
Assumption: agent knows its own epistemic state approximately, with high accuracy
E
Agent's epistemic state
Assumption: the agent's action is a function of epistemic state alone; no other parents
A
Agent's action

CDT=EDT?

1. Smoking Lesion Steelman
2. Smoking Lesion Steelman II
3. Smoking Lesion Steelman III: Revenge of the Tickle Defense
4. Comparing LICDT and LIEDT
5. Mixed-Strategy Ratifiability Implies CDT=EDT
6. XOR Blackmail & Causality
7. A Rationality Condition for CDT Is That It Equal EDT (Part 1)
8. A Rationality Condition for CDT Is That It Equal EDT (Part 2)
9. Dutch-Booking CDT
10. CDT=EDT=UDT
11. Troll Bridge
12. Dutch-Booking CDT: Revised Argument
13. My Current Take on Counterfactuals

Smoking Lesion Steelman

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

It seems plausible to me that any example I've seen so far which seems to require causal/counterfactual reasoning is more properly solved by taking the right updateless perspective, and taking the action or policy which achieves maximum expected utility from that perspective. If this were the right view, then the aim would be to construct something like updateless EDT.

I give a variant of the smoking lesion problem which overcomes an objection to the classic smoking lesion, and which is solved correctly by CDT, but which is not solved by updateless EDT.

UDT as originally described involved a "mathematical intuition module" which would take some sort of logical counterfactual. However, I'll be using the term "updateless" purely to describe the decision theory you get by asking another decision theory to choose a policy as soon as it is born, rather than using that decision theory all along. Hence, updateless CDT is what you get when you ask a CDT agent to choose a policy; updateless EDT is what you get when you ask an EDT agent to choose a policy.

I'll also be treating "counterfactual" as synonymous with "causal". There are cases where physical causal reasoning seem to give the wrong counterfactual structure, like Newcomb's problem. I won't be trying to solve that problem here; I'm more trying to ask whether there are any cases where causal/counterfactual reasoning looks like what we really want at all.

The "common wisdom", as I have observed it, is that we should be aiming to construct something like an updateless CDT which works well with logical uncertainty. I'm not sure whether that would be the dominant opinion right now, but certainly TDT set things in this direction early on. From my perspective, I don't think it's been adequately established that we should prefer updateless CDT to updateless EDT; providing some evidence on that is the implicit aim of this post. Explicitly, I'll mostly be contrasting updateful CDT with updateful EDT.

It might be impossible to construct an appropriate logically-updateless perspective, in which case we need logical counterfactuals to compute the effects of actions/policies; but, in some sense this would only be because we couldn't make a sufficiently ignorant prior. (I think of my own attempt at logically updateless decisions this way.) However, that would be unfortunate; it should be easier to construct good counterfactuals if we have a stronger justification for counterfactual reasoning being what we really want. Hence, another aim of this post is to help provide constraints on what good counterfactual reasoning should look like, by digging into reasons to want counterfactuals.

Thanks go to Alex Mennen, Evan Lloyd, and Daniel Demski for conversations sharpening these ideas.

Why did anyone think CDT was a good idea?

The original reasons for preferring CDT to EDT are largely suspect, falling to the "Why Ain'cha Rich?" objection. From the LessWrong/MIRI perspective, it's quite surprising that Newcomb's Problem was the original motivation *for* CDT, when we now use it as a point against. This is not the only example. The SEP article on CDT gives Prisoner's Dilemma as the first example of CDT's importance, pointing out that EDT cooperates with a copy of itself in PD because unlike CDT it fails to take into account that cooperating is "auspicious but not efficacious".

The Smoking Lesion problem isn't vulnerable to "Why Ain'cha Rich?", and so has been a more popular justification for CDT in the LessWrong-sphere. However, I don't think it provides a good justification for CDT at all. The problem is ill-posed: it is assumed that those with a smoking lesion are more likely to smoke, but this is inconsistent with their being EDT agents (who are not inclined to smoke given the problem setup). (Cheating Death in Damascus) points out than Murder Lesion is ill-posed due to similar problems.)

So, for some time, I have thought that the main effective argument against EDT and for CDT was XOR blackmail. However, XOR blackmail is also solved by updateless EDT. We want to go updateless either way, so this doesn't give us reason to favor CDT over EDT.

