem can affect the verdicts of EDT, CDT, and FDT.[2] The primary goal for this paper is to show that just as Eells' "tickle defense" allows EDT to mimic the verdicts of CDT in some intuitive presentations of Newcomb's problem, features of iteration and habits will allow CDT to mimic the verdicts of FDT. Just as EDT recommends one-boxing in precise versions of the "medical Newcomb problem", but two-boxing in more intuitive versions, we will show that CDT recommends two-boxing in precise versions of the classic Newcomb problem, but one-boxing in more intuitive versions.

This analysis reveals that the real dispute between CDT and FDT may not lie in their divergent recommendations about specific cases, but rather in how they conceptualize the unit of rational evaluation. Where CDT evaluates individual acts, FDT evaluates decision algorithms. Our cases suggest that when we incorporate realistic features of human psychology and decision-making — particularly the way current choices shape future behavior through habit formation — CDT's act-based evaluation often yields similar results to FDT's algorithm-based approach. This insight has important implications for how we understand rational agency and the relationship between one-off decisions and broader patterns of behavior.

This challenges the common view that intuitions about Newcomb-like problems can adjudicate between these decision theories. Instead, we argue that the choice between them must rest more on theoretical considerations about the proper unit of rational evaluation and how we understand the connection between our present and future selves.

The outline of the paper is as follows. In Sect. 2 we introduce the details of EDT, CDT, and FDT. In Sect. 3, we discuss Nozick's original presentation both of the standard and "medical" Newcomb problem. Then, in Sect. 4, we consider two cases that have since been discussed in the literature, in which CDT recommends

[1] The idea behind this theory goes back several decades, but the name was first introduced by Yudkowsky and Soares (2017) and has appeared in philosophical literature in (Levinstein & Soares, 2020) A similar approach is found in the "timeless decision theory" developed in (Yudkowsky, 2010).

[2] For instance, see (Easwaran, 2021) (Levi, 1975, 1982), (Li & Zenker, 2021), (Mackie, 1977), (Maitzen & Wilson, 2003), (McKay, 2004), (Slezak, 2006), (Talbott, 1987), and (Wolpert & Benford, 2013).

one-boxing. Such cases, however, may be reasonably dismissed as distinct "pseudo-Newcomb problems",[3] because one involves backwards causation, and the other involves a choice made before the predictor has made their presence known. Both are incompatible with Nozick's original description of Newcomb's Problem. Still, this will help motivate considerations involving iterated choice (in Sect. 5), and of habit (in Sect. 6) that can be used to develop cases that are consistent with Nozick's original description.

In those sections, we show that when such features are present, CDT will recommend one-boxing, just as FDT does. That is, these considerations enable CDT to mimic the decisions of FDT, just as the "tickle" enables EDT to mimic the decisions of CDT. Depending on how widely these iterations and habits recur, there may be few cases where clear enough intuitions help decide between CDT and FDT, just as a sufficiently widespread occurrence of tickles may make it difficult for intuitions to distinguish EDT and CDT.

Instead, we think that the real dispute must come down to the substantive features that play a role in decision theory. Even though the tickle defense, if successful, would allow EDT to mimic many of the recommendations of the other two theories, we reject EDT because we believe that it is not adequately sensitive to the important role causal information may play in deliberation. But we are less clear whether idealized FDT, or CDT supplemented with the plausible roles of iteration and habits in real-world decisions, is a better fit for the theory of rational choice. We discuss this further in Sect. 7.

2 EDT, CDT, and FDT

In this section we give an overview of the the framework we use for decision theory, and explain how EDT, CDT, and FDT work. In the next section, we illustrate the differences between these theories by showing how they deal with certain precisified versions of Newcomb's problem.

In general, we take decision problems to comprise a set of acts, $\mathcal{A}$, a set of states, $\mathcal{S}$, and a specification of the outcomes $\mathcal{O}$. The set of worlds $\mathcal{W} = \mathcal{A} \times \mathcal{S}$ is the set of fine-grained, fully specified results of an action and a state, while the set of outcomes is the set of equivalence classes of worlds under the agent's indifference relation. The outcome resulting from act a in state s is $o[a, s]$. We measure the desirability of outcomes with a utility function u unique up to positive affine transformation.

All three theories agree that rational choice consists in the maximization of expected utility, but disagree about which probabilities are relevant for its calculation. EDT recommends that agents maximize *evidential* expected utility:

$$\mathrm{EU}_{\mathrm{EDT}}(a) := \sum_{s \in \mathcal{S}} \Pr(s \mid a) \cdot u(o[a, s])$$

[3] See (Joyce, 2018). Note that Levi (1982) uses the same phrase for a variant that we would classify as a "real" Newcomb Problem.

Here, for each s, the utility of $o[a, s]$ is weighed by the (evidential) conditional probability of s given a. Since Newcomb's Problem involves a reliable predictor who determines that the money is present/absent just in case they predict that you will one-/two-box, EDT recommends one-boxing, since choosing one box provides strong evidence that money is present.

By contrast, CDT does not regard all evidential correlations as decision relevant. Instead, agents should consider how much performing a causally promotes/inhibits s. This is often expressed using a causal conditional "→", where $\Pr(a \rightarrow s)$ is the agent's subjective probability that s would obtain were the agent to perform a. (In the notation of Pearl (2000), this is $\Pr(a \mid \mathrm{do}(s))$). So, $\Pr(a \rightarrow s) > \Pr(s)$ implies that the agent thinks that performing a would causally promote s, $\Pr(a \rightarrow s) < \Pr(s)$ implies that they think that performing a would causally inhibit s, and $\Pr(a \rightarrow s) = \Pr(s)$ implies that they regard a as causally independent of s. Thus, according to CDT, agents should maximize *causal* expected utility:

$$\mathrm{EU}_{\mathrm{CDT}}(a) := \sum_{s \in \mathcal{S}} \Pr(a \rightarrow s) \cdot u(o[a, s])$$

In standard presentations of Newcomb's Problem, where there is merely evidential correlation between acts and states, this results in the recommendation that agents two-box. This is because the choice is causally independent of the state and either way two-boxing results in an outcome with a greater payout.

While EDT and CDT disagree in their determinations of expected utility, they agree that *acts* are the fundamental units of rational evaluation. For both, the rationality of an agent depends solely on whether each of their time slices maximizes expected utility. FDT offers a different picture. Rather than evaluating the agent as a sequence of time slices, FDT evaluates agents as decision algorithms (with multiple possible physical instantiations).[4]

To see how FDT calculates expected utility in a choice scenario, we need to express how the world would change if the agent (i.e., the decision algorithm) were to result in a different verdict when fed the same decision problem. Let `myAlg(·)` be the function from decision problems to recommended acts that the agent (qua decision algorithm) implements, and let `problem` be the decision problem. According to FDT, agents should maximize *functional* expected utility:

$$\mathrm{EU}_{\mathrm{FDT}}(a) := \sum_{s \in \mathcal{S}} \Pr((\texttt{myAlg}(\texttt{problem}) = a) \rightarrow s) \cdot u(o[a, s])$$

[4] There is a philosophically rich discussion to be had regarding whether commitment to EDT/CDT (or FDT) requires accepting (or rejecting) so-called "time-slice rationality" developed by Hedden (2015a, 2015b). For the purposes of this paper, however, we simply draw the reader's attention to this connection and leave its consideration for future work. Further discussion of time-slice rationality is provided in (Moss, 2015) and (Sullivan, 2018).

That is, where CDT considers how likely a state would be if her *act* were a, FDT considers how likely a state would be if her *decision algorithm* recommended a.[5] Note that the supposition that $\texttt{myAlg(problem)} = a$ also involves the agent doing a in other instances of the same decision algorithm, and involves predictors of the agent all predicting that she does a.

3 Newcomb's problem and the medical Newcomb problem

We now describe the general structure of what we take to be a "Newcomb-type problem", and illustrate the differences between EDT, CDT, and FDT by showing their different recommendations in certain precise versions of such a problem. While some decision theorists have tried to use intuitions about these precise versions of the problem to settle the debate as to which of these theories is normatively correct, we want to be more careful. In this section, we just use these problems to show that the theories are in fact distinct, and give some examples of the reasoning these theories use to recommend their different verdicts. In subsequent sections, we will explore a variety of other Newcomb-type problems, to show that the verdicts of these three theories actually agree more often than one might at first think on the basis of these precisifications.

Table 1 represents the typical payoff matrix of a Newcomb-type problem. There are two available acts, $\mathcal{A} = \{\texttt{act1}, \texttt{act2}\}$, two possible states, $\mathcal{S} = \{\texttt{state1}, \texttt{state2}\}$, and outcomes with the following schematic relations.

The important feature that is shared by all Newcomb-type problems, beyond this schematic payoff matrix is that `act1` is correlated with `state1`, and `act2` is correlated with `state2`, even though the acts are not in any way causally upstream from the states. Different versions of the problem differ in the specification of the acts, states, and outcomes, and importantly in the causal structure that sets up the essential correlation between acts and states. As we will show later, these details will matter to the recommendations of some decision theories, as well as to our intuitions.

Before offering precise versions, we quote Nozick's (1969, p. 114) setup of the problem:

> Suppose a being in whose power to predict your choices you have enormous confidence. (One might tell a science-fiction story about a being from another planet, with an advanced technology and science, who you know to be friendly, etc.) You know that this being has often correctly predicted your choices in the past (and has never, so far as you know, made an incorrect prediction about

[5] Yudkowsky and Soares (2017) observe an interpretive difficulty based on the fact that the output of an algorithm is logically determined, so that the counterfactual conditional here has a counter-logical antecedent. But Levinstein and Soares (2020) suggest we can understand this analogously to the counter-possibles that occur in compatibilist accounts of free will, and Easwaran (2021) suggests that we can think of it on the model of an "intervention" at the level of the (multiply-realized) decision algorithm rather than at the level of the act. Kocurek (2024) argues that these are usually not actually counter-possibles, and that when they are, the problems are solvable.

Table 1 Standard decision matrix for Newcomb-type problem

	`state1`	`state2`
`act1`	Very good	Bad
`act2`	Best	Slightly good

> your choices), and furthermore you know that this being has often correctly predicted the choices of other people, many of whom are similar to you, in the particular situation to be described below. One might tell a longer story, but all this leads you to believe that almost certainly this being's prediction about your choice in the situation to be discussed will be correct.

Nozick then describes a choice. There are two boxes in front of you: The first contains either \$1,000,000 or nothing, and the second \$1,000. You can either take only the first box (`act1`) or take both (`act2`). You are told that the amount in the first box was determined on the basis of what the being predicted about your choice. If they predicted you would only take the first box, then it contains \$1,000,000 (`state1`), but if the being predicted that you would take both, then it contains nothing (`state2`). The prediction has already been made and, thus, the value in the box is already determined. As Nozick (1969, p. 117) observes, "To almost everyone it is perfectly clear and obvious what should be done. The difficulty is that these people seem to divide almost evenly on the problem, with large numbers thinking that the opposing half is just being silly."[6]

3.1 Standard precisification

In the relatively standard precisification that follows, both EDT and FDT recommend `act1`, while CDT recommends `act2`.

Suppose Alex is an AI robot with open source code.[7] Omega is a predictor who will run the algorithm Alex instantiates, and predict what she will do when she shows up. He will put \$1,000,000 in an opaque box if he predicts Alex will take just that box, and \$0 if he predicts she will take that box plus a second box containing \$1,000. Then Alex will show up, and Omega will tell her these facts, and let her choose whether to take just the opaque box, or the opaque box plus the second box. After Alex makes her choice and gets her reward, Omega will take away anything left over, they will never interact again, and no one else will observe what happened.

If Alex follows CDT, she will reason as follows:

> If I were to take just the one box, I'd get whatever is in there. But if I were to take both, I'd get \$1,000 more. After all, the first box has already been filled

[6] Nozick's observation is (mostly) vindicated by the two recent PhilPapers surveys as outlined in (Bourget & Chalmers, 2014, 2023) In the earlier survey, 31.4% of respondents were two-boxers, 21.3% were one-boxers, while 47.4% indicated "other"; in the more recent, 39% were two-boxers, 31.2% were one-boxers, and 30.2% indicated "other".

[7] Or, in keeping with more modern machine learning paradigms, she is a neural net whose activations and architecture are public.

> or not, so what is in there doesn't depend on what I do. Since nothing else depends on what I do, I should take both boxes.

CDT says to two-box because the only thing affected by her act on this occasion is whether or not she gets the extra $1,000.

However, if Alex follows FDT, she will reason as follows:

> If my algorithm were to output `act1` (one-boxing), then I'd get $1,000,000. After all, Omega is running my algorithm too (as a subroutine). So, the outputs will be the same regardless of whether I run it or Omega does. On the other hand, if it were to output `act2` (two-boxing) here, then I'd get $1,000. I like $1,000,000 more, so I should output `act1`.

In this case, FDT says to one-box because if Alex's algorithm were to output `act1` when it is run on her hardware, then it also would output `act1` when run on Omega's, and this would bring about the presence of the million dollars.

Note that this is somewhat different from the reasoning Alex would do if she follows EDT:

> If I take just the one box, that is strong evidence that Omega in fact already predicted that I would do this, which would mean that the million dollars is there. But if I take both boxes, that is strong evidence that Omega has in fact already predicted that I would do that, which would mean that the million dollars is not there. Thus, one-boxing is evidence that something good has already happened, so I should do that.

EDT and FDT both say to one-box, but EDT does because one-boxing is *evidence* of a good outcome, while FDT does because being the kind of algorithm that outputs `one-box` *brings about* a good outcome.

Note that while CDT and FDT yield different recommendations in this case, they both rely on counterfactual conditionals. The different recommendation comes about because each involves a different type of counterfactual dependence. FDT's counterfactuals ask, "What if my decision procedure were to output one thing rather than another?" CDT's counterfactuals ask "how would the world be different if I were to move my limbs one way rather than another?" Where CDT thinks of the unit of analysis as the time slice of the agent, FDT thinks of the unit of analysis as the agent herself, fully located at every time and place where her decision algorithm is instantiated.[8] In essence, FDT's counterfactuals treat agents as repeatable software rather than one-off hardware, while CDT's treat agents as physical machines.

There is a certain deep similarity between FDT and a precursor view developed outside of philosophy by Douglas Hofstadter (1985, p. 737–780). He introduces a concept of "superrationality" that treats the unit of rational evaluation as the theory of rationality itself. If an *individual* were to engage in the act of two-boxing, then *she* would receive an additional thousand dollars. But if the *theory of rationality* were to

[8] There may be an interesting connection to the metaphysical debate between endurantists and perdurantists about personal identity — see (Sider, 2001) for exposition of that debate. However, our concern is just with the theory of rationality, and we leave the connections to metaphysics for future research.

recommend one-boxing, then *rational individuals* would be predicted to one-box, and would *all* get an additional million dollars. Thus, Hofstadter argues that the theory of rationality ought to recommend one-boxing.[9] Rather than disputing the distinction between causal and evidential correlations between acts and states, this alternative (like FDT) disputes the claim that the unit of rational evaluation is the act.

3.2 The medical Newcomb problem

Nozick himself endorsed two-boxing in his original incompletely-specified case. In order to motivate this, he presented another case, which has since come to be known as the “medical Newcomb problem”. Lewis (1981, p. 9–10) describes one version as follows:

> Suppose you like eating eggs, or smoking, or loafing when you might go out and run. You are convinced, contrary to popular belief, that these pleasures will do you no harm at all. (Whether you are right about this is irrelevant.) But also you think you might have some dread medical condition: a lesion of an artery, or nascent cancer, or a weak heart. If you have it, there’s nothing you can do about it now and it will probably do you a lot of harm eventually. In its earlier stages, this condition is hard to detect. But you are convinced that it has some tendency, perhaps slight, to cause you to eat eggs, smoke, or loaf. So if you find yourself indulging, that is at least some evidence that you have the condition and are in for big trouble. But is that any reason not to indulge in harmless pleasures?

Many discussions of medical Newcomb cases — e.g. that of Skyrms (1980, p. 129) — suggest that inspiration is drawn from the hypothesis proposed by R.A. Fisher (1959) that the correlation between smoking and cancer is best explained by a common genetic cause.[10]

Let us precisify the case as follows, which we will call `medical`.

> Scientists have discovered a strange gene that comes in two variants. The first variant (`state1`) both deterministically prevents a person from ever developing lung cancer, and deterministically causes a person to not smoke (`act1`). In particular, it causes this by causing them to decide according to a decision algorithm `myAlg`(·) such that `myAlg(medical)` = `act1`. The second variant (`state2`) both deterministically causes a person to develop lung cancer later in life, and deterministically causes a person to smoke (`act2`). In particular, it causes this by causing them to decide according to a decision algorithm `myAlg`(·) such that `myAlg(medical)` = `act2`. Regardless of whether or not a person has the gene, everyone likes smoking, but would really rather not have cancer, so the outcomes have the utilities of the standard table. Because of current legal and ethical guidelines, labs will not do genetic testing to determine which

[9] A similar analysis by a philosopher is given later by Gauthier (1994), and is disputed by Velleman (1997) and Barnes (1997).

[10] For simplicity, we adopt the accompanying terminology when discussing medical Newcomb cases.

> variant a person has, and the only evidence a person can gather about which variant they have is by noting whether they choose to smoke or not.

If a person follows EDT, they will reason as follows:

> If I don't smoke, that is strong evidence that I have the first variant, which would mean that I'm never going to develop lung cancer. But if I do smoke, that is strong evidence that I have the second variant, which would mean that I will develop lung cancer. Although I would like smoking, it would be evidence of something really bad, while not smoking is evidence of something good, so I should not smoke.

Just as in the original Newcomb problem, EDT recommends `act1` because it is evidence of a good outcome.

If a person follows CDT, they will reason as follows:

> If I don't smoke, my only outcome is either cancer or no cancer, based on which variant of the gene I carry. But regardless of which variant I carry, I'd be happier if I smoked. After all, either I already have the gene or not, so it doesn't depend on what I do. Therefore, I should smoke.