Regardless, I'm quite sympathetic to the *intuition* behind CDT, namely that it's important to consider the counterfactual consequences of your actions rather than just the conditional expected utilities. Furthermore, the following idea (which I think I got from an academic paper on CDT, but haven't been able to track down which one) seems at least plausible to me:

> *If we were dealing with ideal decision agents, who could condition on all the inputs to their decision process, CDT would equal EDT. However, imperfect agents often cannot condition on all their inputs. In this situation, EDT will get things wrong. CDT corrects this error by cutting the relationships which would be screened off if only we could condition on those inputs.*

This intuition might seem odd given all the cases where CDT doesn't do so well. If CDT fails Newcomb's problem and EDT doesn't, it seems CDT is at best a hack which repairs some cases fitting the above description at the expense of damaging performance in other cases. Perhaps this is the right perspective. But, we could also think of Newcomblike decision problems as cases where the classical causal structure is just wrong. With the right causal structure, we might postulate, CDT would always be as good as or better than EDT. This is part of what TDT/FDT is.

I'll charitably assume that we can find appropriate causal structures. The question is, even given that, can we make any sense of the argument for CDT? Smoking Lesion was supposed to be an example of this, but the problem was ill-posed. So, can we repair it?

A Smoking Lesion Steelman

Agents who Don't Know their Utility Function

I'll be assuming a very particular form of utility function ignorance. Suppose that agents do not have access to their own source code. Furthermore, whether CDT or EDT, the agents have an "epistemic module" which holds the world-model: either just the probability distribution (for EDT) or the probability distribution plus causal beliefs. This epistemic module is ignorant of the utility function. However, a "decision module" uses what the epistemic module knows, together with the utility function, to calculate the value of each possible action (in the CDT or EDT sense).

These agents also lack introspective capabilities of any kind. The epistemic module cannot watch the decision module calculate a utility and get information about the utility function that way. (This blocks things like the tickle defense).

These agents therefore have full access to their own utility functions for the sake of doing the usual decision-theoretic calculations. Nonetheless, they lack knowledge of their own utility function in the epistemic sense. They can only infer what their utility function might be from their actions. This may be difficult, because, as we shall see, EDT agents are sometimes motivated to act in a way which avoids giving information about the utility function.

While I admit this is simply a bad agent design, I don't think it's unrealistic as an extrapolation of current AI systems, *or* particularly bad as a model of (one aspect of) human ignorance about our own values.

More importantly, this is just supposed to be a toy example to illustrate what happens when an agent is epistemically ignorant of an important input to its decision process. It would be nice to have an example which doesn't arise from an obviously bad agent design, but I don't have one.

Smoking Robots

Now suppose that there are two types of robots which have been produced in equal quantities: robots who like smoking, and robots who are indifferent toward smoking. I'll call these "smoke-lovers" and "non-smoke-lovers". Smoke-lovers ascribe smoking +10 utility. Non-smoke-lovers assign smoking -1 due to the expense of obtaining something to smoke. Also, no robot wants to be destroyed; all robots ascribe this -100 utility.

There is a hunter who systematically destroys all smoke-loving robots, whether they choose to smoke or not. We can imagine that the robots have serial numbers, which they cannot remove or obscure. The hunter has a list of smoke-lover serial numbers, and so, can destroy all and only the smoke-lovers. The robots don't have access to the list, so their own serial numbers tell them nothing.

So, the payoffs look like this:

Smoke-lover:

- Smokes:
  - Killed: -90
  - Not killed: +10
- Doesn't smoke:
  - Killed: -100
  - Not killed: 0

Non-smoke-lover:

- Smokes
  - Killed: -101
  - Not killed: -1
- Doesn't smoke:
  - Killed: -100
  - Not killed: 0

All robots know all of this. They just don't know their own utility function (epistemically).

If we suppose that the robots are EDT agents with epsilon-exploration, what happens?

Non-smoke-lovers have no reason to ever smoke, so they'll only smoke with probability epsilon. The smoke lovers are more complicated.

The expected utility for different actions depends on the frequency of those actions in the population of smoke-lovers and non-smoke-lovers, so there's a Nash-equilibrium type solution. It couldn't be that all agents choose not to smoke except epsilon often; then, smoking would provide no evidence, so the smoke-lovers would happily decide to smoke. However, it also can't be that smoke-lovers smoke and non-smoke-lovers don't, because then the conditional probability of being killed given that you smoke would be too high.

The equilibrium will be for smoke-lovers to smoke just a little more frequently than epsilon, in such a way as to equalize the EDT smoke-lover's expected utility for smoking and not smoking. (We can imagine a very small amount of noise in agent's utility calculations to explain how this mixed-strategy equilibrium is actually achieved.)

As with the original smoking lesion problem, this looks like a mistake on the part of EDT. Smoking does not increase a robot's odds of being hunted down and killed. CDT smoke-lovers would choose to smoke.

Furthermore, this isn't changed at all by trying updateless reasoning. There's not really any more-ignorant position for an updateless agent to back off to, at least not one which would be helpful. So, it seems we really need CDT for this one.

What should we think of this?

I think the main question here is how this generalizes to other types of lack of self-knowledge. It's quite plausible that any conclusions from this example depend on the details of my utility-ignorance model, which would mean we can fix things by avoiding utility-ignorance (rather than adopting CDT).

On the other hand, maybe there are less easily avoidable forms of self-ignorance which lead to similar conclusions. Perhaps the argument that CDT outperforms EDT in cases where EDT isn't able to condition on all its inputs can be formalized. If so, it might even provide an argument which is persuasive *to an EDT agent*, which would be really interesting.

Smoking Lesion Steelman II

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

After Johannes Treutlein's comment on Smoking Lesion Steelman, and a number of other considerations, I had almost entirely given up on CDT. However, there were still nagging questions about whether the kind of self-ignorance needed in Smoking Lesion Steelman could arise naturally, how it should be dealt with if so, and what role counterfactuals ought to play in decision theory if CDT-like behavior is incorrect. Today I sat down to collect all the arguments which have been rolling around in my head on this and related issues, and arrived at a place much closer to CDT than I expected.

CDT differs from EDT only in those cases where a parent of the decision node is not directly observed. If the causal parents of a decision are all observations, then there is no probabilistic relationship for causal reasoning to remove. So, there are two critical questions to motivate CDT as an alternative to EDT:

1. (Existence of cases where CDT ≠ EDT) Are there necessarily parents of the decision which cannot be observed?
2. (Correctness of CDT) If so, is CDT the correct way to account for this?

Existence of Cases

Smoking Lesion Steelman was an attempt to set up a scenario answering #1. However, it was very contrived; I had to specifically rob the agents of all introspection and set up an important piece of knowledge about their own decision process which they lacked. Are there more natural examples?

Certainly there will be influences on a real agent's decision process which the agent cannot observe. For example, it is always possible for a cosmic ray to hit a computer and change the output of decision calculation. However, CDT doesn't seem to help with this. The right way to guard against this is vie fault-tolerant implementations which use redundant calculations and memory; but to the extent that a decision procedure might guard against it by having checks in the world-model saying "if things look too goofy I've probably been hit in the head", CDT would potentially sever those probabilistic links, keeping the agent from inferring that it is in an error state. (More on this kind of failure of CDT later.)

Sam Eisenstat has argued (in an in-person discussion) that an example of #1 is the consistency of the logic an agent uses. In the counterexample to the trolljecture, an agent reasons as if if had control over the consistency of its logic, leading to an intuitively wrong decision. Perhaps consistency should be thought of as a parent of the decision, so that CDT-like reasoning would avoid the mistake. Sam also has a probabilistic version of this, to show that it's not just a peculiarity of MUDT. The probabilistic version is not consistent with logical induction, so I'm *still* not sure whether this situation can arise when an agent is build with best practices; but, I can't entirely dismiss it right now, either. (It's also unclear whether the consistency of the logic should be treated as a causal parent of a decision in general; but, in this case, it

does seem like we want CDT-like reasoning: "my action cannot possibly *influence* whether my logic is consistent, even if it provides *evidence* as to whether my logic is consistent".)

Correctness

As for #2, Johannes' comment helped convince me that CDT isn't doing exactly the right thing to account for inputs to the decision process which can't be observed. A discussion with Tom Everitt also helped me see what's going on: he pointed out that CDT fails to make use of all available information when planning. Johannes' example fits that idea, but for another example, suppose that an absent-minded football fan always forgets which team is their favorite. They know that *if* they buy sports memorabilia, they almost always buy their favorite team's gear on impulse. They also know that if they purchased the wrong team's gear, they would hate it; and on the other hand, they love having their favorite team's gear. They find themselves choosing between mugs for several different teams. Should they make a purchase?