CDT's recommendation of `act2` will also remain unchanged from the original Newcomb problem because the only thing affected by their act is the small sweetener.

However, if a person follows FDT, they will reason as follows:

> If my algorithm were to output `act1` (not smoking), then my only outcome is either cancer or no cancer, based on which variant of the gene I carry. But regardless which variant I carry, if my algorithm were to output `act2` (smoking), I'd be happier. After all, my algorithm doesn't affect my genetics (though it is affected by it). Therefore, my algorithm should output `act2`.

Unlike the original Newcomb problem, FDT recommends `act2`. In the original problem, the output of their algorithm was causally upstream from the predictor's prediction, and thus the state, while in the medical version, their algorithm is causally downstream from the factor that determines the state.

We don't know the details of how the gene works on the decision algorithm of the agent. It could be that the first variant causes people to use EDT while the second variant causes people to use CDT, or it could be that the first variant causes people to use EDT while the second variant causes people to use FDT. It could be that there are multiple other decision algorithms beyond the ones that we have described, some of which output `act1` and others of which output `act2`, and the first variant causes people to adopt a random element of the first set while the second variant causes people to adopt a random element of the second set. There is undoubtedly something odd about the fact that all of these decision algorithms operate by having the agent consider what might happen if they turn out to have genetics that caused them not to use this very decision algorithm, but some of this oddness occurs whenever decision theory is applied in a deterministic world.

Nozick thought that the original problem and the medical version were structurally similar enough (same payoff table, both with an act-state correlation in which the acts are not causally upstream of the states) that any decision theory should make the same recommendations in both. Since `act2` seems more intuitive in the medical Newcomb problem, he thought this was suggestive that `act2` was right in the first version. But now we have seen that FDT recommends `act1` in the first version, and `act2` in the medical version. So, FDT is genuinely a third theory whose recommendations sometimes differ from both EDT and CDT.[11] So more careful argument about the structure of the theory is needed.

3.3 The "tickle defense"

Despite agreeing that two-boxing is rational in the medical Newcomb problem, some philosophers argued[12] that there is theoretical reason to prefer EDT over CDT — in particular, EDT obviates the need to distinguish causal correlations from merely evidential ones. Furthermore, Eells (1981, 1982) argued that EDT can recover the judgment CDT gives in the medical Newcomb problem. He argued that although there is no available evidence about which variant one has in a highly idealized presentation, like `medical`, in more realistic presentations, evidence is available.

We don't know how it is that genetics could possibly predispose people to deciding according to one decision algorithm or another. But Eells argued that there would likely be some introspectible manifestation. Thus, he suggested that instead of `medical`, a more plausible version of the problem, which we might call `tickle`, would have the following last sentence:

> Although legal and ethical guidelines prevent labs from doing genetic testing, we have noticed that people with the second variant can detect a "tickle in the throat" that is not present for people with the first variant.

In this version of the problem, CDT and FDT both still recommend smoking (`act2`), because they care only about the causal structure, and not the evidential structure. But a person who follows EDT would reason as follows:

> First, let me introspect as to whether or not I have the "tickle in the throat". I find that I [do/don't]. This means that I am in [`state2`/`state1`]. Thus, I [will/won't] get lung cancer. Regardless, I will be a little happier if I smoke, so I should smoke.

Thus, although EDT recommended `act1` in `medical`, it recommends `act2` in `tickle`.

The general theoretical move of arguing that act-state dependence must obtain in virtue of some introspectible feature has since been called "the tickle defense". To the extent their recommendations agree[13], arguments about whether EDT or CDT is

---

[11] Other cases that help illustrate the ways that FDT differs from both EDT and CDT are discussed by Isaacs and Levinstein (2024) and Levinstein and Soares (2020).

[12] See Ahmed (2014).

[13] It should be said that disagreement remains as to whether the tickle defense can be successfully applied in all cases. See (Huttegger, 2023) and (Oesterheld, 2022) for reasons to think that it may not.

the better decision theory must then turn on theoretical considerations. In the early 1980s, it may have been plausible to lean towards EDT on these grounds due to the lack of sophisticated understanding of causal concepts. But with more modern analyses — e.g. (Pearl, 2000) or (Spirtes et al., 2000) — resisting CDT due to discomfort with causal concepts is no longer compelling.

It is important to note that, since the theory of rationality doesn't affect one's genetics, superrationality and FDT behave just like CDT with respect to the medical Newcomb problems, regardless of how they behave in the original Newcomb problem. We find the recommendations of FDT to be intuitive — one-boxing (`act1`) in ordinary Newcomb problems, and smoking (`act2`) in medical Newcomb problems. But as we will suggest, these intuitions are not sufficient to decide between FDT and CDT. Rather, the important question is, as argued by Velleman (1997), that of whether the proper unit of rational evaluation is the act, or the theory of rationality itself.

4 Two pseudo-Newcomb problems

As we will now see, CDT does not lead to two-boxing in every Newcomb-like situation. In the original case, it is important that (1) there is no backwards causation, and (2) the decision under evaluation occurs after the prediction. If either condition fails, we have a different type of problem in which causal decision theorists are typically content to endorse one-boxing.

4.1 Backwards causation

Consider a variant where, although the prediction is *temporally prior* to your decision, the being's prediction is *causally downstream* from your decision. This may be because the predictor has a crystal ball directly accessing the future (Hall, 1994), or because of some consistent time travel story (Egan, 2007), or because you and the predictor lie on some closed timelike curve in a Gödel spacetime (Gödel, 1949). Regardless of which, CDT recommends one-boxing because your choice straightforwardly *causes* the box to contain $1,000,000.

Although not ruled out in his initial description (he mentions "a science-fiction story" after all), Nozick (1969, p. 134) clarifies:

> [S]uppose we establish or take as given that there is no backwards causality, that what you actually decide to do does not affect what he did in the past, that what you actually decide to do is not part of the explanation of why he made the prediction he made. So let us agree that the predictor works as follows: He observes you sometime before you are faced with the choice, examines you with complicated apparatus, etc., and then uses his theory to predict on the basis of this state you were in, what choice you would make later when faced with the choice.

The variants we discuss later will respect the prohibition on backwards causation, though the cases in Sect. 5 allow the "complicated apparatus" to include your past selves, or duplicates, whose decision *does* cause the prediction.

4.2 Binding

CDT will also recommend one-boxing when the decision under evaluation is a binding choice *before* the prediction. Consider the following:

> You have never met a being with the predictive powers mentioned above. However, you have heard of other people encountering such a situation and want to be prepared. A neuroscientist offers you a pill that will rewire your neurons leaving you mostly unchanged, but with one small effect: If you do ever encounter such a situation, your autonomous reasoning abilities are bypassed, causing you to one-box. Moreover, this rewiring is done in an obvious way so that the being will find it easy to predict that you will one-box.

CDT recommends taking the pill since this will affect any future predictions of your behavior and, thus, help bring it about that the box contains $1,000,000. Accordingly, CDT recommends one-boxing in this case.[14]

However, in this version of the story, the decision under evaluation isn't whether to one- or two-box — it's about taking the pill. At its core, Newcomb's Problem is about what to choose *after* the being has already made their prediction. If you've already taken the pill, this question turns out to be uninteresting since taking the pill preempts this decision. We might say that CDT still *recommends* two-boxing, but it is an idle recommendation, since you no longer have the ability to decide. Or we might instead say that since your abilities are now limited, CDT now recommends you take the best available option, which happens to be one-boxing, as it is the *only* available option. Regardless, the decision that *causes* the one-boxing in this pseudo-Newcomb problem is a decision that occurs before the predictor makes the prediction.

The variants of the problem we discuss later will involve decisions made after the start of the relevant problem, rather than decisions made in advance, though the versions in Sect. 6 will have similar causal impacts on your future behavior.

4.3 Preview

The cases that will be the focus of our attention will not be pseudo-Newcomb problems as discussed above: they all avoid appealing to backwards causation, and involve decisions made from inside the problem rather than in advance.

[14] Similar insights are found elsewhere in the literature: An early statement that CDT would recommend such self-modification is given by Skyrms (1982, p. 707); Arntzenius et al. (2004) provides a discussion of the rational benefits of having the ability to self-bind; Meacham (2010) considers the consequences of binding for various debates in decision theory; Carr (2017) considers a similar case in discussion of epistemic utility theory; and Oesterheld and Conitzer (2021) discuss binding in connection with diachronic Dutch books.

Nonetheless, some structure from these other modifications is preserved. In all of the cases we discuss, there will be some sort of repetition, duplication, or iteration. Although CDT only cares about the causal impacts of the one act, uncertainty about *which* act is in question, or causal connections between the acts, will lead it to make similar recommendations to FDT, which evaluates the multiply-realized decision algorithm.

Decision theorists, following Savage (1954), sometimes distinguish between "grand world" and "small world" decision problems. In a "grand world" decision problem, states and acts are so finely differentiated that, together, a state and an act suffice to determine everything that the agent ever cares about. This is sometimes taken to mean that in a grand world decision problem, there is only one decision to be made, which then determines the agent's strategy at all future choice points — if any future choice points are left open, then the current act and state haven't determined the overall goodness that results, since future choices will play some role. This leaves little room for Newcomb problems in the grand world setting. The decision in Newcomb's Problem is one that is made after the prediction has been made, which is likely after other decisions, and so it was itself once one of the future choice points. Furthermore, the outcome of Newcomb's Problem is usually described in terms of the monetary payout received, which thus leaves open the details of future decisions the agent may make. Newcomb's Problem thus usually occurs in a small world. We will argue that if there is uncertainty about which small world context one is in, or causal connections between these contexts, these can often lead CDT to make recommendations similar to those given by FDT.

5 Causal benefits without backwards causation

Our first set of cases build on the backwards causation pseudo-Newcomb problem. Here, one's decision is effective at bringing about good outcomes by means of influencing some future prediction, rather than by means of influencing the current payoff, or one's future choices.

5.1 Iterated Newcomb problems

Nozick says little about how the being is so accurate, but he does say that they have previously made many correct and no incorrect predictions about both you and others. Nozick avoids saying whether the being has ever put *you* in this same situation before. However, we can imagine a variant where they have, such that CDT clearly recommends one-boxing.

Consider a case where the being offers you (and many others) the choice of one or two boxes every day. Instead of $1,000,000 or nothing in the first box, and $1000 in the second box, let us say that it is either $3 or nothing in the first box, and $1 in the second box. Let us also relax the assumption that the being has never been wrong, and instead just say that the being has *rarely* been wrong. Finally, let us assume the being does nothing at all sophisticated in forming their predictions.

> Each day, the being makes their prediction about how someone will behave based on what they did on the previous the day. Since most people are fairly stable in their decision-making, this prediction is fairly accurate most of the time.

Now, your choice on any given day has *two* causal influences on outcomes. While it *doesn't* causally influence what is in the first box *that* day, only influencing whether you get the additional $1, it *does* causally influence the *next* day's prediction, and thus what is in the first box the next day. While one-boxing deprives you of $1 today, it grants you $3 tomorrow. So long as you don't substantially discount future earnings, CDT now recommends one-boxing each day.

Even without building in knowledge of such iteration, Nozick's original specification does not include much about what to expect after your choice. We are presumably meant to assume that there will *certainly* be *no* further interactions with the being (which actually sounds kind of sad, given knowledge that it is friendly and the mention of your extended past interactions). But if we instead think that it is *possible* that the being will offer the choice again on a future occasion and we think that the choice this time *might* play a role in the being's predictions of future behavior, then CDT can recommend one-boxing.[15]

Described in this way, the case is nearly identical to an iterated Prisoner's Dilemma.[16] We have specified that, if the being predicts you will one-box, you get $3 from the first box (regardless of your choice), and $0 if they predict otherwise; and you also get an additional $1 if you in fact two-box. In a standard Prisoner's Dilemma, each player gets 3 units of payoff if the other person cooperates and nothing otherwise, and gets 1 additional unit of payoff if they themselves defect and nothing otherwise. Your *act of one-boxing* and the being's *prediction of one-boxing* can be regarded as playing `cooperate` and your *act of two-boxing* and the being's *prediction of two-boxing* as playing `defect`. The being's prediction that you will do the same as what you did last time just means that the being will play `cooperate` if you previously played `cooperate`, and will play `defect` if you previously played `defect`. That is, they are playing the Tit-for-Tat strategy, which has empirically done well in simulated tournaments of iterated Prisoner's Dilemma (Axelrod, 1980).

Questions about the rational strategies in one-off and iterated Prisoner's Dilemmas have been subject of much debate. One standard response is to say that `defect` is the only rational option when one is confident that the current round has no effect on the opponent's choice in the next round. On such an account, if it is common

[15] In particular, let I be your credence that there will be at least one more iteration of the offer, and let x be the difference between your credence that the predictor will predict you to one-box next time given that you one-boxed this time, and your credence that the predictor will predict you to one-box next time given that you two-boxed this time. Assume that the current decision will have only neutral causal impacts on expected values for anything other than potential future Newcomb choices. With payoffs of $3 and $1, CDT will recommend one-boxing on this occasion as long as $I \cdot x > 1/3$, but with the original payoffs of $1,000,000 and $1000, CDT will recommend one-boxing on this occasion as long as $I \cdot x > 1/1000$.

[16] The connections between Newcomb's Problem and the Prisoner's Dilemma have been extensively discussed in the literature, e.g., by Brams (1975), Lewis (1979), Sorensen (1985), Hurley (1991), and the volume edited by Campbell and Sowden (1985).

knowledge that both players are rational and common knowledge what finite number of rounds will be played, a standard backwards induction argument shows that rationality requires both players to `defect` on every round (Kuhn, 2019). But if one player is sufficiently confident that the other is playing Tit-for-Tat in every round, regardless of whether that strategy is rational, then CDT says that `cooperate` is rational in every round that is not likely to be terminal.

Nozick's setup of Newcomb's Problem is compatible with this stipulation and, thus, compatible with CDT recommending one-boxing. Although your act doesn't causally impact the current prediction, we can't rule out the possibility of another round that it *will* causally impact. Just as a tickle would make EDT behave like CDT in the medical Newcomb problem, the possibility of iteration would make CDT behave like FDT in the original Newcomb problem.

5.2 The Sleeping Beauty Newcomb problem

To see another way of grounding the reliability of the being's predictions, we consider another case, adapted from the Sleeping Beauty problem. The details of the original Sleeping Beauty problem are not essential here, since our case borrows only some of its apparatus.[17]

> It's Sunday and you've arrived at a psychology lab where you'll spend a few days participating in a study that will conclude on Wednesday. You haven't been told many details, though you have been credibly assured that it has been approved by the institutional review board (IRB) and won't lead to any harm or embarrassment. They say you'll sleep Sunday, wake Monday, and have some (safe and non-embarrassing) interactions that they won't tell you about yet. After you fall asleep on Monday night, they will administer a drug erasing memories of the past twenty-four hours and you'll awake on Tuesday with your body and brain reset to precisely their state from Monday morning. However, at this point, they will inform you that it is Tuesday, and will place two boxes before you. You are given the familiar choice between the first box (containing either \$1,000,000 or \$0) or the first box together with the second, which you know to contain \$1000. The contents of the first box depend on a prediction that they have already made. If they predicted you will one-box, then the first box contains the \$1,000,000. If they predicted that you will two-box, then it contains nothing. After you choose, they'll reveal what's in the boxes. You'll sleep in the lab one more night and will depart on Wednesday with your earnings.

So far, this sounds like the standard Newcomb Problem and that CDT will recommend two-boxing. However, there is this mysterious gap regarding the events of Monday, which you won't remember, and haven't been described to you. The

[17] For the description of the original problem, see (Elga, 2000). For other work analyzing similar decision theories in this and related contexts, see (Piccione & Rubinstein, 1997), (Aumann et al., 1997), (Briggs, 2010), (Conitzer, 2015), (Schwarz, 2015), and (Korzukhin, 2021).

experimenters also haven't explained how they've managed high accuracy in predicting choices of past participants (which you have, in fact, verified).

Before going to bed on Sunday (early enough to be safe from the memory-eraser, which only erases Monday memories), you get an encrypted message from a whistle-blower informing you of some deception. On Mondays, they've been lying to people by telling them it's Tuesday and offering them the same choice! In fact, they've been running both Mondays and Tuesdays off the same script. The predictions they have made for these Monday interactions have often been incorrect. But Mondays, after they've revealed the contents of the boxes and told the participants that their earnings are safe, they've removed these earnings, and made their prediction for Tuesday based on what the person chose on Monday. Thanks to the tightly-scripted interaction and the reset of the brain and body states, participants have behaved very consistently, so the lab's Tuesday predictions have always been accurate. Also, thanks to the IRB regulations, the Tuesday discussion has in fact turned out to be accurate and participants have in fact received their Tuesday earnings, even if their Monday earnings were a lie. The whistle-blower says that you should just keep quiet, do your part, collect your earnings, and wait until Wednesday to bring in the authorities.

Now, you find yourself awake in the lab, being told that it is Tuesday, and offered the choice. But, knowing that Mondays and Tuesdays are run from the same script (which is honest on Tuesday but deceptive on Monday), you wonder how to maximize your earnings. They've said that two-boxing carries a net benefit of $1000, but you know that this is only true if it's Tuesday. If it's Monday, then they're lying, and *one*-boxing will actually increase your eventual earnings by $1,000,000.

In the iterated Newcomb problem a decision to two-box definitely causes you both to get the small prize now and to lose the big prize on the next run, while a decision to one-box definitely causes you both to lose the small prize now and to gain the big prize on the next run. Thus, in the iterated problem it is clear that one-boxing causes a greater net improvement in your fortune than two-boxing.