EDT says that conditioned on making a purchase, it was very likely to be the right team. Therefore, it is a good idea to make the purchase. CDT, on the other hand, treats the favorite team as an un-observed input to the decision process. Since it is a causal parent, the probabilistic connection gets cut; so, the CDT agent doesn't make the inference that whatever mug they pick is likely to be the right one. Instead, since they're choosing between several mugs, each mug is individually more likely to be the wrong one than the right one. So, CDT doesn't buy a mug.

(The reader may be concerned that this problem is ill-specified for the same reason the standard Smoking Lesion is ill-specified: if the agent is running CDT or EDT, how is it possible that it has a bias toward choosing the right team? We can address this concern with a less extreme version of the trick I used on Smoking Lesion: the decision procedure runs CDT or EDT utility calculations, but then adds epsilon utility to the expected value of any purchase involving the favorite team, which breaks ties in the expected value calculation. This is enough to make the EDT agent reliably purchase the right gear.)

Intuitively, what's going on here is that CDT is refusing to make use of important information in making its decision. EDT goes wrong by confusing evidence for influence; CDT goes wrong by discarding the evidence along with the influence. To the extent that which action we take tells us important information about ourselves which we didn't already know, we want to take this into account when making a decision.

But how can we "take this information into account" without confusing causation and correlation? We can't *both* change our expectation to account for the information *and* hold the probability of causal parents still when making a decision, can we? What sort of decision rule would allow us to get all the examples so far right in a principled way?

An Idea from Correlated Equilibria

I won't be very surprised if the following proposal is already present in the literature. This is a revision of what I proposed in response to Johannes' comment, which he pointed out did not work (which was similar to *Gandalf's Solution to the Newcomb Problem* by Ralph Wedgewood, which Johannes pointed me to). It also bears some resemblance to a post of Scott's.

It seems to me that the problem comes from the fact that when we condition on the *actual* action we will take, we get *good* information, which we *want* to incorporate into our beliefs for the sake of making the right decision. However, when we condition on taking some *other* action, we get *bad* information which we *don't* want to use for making the kind of inference about ourselves which we need in order to get the football-mug example right.

This reminds me of correlated equilibria, which have been discussed quite a bit around MIRI lately. In one way of setting them up, some outside information tells each agent what action to take, with a joint distribution on actions known to all agents. If all agents would accept this advice, then the joint distribution is a stable equilibrium. So, (1) you know which action you will take -- not just what probability you have on different actions, as in Nash equilibria; and (2) a max-expected-value calculation still decides to take that action, after knowing what it is.

This inspired me to try the following decision rule:

$$a_{chosen} = \text{argmax}_a E(do(a)|a_{chosen})$$

IE, the chosen action must be the best causal intervention, after taking into account (as *evidence*) the fact that we chose it. ~~I'll call this CEDT (correlated equilibrium DT) for now.~~ *Scott pointed out that this is a pretty bad name, since many concepts of rationality, including Nash equilibria and logical induction, can be seen as having this property of approving of their output under the condition that they know their output. I therefore re-name it to* ***causal-evidential decision theory*** *(CEDT).*

Note that there may be several consistent fixed-points of the above equation. If so, we have a problem of selecting fixed-points, just like we do with Nash equilibria. Bad choices may be self-fulfilling prophecies.

In the football-mug scenario, every mug becomes an equilibrium solution; taking the mug is evidence that it is the right one, so switching to other mugs looks like a bad idea. Not buying a mug is also an equilibrium. To keep my assumption that the decision is almost always right if the merchandise is purchased, I must suppose that the right equilibrium gets selected (by whatever process does that): the right mug is purchased.

In the cosmic ray example, if some particular action is strong evidence of cosmic-ray-induced malfunction, such that the AI should enter a "safely shut down" mode, then this action cannot be an equilibrium; the AI would switch to shutdown if it observed it. Whether the fact that it isn't an equilibrium actually helps it to happen less given cosmic ray exposure is another question. (The CDT-style reasoning may still be cutting other links which would help detect failure; but more importantly, decision theory still isn't the right way to increase robustness against cosmic rays.)

In the Smoking Lesion Steelman scenario, non-smoke-loving CEDT robots would decline to smoke no matter what action they thought they'd take, since it only costs them. So, the only equilibrium is for them not to smoke. The smoke-loving CEDT robots would smoke if they were told they were going to smoke, or if they were told they wouldn't; so the only equilibrium for them has them doing so.

In Suicidal Smoking Lesion Steelman (Johannes' variant), the non-smoke-lovers again don't smoke. On the other hand, if I've thought this through correctly, smoke-lovers may either smoke or not smoke: if they're told they don't smoke, then they aren't sure

whether they'll be killed or not; so, switching to smoking looks too dangerous. On the other hand, if they're told that they *do* smoke, then they know they *will* die, so they're happy to smoke.

So, in the examples I've looked at, it seems like CEDT always *includes* the desired behavior as a *possible* equilibrium. I'm not sure how to measure the "best" equilibrium and select it. Selecting based on expected value would seem to re-introduse EDT-style mistakes: it might select equilibria where its actions hide some information, so as to increase EV. Another candidate is to select based on the expected benefit, *after* updating on what action is taken, of taking that action rather than alternatives; IE, the gain. This isn't obviously correct either.

This isn't a final solution, by any means, but it is enough to make me willing to say that CDT isn't so far off the mark. As long as a CDT agent knows its own action before it has decided, *and* is given the right causal representation of the problem, then it at least *can* get examples like this right, supposing it happens to be in the right fix-point.

Thanks to Johannes, Tom, Sam, Scott, Patrick, and many participants at this year's AISFP program for discussions leading to this post.

Smoking Lesion Steelman III: Revenge of the Tickle Defense

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

I improve the theory I put forward last time a bit, locate it in the literature, and discuss conditions when this approach unifies CDT and EDT.

Improvements

Where I left off last time, I proposed that decisions should obey the constraint

$$a_{chosen} = \text{argmax}_a E_{P(.|A=a_{chosen})}(U|do(A = a))$$

IE, *evidentially* conditioning on the chosen action, we should still find that if we *causally* evaluate different actions, we make the same choice. (My notation of this continues to be rather awkward, sorry.) This is supposed to combine the advantages of CDT and EDT, since we make maximal use of the information revealed by our choice of action, but avoid confusing evidence with influence in doing so.

Really, I should have instead said that the chosen action must be *among* the maximal choices, so that tie-breaking behavior doesn't matter.

This constraint may be satisfied by several actions; I referred to this as a challenge of "choosing an equilibrium" by analogy to game-theoretic equilibria. I neglected to consider that it may also be satisfied by *no* actions.

Choosing Between Multiple Options

I mentioned a concern that choosing between equilibria using expected value might sometimes select equilibria which hide information in order to look better. This doesn't actually happen in any of the examples I discussed, because being consistent with CDT-style choice prevents the information-hiding actions from being equilibria. However, we can easily make an example where it does happen. In the suicidal smoking lesion problem, we might *additionally* stipulate that the smoke-lovers intrinsically prefer to be non-smoke-lovers, placing -1000 utility on smoke-loving. Then, an equilibrium which leaves a significant chance of not being a smoke-lover looks much better in expected value. So, this method of equilibrium selection would choose the equilibrium where the agent doesn't smoke (and CDT doesn't know enough to switch to smoking, since it looks possible that the agent could smoke and not die).

This makes the solution seem clear to me: we require CDT at a higher level. The error is in choosing equilibria as if the modification of information states actually modified our attributes; but, alas, the self-hating smoke-lover is a smoke-lover whether they smoke or not.

Choosing equilibria via CDT just means selecting an action by CDT, out of those actions which could be equilibria. A question arises as to what information state we choose from. We can again make the argument that we need to choose "using all information" (including an evidential update on which equilibrium is chosen), bringing back the question of equilibrium selection in full force. But, this adds no information. Apparently there is no good answer to the equilibrium-selection problem, except that it is equivalent to the action-selection problem, which we solve by choosing from within some equilibrium.

In any case, this line of thinking made me realize that my "CEDT" is just the ratifiability condition. Ratifiability was first introduced by Jeffrey in *The Logic of Decision*. Skyrms showed in Ratifiability and the Logic of Decision that a reasonable interpretation of ratifiability (perhaps the only reasonable interpretation) makes Jeffrey's EDT select actions which are CDT-optimal when CDT chooses via a probability distribution which has been updated on knowledge of which action is to be selected.