In this version, however, only *one* of these two effects will result from your current decision, but you don't know which. In the Sleeping Beauty problem that inspired this case, Elga (and others) argue that if you're sure you'll wake on both Monday and Tuesday, then you should have credence 0.5 about whether it is Monday or Tuesday, and thus about whether one-boxing will cause you to gain the large prize or cause you to lose the small prize. But the recommendation is still the same — as long as the large prize is larger than the small prize, CDT will recommend one-boxing. In fact, FDT makes the same recommendation, since the repeatable algorithm is realized on both dates and causes *both* consequences.

The important difference between the CDT and FDT evaluation of this case is that the two centers get equal weight in the FDT evaluation, while their weight in the CDT evaluation depends on the credence that one is in either center. This may be relevant for a person who went through the experiment without the information from the whistle-blower. It might have *seemed* like CDT must have recommended two-boxing for them, because only the Tuesday center was salient. But, is this right? They were told that Monday was going to be a mystery, and the accuracy of the predictions was unexplained. Should they have considered the possibility that the whistle-blower revealed? If the probability of this structure was greater than 1/500, and they were conditionally 50/50 about whether it was Monday or Tuesday, then there would be greater than 1/1000 chance of one-boxing causing a gain of $1,000,000 without causing any corresponding loss. If the possibility that the whistle-blower revealed was not much less likely than alternative causal structures for the prediction, it may well be that CDT would recommend one-boxing in the original unspecified version of the case, due to this uncertainty. Although CDT only cares about the causal results of *this* act, uncertainty about *which* act this is can lead it to make the same recommendation as FDT, which cares about the causal results of *all* runs of the decision algorithm. The difference is that CDT weights these acts according to one's credence, while FDT counts them all equally, as long as they are runs of the same algorithm.

5.3 The simulator Newcomb problem

Return now to Nozick's original specification. There is some being, who says they have already predicted whether you would one-box or two-box, and has already put the money in the box (or not). But this being hasn't revealed how they make their predictions. However, suppose that you have discovered that the being has a dark secret:

> The being has access to another planet, and lots of energy and resources. Every time they want to predict how someone will make a choice on Earth, they create a full physical copy of that person on that planet. They copy the full circumstances of this choice, and see how the duplicate behaves. However, once the duplicate has made their choice, the being destroys the copy, and uses that result for their prediction here on Earth.[18]

Having learned this, when you find yourself facing the choice, you now realize that you might actually be the duplicate rather than the original. If you are in fact the original, then your choice will be causally efficacious only at determining whether you get the additional $1000 or not. But if you are the duplicate, then you realize that you are about to be destroyed soon, regardless of what you do, and will never receive any winnings. Your choice will only be causally efficacious insofar as it leads the being to either put the $1,000,000 into the box for the original or not. This puts a damper on the emotional impact of $1,000,000. But if the duplicate and the original both care about each other as much as they care about themselves (as a parent on their deathbed might care about the well-being of their child as much as themselves), then like in the Sleeping Beauty version, it appears that CDT recommends one-boxing.[19]

A version of this reasoning is discussed by Elga (2004), though without the Newcomb-style decision problem. (Oddly, the *victim* of the duplication in his paper is

[18] Formulating the problem by making the predictor run a simulation of the agent has been discussed at least as early as 2005, https://scottaaronson.blog/?p=30.

[19] Note that the duplicate doesn't actually get any benefit based on what is actually in *her* boxes — the only payoff is the vicarious payoff to the original. This makes the role of the duplicate in this version of the problem importantly different from that of Lewis (1979).

Dr. Evil, and Elga makes no comment on the moral status of the perpetrator of the duplication.) Interestingly though, Elga (2004, p. 386–388) suggests that some of the details of *how* such a simulation is realized might be relevant to whether one should take it as seriously possible that one is the duplicate, or if some anti-skeptical principle can let one know that one is the original.[20] But if credences should be equal, then CDT will make the same recommendation as FDT for any specification of the outcomes. Even if credences can be unequal (for instance, if we are uncertain whether the predictor works in this way, so we are uncertain whether the duplicate exists, even though we are certain that the original exists), as long as their ratios are not larger than the ratios of the outcomes, the recommendation will be the same.

5.4 Wrap-up

Such cases can be proliferated. Each involves the possibility of your current decision affecting the outcome for other decisions structurally like this one: in the Iterated Newcomb Problem, your current decision affects the outcome of possible future decisions; in the Sleeping Beauty Newcomb Problem, the Monday decision affects the Tuesday outcome; and in the Simulator Newcomb Problem, the decision of the duplicate affects the original. Since some of the salient possibilities in each would render your choice causally relevant to the future prediction, CDT recommends one-boxing. Many authors take it as implicitly stipulated that there is no causal impact beyond the present outcome — in such cases CDT recommends two-boxing. But if an agent thinks that a choice offered once may be offered again, and that behavior now might affect the choices that are offered in the future or the predictions about how they would be made, then they may follow CDT intuitions but still end up one-boxing.

The reasoning by which CDT recommends one-boxing in these cases is in some important ways analogous to that by which FDT does. CDT recommends one-boxing because it considers what the outcome would be if we imagined changing the output of just one instance of the decision algorithm. But because of the setup of the problem, we don't know *which* instance of the algorithm is being changed, so we weigh all their effects according to our credence in each (and if an indifference principle is available, then these weights are equal). By comparison, FDT recommends one-boxing because it considers what the outcome would be if we imagined changing the output of *all* these instances of the decision algorithm, each weighted fully.

6 Behavior modification without binding

In the versions of Newcomb's Problem from the previous section, one-boxing was effective at bringing about good outcomes due to possible direct influence on the being's future prediction. The ones that we see in this section influence the prediction by affecting your own future behavior. Where the former were analogous to the pseudo-Newcomb problem involving backwards causation, these will be parallel to

[20] Various versions of this principle, in skeptical and anti-skeptical formulations, are considered by Smithson (2017) and Builes (2024).

the one involving binding. Unlike in the original binding case, your future behaviors will not be affected by directly binding yourself to some future choice. Rather, these cases capitalize on the ways our current choices shape our future selves. Metaphorically, the contrast is between, on the one hand, Odysseus binding himself to the mast to resist the call of the sirens and, on the other, undertaking a course of exposure therapy to develop the ability to resist their call.

But, undergoing exposure therapy is not necessary for this to obtain — the mere fact of acting in a certain way once can help make this way of acting habitual. There is significant evidence that the choices made by real-world, human agents on the basis of cognitively intensive deliberation can have downstream impacts on future habitual behavior.[21] Idealized agents capable of complete, rational deliberation at every moment in time may not have their behavior shaped in these ways. But we find it plausible that any finite, physical being will have some analogue of some of these effects.[22] If a being doesn't usually have the time and energy to perform a complete, rational deliberation at every moment in time, it would make sense for the outputs of deliberation to be cached in some way for easy retrieval in case another similar situation arises.[23]

The cases we consider in this section take advantage of this effect. Sometimes our present interests and preferences align with our long-term ones. But, often the options that seem most appealing at a given moment are in conflict with those that contribute to our overall well-being. The cases that we will now explore hinge on precisely this sort of inconsistency between what we currently prefer and the patterns of behavior that we would most prefer to embody in the long-run.[24]

6.1 Iterated self-prediction

Many of our previous cases involved some amount of science fiction; the ones we will now consider are more realistic (in some, perhaps uncomfortably so). Self-prediction will play a significant role since your current choice may foreseeably lead to the formation of habits guiding your future behavior. Self-prediction and habit formation are valuable tools for shaping our future selves and contributing to our overall long-term well-being. One way to cash this out is as follows: Your current time slice values the well-being of *you-now* over *you-tomorrow*. However, each of your time slices also values *you* — the four-dimensional object composed of your various temporal parts. There is no conflict when the interests of a particular time slice align

[21] See (Smith & Graybiel, 2016) for a review of the evidence surrounding the mechanisms of habit formation and (Kruglanski & Szumowska, 2020) for a discussion of the connection between habitual and goal-driven behavior.

[22] Indeed, a common method in reinforcement learning involves *dynamic programming* where previously stored estimates of the value of a choice are used to guide and optimize future decisions instead of being recomputed every time — see (Sutton & Barto, 2018).

[23] See (Thorstad, 2024) for a recent discussion in the context of the accuracy-coherence tradeoff for bounded rationality.

[24] Much of the discussion in this section is based on the accounts provided by Ainslie (2001) and Rachlin (2000). There are also some interesting connections between these ideas and the discussion provided by Oddie (1993).

perfectly with those of the series. However, trouble arises when they don't, as in the following case that is likely familiar to most readers:

> You want to establish a new routine of regularly going to the gym. You know that doing so will have significant benefits for your long-term health and overall well-being. However, you also know that on any given day you simply don't feel like going to the gym.

From any given time slice's perspective, the best outcome is to stay home that day, but for all future time slices to go to the gym. Importantly, while any time slice cares both about its own momentary enjoyment and about the fitness of the self extended through time, no time slice cares about the momentary enjoyment of any other time slice. The momentary enjoyment of staying home contributes slightly more than 1 unit to the payoff, and costs a tiny bit of fitness, leading to an overall gain of a single unit of payoff. Each 25% of the other time slices going to the gym will contribute an amount of fitness that adds 10 units to the payoff. For simplicity, then, we can think of the current time slice's payoff as determined by (i) whether it goes to the gym, and (ii) what percentage of the future time slices go (Table 2).

As with two-boxing in Newcomb's Problem, staying home dominates going to the gym. However, as with the prediction in Newcomb's Problem, the behavior of one's other time slices is correlated with one's current behavior. The idea we want to draw attention to here is that one's current behavior may not merely be *correlated* with one's future behavior, but may in fact *causally* influence it by helping build a habit. Going to the gym today can make it more likely that you'll go to the gym in the future, and staying home today can make it more likely that you'll stay home. FDT evaluates these cases by considering which instances of the decision are controlled by one's current algorithm, while CDT evaluates them by evaluating the causal influence of one decision on all future decisions.

Another example is provided in the following Public Goods Game adapted from (Ainslie, 2001, p. 93):

> There is a large stadium filled with people who are a lot like you, who have all been instructed to play this game to maximize their individual earnings. Players are given the opportunity to announce that they choose `cooperate` and reward 10 cents to themselves and all other players, or `defect` and reward $1 to only themselves. When making their choice, each player is aware of the choices made by each of the previous players. The game will conclude at a random point after at least 20 players have made their announcements, at which point everyone is invited to collect their earnings. Supposing you are the first player, what should you do?

Table 2 Gym payoffs for *you-now* that depend on whether you go to the gym or stay home at this time, and on what fraction of your other times go to the gym

		You-other-times going				
		0%	25%	50%	75%	100%
You-now	Go	0	10	20	30	40
	Stay	1	11	21	31	41

If there were common knowledge of when the game would conclude, and that all players were in fact choosing optimally according to CDT, then a backwards induction argument shows that everyone will `defect`. Thus, CDT recommends that you also `defect`.

But, absent such knowledge, the choice is not so simple. The optimal choice depends on a prediction about how many subsequent players there will be, and how they will choose. If one player already knew what all successive players would choose, then CDT would clearly recommend that they `defect` since there would be no gain from `cooperate`. However, if they suspect that some of the future players might be causally influenced by the behavior of past players, then CDT takes this into account. If choosing `cooperate` increases their confidence that future players do as well, then CDT may well recommend `cooperate`. Likewise, CDT can recommend a time slice to go to the gym if it's not common knowledge among your time slices that they all follow the recommendations of CDT, and there is some possibility that going to the gym now will cause an increase in probability that future time slices go to the gym. The FDT evaluation of this case depends on how the behavior of each player or time slice is influenced by the output of the decision algorithm in question, which may or may not track these causal influences.

6.2 Masking in a pandemic

Our next case makes the one-boxing-like recommendations of CDT more explicit.

> A highly contagious respiratory illness has spread globally. During a spike in cases there's a moderate chance you will get sick from entering an indoor public space. This risk can be significantly cut by wearing a mask. In the period between spikes, the chance of getting sick is negligible (regardless of a mask). Wearing a mask is uncomfortable, but the complications of getting sick are far worse. You judge that the benefit of even the incomplete risk mitigation of masking during a spike is a thousand times as large as the unpleasantness of wearing a mask, but the benefit is negligible between spikes. If you were always sufficiently attentive, the decision would be simple: Wear a mask during a spike, and don't when there's not. However, you're an absent-minded philosopher and you've observed that half the time you forget to check case levels before going out. When you do remember, you choose deliberately; when you forget, you act out of habit. The probability that your habitual behavior results in wearing a mask is equal to the fraction of the last ten times you went out and wore a mask (whether attentively or habitually). You are about to enter a public space and, having just checked the case numbers, you are making a deliberate choice rather than acting habitually.

If there is currently a spike in cases, you should obviously wear the mask. But if there isn't, then the decision isn't quite as simple. Although wearing the mask comes with slight discomfort and offers no substantive present benefit, it does causally influence your future behavior. If spikes tend to come and go quickly compared to how frequently you enter indoor public spaces, then you probably will soon do so during a spike. The 50% chance that you will forget to check during such a spike, combined with the 10% impact on your habitual behavior that your current mask use causes, plausibly gives you reason enough to wear the mask now. Even if you are confident that your *next* visit will be before the next spike, the same reasoning still applies so long as there is a good chance that *at least one of your next ten visits* will occur during a spike.

Interestingly, the reasoning generalizes even further! Even if you are certain that none of the next ten visits will occur during a spike, if there's a good chance that the *eleventh* visit will be in a spike, then wearing the mask this time has a reasonable chance of affecting your habitual behavior on that eleventh visit, by causing habitual mask use on some of the next ten visits. However, if you are confident that the next spike is sufficiently far off and that you will at least *suspect* it in advance, then it will be causally rational now not to wear a mask on this visit. The details of a complete analysis seem quite complicated, but relevant factors include the number of visits it takes to build a habit (ten, in the specification above), the probability of acting habitually rather than deliberately on any particular visit (50% as specified above), your current confidence that a spike is coming (unspecified above), and your confidence that you will have some suspicion of a spike in advance (similarly unspecified).

We can now use this to generate a version of Newcomb's Problem with the same structure.

> You are often offered choices between one and two boxes. Sometimes, the choice is offered by Nozick's being, as in the original problem, having made a prediction about your decision and put $1,000,000 in the first box or not on the basis of this prediction. But sometimes the choice happens without the being, and in such cases there is always nothing in the first box, though there is still $1000 in the second box. As in the masking case, you know there is generally a 50% chance that you forget to check whether or not the being is involved (we might need smaller payoffs to make it plausible that you would ever forget to check), and you act habitually with a probability specified by the fraction of the past 10 decisions you made. The being is generally able to foresee which action you will make, regardless of whether it is a conscious choice or habitual behavior governed by a process that you think of as probabilistic.

In this example, wearing a mask is analogous to one-boxing and not wearing a mask to two-boxing. The discomfort of wearing a mask (both now and on a future visit where you might act habitually rather than deliberating) stands in for the $1000 that you forego by taking only one box. The benefit of mitigating your risk of infection (either if there is a spike now, or on the chance there is a spike during a future visit when you are acting habitually) stands in for the $1,000,000. When the being is not predicting, one-boxing has only negative effects now, just as between spikes, masking has only negative effects now. However, it may still be recommended because of the benefits accrued from building habitual probabilities of one-boxing or masking on future visits when the being or a viral spike is present. This pattern of behavior (of doing something unpleasant with no present benefit in order to make it more likely that you'll do it when it actually matters) is common in rehearsals, fire drills, and many other cases of extended self-control, like exercise and masking.

As in the original binding case, CDT recommends one-boxing due to the predictable effects of your current choice on your future behavior. FDT's analysis must settle whether the decisions to mask or not at each moment are controlled by the same algorithm or distinct ones, and whether the input strings to the algorithm are the same or relevantly different. CDT allows for intermediate weight based on intermediate causal control.

6.3 Self-prediction and policy selection

Other examples involving willpower and self-control follow a similar dynamic to the problem of going to the gym. For example:

- You come home after a long day. You could either spend further time cooking dinner for cheap, or order delivery for $40.
- You're a poor graduate student who is out late and had a few drinks after a colloquium dinner. You can either take the subway for cheap or spend more on an Uber.
- You go out for a short bike ride, but it's a hot day. You can either wear an uncomfortable helmet for safety or go helmet-free at a slightly higher risk.

In each, your current decision appears minor. Spending a few extra bucks on dinner won't make a big difference in your life. Nonetheless, we often agonize over these sorts of decisions far more than would seem justified given their low stakes. Perhaps this is due to worry we are influencing a larger personal policy. If you take an Uber today, that makes it more likely that you'll more frequently take an Uber. If you don't wear a helmet now, you might stop wearing a helmet generally when it's just a bit uncomfortable, which may add up to more risk than one wants in the long run.

The optimal long-run behavior is probably to order delivery a certain percentage of the time, occasionally take an Uber, and rarely leave off your helmet (Icard, 2021). But in these cases, each time slice wants to be the exception to the onerous rule. As Ainslie (2001, p. 88) writes, "Long range interests define principles, and short-range interests find exceptions."

When we make any of these repeated decisions, we are both predictor and predictee. Each time slice has the option of taking their own utility (e.g., by ordering delivery) or not. But this increases the likelihood that future time slices will do so as well, thereby reducing the amount of utility that you'll end up with in the long run. It is often better to avoid temptation always, or nearly always, than risk a cascading chain of defections. Ainslie (2001, p. 89) argues:

> Principles of choice, or "personal rules", represent self-enforcing contracts with your future motivational states; such contracts depend on your seeing each current choice as a precedent that predicts how you're apt to choose among similar options in the future. Short-range interests evade personal rules by proposing exceptions that might keep the present case from setting a precedent. The will is a recursive process that bets the expected value of your future self-control against each of your successive temptations.