We might imagine, as Skyrms imagined, that the selection of equilibrium proceeds via some kind of convergent process which starts at a knowledge state which is ignorant of the final choice, and ends up in a knowledge state which is in equilibrium. Skyrms raises interesting issues of the Dutch-bookability of such a process.

Choosing in Absence of Equilibria

On the other hand, there are some decision problems which lack ratifiable choices. One example is matching pennies, in the case that we expect the opponent has some skill in predicting us. If we update on taking either action, CDT then wishes to take the opposite one.

Now, I don't mean for my opinion on the EDT vs CDT debate to solve all problems correctly; I still think something like updatelessness is required. But, it does seem like we can naturally extend the spirit of ratifiability to get the right answer in these cases.

Rather than choosing pure actions, we choose mixed strategies, and then pick randomly with the equilibrium probabilities. This guarantees the existence of equilibria, as usual in game theory, since the best-response function is a Kakutani function of the probabilities. This gives the expected solution, choosing randomly with 50-50 odds, in matching pennies.

This might seem to violate the original idea -- aren't we supposed to evidentially condition on the *actual output of our decision process*, ie, the action we take? Aren't we failing to make use of all the information available to us, if we don't? In Regret and Instability in Causal Decision Theory, James Joyce argues in favor of mixed equilibria like this, partly on the basis of making a distinction between what CDT requires you to *believe* vs what CDT requires you to *do*. (He favors a view in which CDT was always intended to yield equilibria like this, and anything else just isn't CDT.)

Conditions for CDT=EDT

The main point in favor of mixed equilibria is that the stronger version requiring full knowledge of actions would be inconsistent, since ratifiable actions don't always exist. However, I think there may be something of deeper importance in Joyce's way of thinking. The "decision", as such, is the formation of the epistemic state about the

action; it is this which is required to be in equilibrium. The *action* occurs after the decision, and may be anything which is consistent with the decision.

To see why this may be important, let's try to put the same self-knowledge condition on EDT. Suppose we are using a version of EDT which epsilon-explores. However, we impose the requirement that EDT's prior knows what policy EDT selects -- IE, it knows what it will do up to the randomness involved in exploration. EDT can still condition on alternate actions, thanks to the epsilon probability remaining for alternate actions. However, knowing its own policy in this way will remove most of the correlations which allow it to make EDT-like decisions. It won't cooperate with a copy of itself in Prisoner's Dilemma if the copy uses different random bits; it knows the policy which the copy will use, and all remaining wiggle room is in randomness, so it no longer sees its action as correlated with the other player's. It will two-box in a version of Newcomb's problem where Omega can't predict the random bits. And so on.

Roughly speaking, we've conditioned on the "output of my decision algorithm" node in the causal graph, which was the hidden parent making CDT different from EDT.

This shouldn't be *too* surprising, given the history of the ratifiability condition. Although I was invoking ratifiability to get CDT to do the right thing in cases where EDT seemed to be doing better, Jeffrey *originally* proposed ratifiability as a corrective measure for making *EDT* act like *CDT* in the cases where they disagreed.

However, this does not always make CDT=EDT. If your exploration bits can be predicted, then EDT and CDT still advise different actions. The "hard mode" of matching pennies is Death in Damascus: Death can predict your every move, so exploration or no, you meet your demise. Under the ratifiable equilibrium, CDT still proposes to flee half the time (as if frantically randomizing in the hope that Death will fail). EDT doesn't do this. In Newcomb with a perfect predictor, CDT two-boxes and EDT one-boxes. And so on. (Note that we still need to make EDT epsilon-explore in these cases, in order to keep the EDT conditional expected utilities well-defined.)

This leads me to propose the following rule for agents obeying ratifiability:

Law of Logical Causality: *If conditioning on any event changes the probability an agent assigns to its own action, that event **must** be treated as causally downstream.*

This is still a bit rough, but constraints on the structure of counterfactuals are hard to come by, so I'll take what I can get.

For example, this means that you must treat Death as causally downstream of your decision in Death in Damascus. You must treat Omega as causally downstream of you in the perfect-predictor version of Newcomb. And so on.