FDT addresses these cases explicitly as ones in which we govern our actions by choosing as if we are choosing a single algorithm for all of our time slices, duplicates, etc. to implement. Classical CDT analyses of these cases presuppose that present acts have no causal effect on future ones, sometimes lamenting the inability of humans to bind themselves effectively. But we think that something like the model of occasionally relying on past cached behaviors will be common for computationally limited agents in a complex world. Thus, while no decision can *perfectly* bind future action, we find the classical model where future action is perfectly *un*-bound by present decision implausible as well. In situations with partial control through practice and habit, CDT often yields the same action recommendation as FDT.

Some defenders of FDT base their defense on intuitions that one-boxing (or the relevant equivalent) is rational. But, as we have seen, CDT may yield the same verdict when paired with a more realistic understanding of human psychology, in which our patterns of long term behavior are partly caused by habits and patterns set up by past behavior. Cases where CDT and FDT clearly yield different recommendations involve unrealistically precise stipulations about the effectiveness of our deliberation and decision algorithms. While theoretical reflection on deliberation and choice might most naturally suggest these precise stipulations, our intuitions about cases are trained in a world where we are imperfect in many ways analogous to the ones explored here.

The main contrast between CDT and FDT is how each decision theory views our intertemporal psychological connectedness. CDT *can* recognize the power our habits and dispositions have over our decision making in the long run, but only when the right sort of causal connection is added to the model. As we saw in the iterated self-prediction cases, CDT's recommendations are often highly sensitive to the precise specification of the case. Tell the story slightly differently and CDT will recommend staying home from the gym. With FDT, because each day's choice is viewed as an instance of the output of the same decision procedure, the connection between your time slices and the importance of good habits is more robust.

FDT requires drawing a clear distinction between which entities in the world are or aren't controlled by one's decision procedure, and imagining what would happen if they all had one output or another. CDT can yield these same recommendations in these cases by attending to the sort of mechanisms described eloquently by William James (1890, p. 127):

> Every smallest stroke of virtue or of vice leaves its never so little scar. The drunken Rip Van Winkle, in Jefferson's play, excuses himself for every fresh dereliction by saying, 'I won't count this time!' Well! he may not count it, and a kind Heaven may not count it; but it is being counted none the less. Down among his nerve-cells and fibres the molecules are counting it, registering and storing it up to be used against him when the next temptation comes. Nothing we ever do is, in strict scientific literalness, wiped out. Of course, this has its good side as well as its bad one. As we become permanent drunkards by so many separate drinks, so we become saints in the moral, and authorities and

experts in the practical and scientific spheres, by so many separate acts and hours of work.

The theorist's sensibilities will likely determine whether the fragile intertemporal connections of CDT or the resilient connections of FDT are more attractive. But we see the question of how to view the connection between various time slices and the value of habits as central for determining the correct theory of rationality.

7 Conclusion

We have now seen several ways of specifying the underlying causal story in Newcomb's Problem in which CDT recommends one-boxing. To be clear, in a classic Newcomb case, with no backwards causation, the decision made after the prediction, and no relevant causal impacts of the decision beyond this current case, either on your own future behavior or future predictions about you (a scenario we like to think of as "the last Newcomb problem ever"), CDT clearly recommends two-boxing. Our central claim is just that there are many other apparently similar cases in which CDT recommends one-boxing. Some work by iterations (or duplications) that mimic backwards causation. Others work by habits that mimic binding. Moreover, the influences of habit and iteration are common features of everyday human interactions, and thus our intuitions might respond to them subconsciously even when we try to explicitly stipulate them away.

As we have suggested, careful consideration of these cases helps elucidate the relationship between CDT and contemporary rivals. One way to characterize the difference between CDT and FDT is that CDT takes the *act* to be the bearer of rationality, while FDT takes the *decision rule* to be the bearer of rationality.[25] Some defenders of FDT use intuitive judgments about cases to motivate their view, but if CDT can actually yield the same judgments, then the evaluation of the theories needs to focus more conceptually on this point about the unit of evaluation. Proponents of CDT have ignored this question in the past, since their main interlocutors have supported EDT, which agrees with CDT on the unit of evaluation but disagrees on how it is evaluated. One way to characterize the distinction between CDT and FDT is as agreeing on the *method* of evaluation, but disagreeing on the *unit* of evaluation. Rather than seeing which *acts* will *cause* good outcomes (CDT) or which *acts* are *correlated* with good outcomes (EDT), we see which *decision rules* will *cause* good outcomes (FDT).

We end by reconsidering the original Newcomb Problem. What have we learned? EDT naively seems to recommend one-boxing in all the Newcomb problems and not smoking in the medical Newcomb problem. But, if Eells' tickle defense is deemed successful, EDT recommends smoking in the medical Newcomb problem as well. CDT naively seems to recommend two-boxing in all the Newcomb problems, and smoking in the medical Newcomb problems. But, with our observations about iteration and habits, it may in fact recommend one-boxing in many Newcomb problems

[25] Though see (Yudkowsky & Soares, 2017) for a different interpretation.

(though not in fully idealized ones). Both sets of recommendations are similar to those made by FDT.

Since people are used to making intuitive decisions when embedded in a larger world with repeated interactions, and often form habits to simplify many complexities, intuitive judgments can easily track these versions rather than the case as officially stipulated. Thus, if these cases are intended to test intuitions, then it is important to specify the case very carefully to ensure that these other features don't interfere (or else specify that these other features are clearly involved, if their "interference" is the point the case is meant to test). Furthermore, the fact that many specifications lead to this conclusion and that such complications permeate real world cases may mean that CDT really does recommend one-boxing in many of the apparently Newcomb-like situations we face because of reasonable uncertainty of the prediction mechanism and whether it is iterated.

Acknowledgements We would like to acknowledge the valuable feedback provided by audiences at CU Boulder's Center for Values and Social Policy (2022), University of Cologne's CONCEPT Brown Bag Meeting (2022), and the Chambers Philosophy Conference at UN Lincoln (2023), where this paper was presented. We also express particular appreciation to Arc Kocurek, Graham Oddie, Caspar Oesterheld, and Brian Skyrms for their helpful feedback on earlier drafts of this paper. Lastly, we express gratitude to Branden Fitelson for hosting the Northeastern Epistemology Workshop (2021), where we first had the conversation that gave rise to this paper.

Data availability We do not analyse or generate any datasets, because our work proceeds within a theoretical and mathematical approach.

References

Ahmed, A. (2014). *Evidence decision and causality*. Cambridge University Press.

Ainslie, G. (2001). *Breakdown of the Will*. Cambridge University Press

Arntzenius, F., Elga, A., & Hawthorne, J. (2004). Bayesianism, infinite decisions, and binding. *Mind, 113*(450), 251–283. https://doi.org/10.1093/mind/113.450.251

Aumann, R. J., Hart, S., & Perry, M. (1997). The absent-minder driver. *Games and Economic Behavior, 20*(1), 102–116. https://doi.org/10.1006/game.1997.0577

Axelrod, R. (1980). Effective choice in the Prisoner's Dilemma. *Journal of Conflict Resolution, 24*(1), 3–25. https://doi.org/10.1177/002200278002400101

Barnes, R. E. (1997). Rationality, dispositions, and the newcomb paradox. *Philosophical Studies, 88*(1), 1–28. https://doi.org/10.1023/A:1004252605819

Bourget, D., & Chalmers, D. (2014). What do philosophers believe? *Philosophical Studies, 170*(3), 465–500. https://doi.org/10.1007/s11098-013-0259-7

Bourget, D., & Chalmers, D. J. (2023). Philosophers on philosophy: The PhilPapers 2020 survey. *Philosophers' Imprint, 23*, 11. https://doi.org/10.3998/phimp.2109

Brams, S. J. (1975). Newcomb's problem and Prisoners' Dilemma. *The Journal of Conflict Resolution, 19*(4), 596–612. https://doi.org/10.1177/002200277501900402
Briggs, R. (2010). Putting a value on Beauty. In T. Szabó Gendler & J. Hawthorne (Eds.), *Oxford Studies in Epistemology Volume 3* (pp. 3–24). Oxford University Press
Builes, D. (2024). *Center Indifference and Skepticism. Nôus,58*(3), 778–798. https://doi.org/10.1111/nous.12478
Campbell, R., & Sowden, L. (1985). *Paradoxes of rationality and cooperation: prisoner's dilemma and newcomb's problem*. University of British Columbia Press
Carr, J. R. (2017). Epistemic utility theory and the aim of belief. *Philosophy and Phenomenological Research, 95*(3), 511–534. https://doi.org/10.1111/phpr.12436
Conitzer, V. (2015). Can rational choice guide us to correct de se beliefs? *Synthese, 192*(12), 4107–4119. https://doi.org/10.1007/s11229-015-0737-x
Easwaran, K. (2021). A classification of Newcomb problems and decision theories. *Synthese, 198*(27), 6415–6434. https://doi.org/10.1007/s11229-019-02272-z
Eells, E. (1981). Causality, utility, and decision. *Synthese, 48*(2), 295–329. https://doi.org/10.1007/BF01063891
Eells, E. (1982). *Rational Decision and Causality*. Cambridge University Press.
Egan, A. (2007). Some counterexamples to causal decision theory. *The Philosophical Review, 116*(1), 93–114. https://doi.org/10.1215/00318108-2006-023
Elga, A. (2000). Self-locating belief and the Sleeping Beauty problem. *Analysis, 60*(2), 143–147. https://doi.org/10.1093/analys/60.2.143
Elga, A. (2004). Defeating Dr. Evil with self-locating belief. *Philosophy and Phenomenological Research, 69*(2), 383–396. https://doi.org/10.1111/j.1933-1592.2004.tb00400.x
Fisher, R. A. (1959). *Smoking*. The Cancer Controversy: Some Attempts to Assess the Evidence. Oliver and Boyd.
Gauthier, D. (1994). Assure and threaten. *Ethics, 104*(4), 690–721. https://doi.org/10.1086/293651
Gödel, K. (1949). An example of a new type of cosmological solutions of einstein's field equations of gravitation. *Review of Modern Physics, 21*(3), 447–450. https://doi.org/10.1103/RevModPhys.21.447
Hall, N. (1994). Correcting the guide to objective chance. *Mind, 103*(412), 505–517. https://doi.org/10.1093/mind/103.412.505
Hedden, B. (2015). Time-slice rationality. *Mind, 124*(494), 449–491. https://doi.org/10.1093/mind/fzu181
Hedden, B. (2015). *Reasons without persons: Rationality, identity, and time*. Oxford University Press.
Hofstadter, D. (1985). *Metamagical Themas: Questing for the Essence of Mind and Pattern*. Basic Books.
Hurley, S. L. (1991). Newcomb's problem, Prisoners' Dilemma, and collective action. *Synthese, 86*(2), 173–196. https://doi.org/10.1007/BF00485806
Huttegger, S. (2023). Reconciling evidential and causal decision theory. *Philosophers' Imprint, 23*, 20. https://doi.org/10.3998/phimp.931
Icard, T. (2021). Why be random? *Mind, 130*(517), 111–139. https://doi.org/10.1093/mind/fzz065
Isaacs, Y., & Levinstein, B. (2024). Evidential decision theory and the Ostrich. *Philosophers' Imprint, 24*(6), 1–11. https://doi.org/10.3998/phimp.3176
James, W. (1890). *Principles of Psychology*. Henry Holt and Company.
Joyce, J. (2018). Deliberation and stability in Newcomb problems and pseudo-Newcomb problems. In A. Ahmed (Ed.), *Newcomb's Problem* (pp. 138–159). Cambridge University Press.
Kocurek, A. W. (2024). Counterpossibles, Functional Decision Theory, and Artificial Agents. In F. Carcassi, T. Johnson, S. B. Knudstorp, S. Domínguez Parrado, P. Rivas Robledo & G. Sbardolini (Eds.), *Proceedings of the 24th Amsterdam Colloquium* (pp. 218–225).
Korzukhin, T. (2021). A Dutch book for CDT thirders. *Synthese, 198*(12), 11925–11941. https://doi.org/10.1007/s11229-020-02841-7
Kruglanski, A., & Szumowska, E. (2020). Habitual behavior is goal-driven. *Perspectives on Psychological Science, 15*(5), 1256–1271. https://doi.org/10.1177/1745691620917676
Kuhn, S. (2019). Prisoner's Dilemma. In E. N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy* (Winter 2019 Edition), https://plato.stanford.edu/archives/win2019/entries/prisoner-dilemma/.
Levi, I. (1975). Newcomb's many problems. *Theory and Decision, 6*(2), 161–175. https://doi.org/10.1007/BF00169104
Levi, I. (1982). A note on Newcombmania. *Journal of Philosophy, 79*(6), 337–342. https://doi.org/10.2307/2026081

Levinstein, B., & Soares, N. (2020). Cheating death in Damascus. *Journal of Philosophy, 117*(5), 237–266. https://doi.org/10.5840/jphil2020117516
Lewis, D. (1979). Prisoners' Dilemma is a Newcomb Problem. *Philosophy & Public Affairs*, 8(3), 235–240. https://www.jstor.org/stable/2265034
Lewis, D. (1981). Causal Decision Theory. *Australas. J. Philos., 59*(1), 5–30. https://doi.org/10.1080/00048408112340011
Li, Z., & Zenker, F. (2021). Newcomb's problem isn't a choice dilemma. *Synthese, 199*(1–2), 5125–5143. https://doi.org/10.1007/s11229-020-03018-y
Mackie, J. L. (1977). Newcomb's Paradox and the direction of causation. *Canadian Journal of Philosophy, 7*(2), 213–225. https://doi.org/10.1080/00455091.1977.10717014
Maitzen, S., & Wilson, G. (2003). Newcomb's hidden regress. *Theory and Decision, 54*(2), 151–162. https://doi.org/10.1023/A:1026249026940
McKay, P. (2004). Newcomb's problem: The Causalists get rich. *Analysis, 64*(2), 187–189. https://doi.org/10.1093/analys/64.2.187
Meacham, C. J. G. (2010). Binding and its consequences. *Philosophical Studies, 149*(1), 49–71. https://doi.org/10.1007/s11098-010-9539-7
Moss, S. (2015). Time-slice epistemology and action under indeterminacy. In T. Szabó Gendler & J. Hawthorne (Eds.), *Oxford studies in Epistemology Volume 5* (pp. 172–194). Oxford University Press.
Nozick, R. (1969). Newcomb's Problem and Two Principles of Choice. In N. Rescher (Ed.), *Essays in Honor of Carl G. Hempel* (pp. 114–146). Springer.
Oddie, G. (1993). Act and maxim: value-discrepancy and two theories of power. *Philosophy and Phenomenological Research, 53*(1), 71–92. https://doi.org/10.2307/2108054
Oesterheld, C. (2022). Understanding the Tickle Defense in Decision Theory, https://www.andrew.cmu.edu/user/coesterh/TickleDefenseIntro.pdf
Oesterheld, C., & Conitzer, V. (2021). Extracting money from causal decision theorists. *The Philosophical Quarterly, 71*(4), 701–716. https://doi.org/10.1093/pq/pqaa086
Pearl, J. (2000). *Causality: models, reasoning, and inference*. Cambridge University Press.
Piccione, M., & Rubinstein, A. (1997). On the interpretation of decision problems with imperfect recall. *Games and Economic Behavior, 20*(1), 3–24. https://doi.org/10.1006/game.1997.0536
Rachlin, H. (2000). *The science of self-control*. Harvard University Press.
Savage, L. J. (1954). *The foundations of statistics*. John Wiley.
Schwarz, W. (2015). Lost memories and useless coins: Revisiting the absent-minded driver. *Synthese, 192*(9), 3011–3036. https://doi.org/10.1007/s11229-015-0699-z
Sider, T. (2001). *Four dimensionalism: An ontology of persistence and time*. Oxford University Press.
Skyrms, B. (1980). *Causal necessity: A pragmatic investigation of the necessity of laws*. Yale University Press.
Skyrms, B. (1982). Causal decision theory. *Journal of Philosophy, 79*(11), 695–711. https://doi.org/10.2307/2026547
Slezak, P. (2006). Demons, deceivers and liars: Newcomb's Malin Génie. *Theory and Decision, 61*(3), 277–303. https://doi.org/10.1007/s11238-006-9002-6
Smith, K. S., & Graybiel, A. M. (2016). Habit formation. *Dialogues Clin. Neurosci., 18*(1), 33–43. https://doi.org/10.31887/DCNS.2016.18.1/ksmith
Smithson, R. (2017). The Principle of Indifference and Inductive Skepticism. *British Journal for the Philosophy of Science, 68*(1), 253–272. https://doi.org/10.1093/bjps/axv029
Sorensen, R. (1985). The iterated versions of Newcomb's Problem and the Prisoner's Dilemma. *Synthese, 63*(2), 157–166. https://doi.org/10.1007/BF00485365
Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction, and search*. MIT Press.
Sullivan, M. (2018). *Time Biases: A theory of rational planning and personal persistence*. Oxford University Press.
Sutton, R. S., & Barto, A. G. (2018). *Reinforecment learning: An introduction*. MIT Press.
Talbott, W. J. (1987). Standard and non-standard Newcomb Problems. *Synthese, 70*(3), 415–458. https://doi.org/10.1007/BF00414158
Thorstad, D. (2024). The accuracy-coherence tradeoff in cognition. *British Journal for the Philosophy of Science, 75*(3), 695–715. https://doi.org/10.1086/716518
Velleman, J. D. (1997). Deciding how to decide. In G. Cullity & B. N. Gaut (Eds.), *Ethics and Practical Reason* (pp. 29–52). Oxford University Press.

Wolpert, D. H., & Benford, G. (2013). The lesson of Newcomb's paradox. *Synthese, 190*(9), 1637–1646. https://doi.org/10.1007/s11229-011-9899-3

Yudkowsky, E. & Soares, N. (2017). Functional Decision Theory: A New Theory of Instrumental Rationality, https://arxiv.org/abs/1710.05060

Yudkowsky, E. (2010). Timeless Decision Theory. The Singularity Institute.