It does *not* mean that you must treat *all* things correlated with your decision as downstream. In Smoking Lesion Steelman, the lesion is still upstream of your decision. The Law of Logical Causality does not touch it, because an agent obeying ratifiability will know its own decision perfectly, screening off the action from such influences. But, this depends on my special version of Smoking Lesion. Perhaps in a different version of Smoking Lesion, where the increased tendency to smoke is more like a trembling hand after the decision is made, the influence is downstream rather than upstream.

Hence the title of this post. Ratifiability, the very tool which CDT advocates used to fend off a series of counterexamples (including cases like Murder Lesion and the suicidal smoking lesion), can be taken by EDT advocates and turned into a very strong

version of the tickle defense. This strong version of the tickle defense makes EDT act as CDT advises in all but the most extreme cases, where predictors are perfect. From this, we propose a constraint on the causal graphs which would make CDT do as EDT would advise.

The whole structure is far from watertight, and is badly in need of formalization. However, it feels like progress to me.

Comparing LICDT and LIEDT

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

Attempted versions of CDT and EDT can be constructed using logical inductors, called LICDT and LIEDT. It is shown, however, that LICDT fails XOR Blackmail, and LIEDT fails Newcomb. One interpretation of this is that LICDT and LIEDT do not implement CDT and EDT very well. I argue that they are indeed forms of CDT and EDT, but stray from expectations because they also implement the ratifiability condition I discussed previously. Continuing the line of thinking from that post, I discuss conditions in which LICDT=LIEDT, and try to draw out broader implications for decision theory.

Thanks to Scott and Sam for discussions shaping this post. Also thanks to many participants at AISFP for discussions shaping my current view of counterfactuals.

I'm not sure who gets credit for LICDT and LIEDT, but they've been discussed around MIRI since shortly after logical induction itself. LIEDT was sort of the obvious first thing to try; LICDT is a slight variation. (Scott thinks Jessica may have come up with LICDT.) They might be thought of as the punching bag for better logical induction DTs to be contrasted with (although, Tsvi wrote up a likely-better baseline proposal).

Both LICDT and LIEDT use a logical inductor which has been run for n steps, P_n. I'll abbreviate an agent as A_n (parameterizing by the same n as for the inductor), with $A_n = a$ to say that the agent takes action a from some action set A. We can define self-referential sentences $S_n \leftrightarrow P_n(S_n) < \epsilon$. Both LICDT and LIEDT explore when such sentences are true. We can take whichever action the agent least expects itself to do conditioned on its exploring. This forces the agent to take every action with frequency at least $\epsilon/|A|$ in the limit, and also makes the exploration pseudorandom in the sense that the logical inductor cannot predict it much better than to assign probability ϵ to exploration (and therefore, neither can any poly-time computable predictor).

When it isn't taking an action due to the exploration clause, LIEDT chooses actions based on the expected utility conditioning on each action. Utility is defined as a logically uncertain variable (LUV), in the terminology of the logical induction paper. Let U_n be the LUV for the utility of A_n, and $E_n(U_n|A_n = a)$ be the conditional expectation of U_n in P_n given that the agent takes action a. The conditional expectation is always well-defined thanks to the exploration, which ensures that the probability of each action remains above zero.

LICDT is similar, but rather than taking the expectation conditioned on each action, it takes the expectation conditioned on *exploring* and taking that action. Judging actions by what would happen if you took those actions randomly, rather than reliably, is

supposed to remove the kind of correlation which makes EDT cooperate in prisoner's dilemma, one-box in Newcomb, et cetera. We will see that this only party works.

Both LICDT and LIEDT include any observations in their deductive state. (There can be special predicates representing sensory states.) So, they are updateful decision theories.

LICDT and LIEDT aren't very different, and mostly we just talk about LIEDT, calling it LIDT. However, I'm recently realizing just how similar LICDT and LIEDT really are.

LIEDT two-boxes in Newcomb.

Suppose we have an LIEDT agent facing Newcomb's problem. We can specify a sequence of Newcomb problems (for logical inductors of increasing power) by the utility function $U_n := 10P_n(A_n = 1) - I(A_n = 1)$, where $A_n = 1$ is the proposition stating that the agent (of power n) one-boxes, and I() is the indicator function which returns 1 for true propositions and 0 for false. This is a Newcomb problem where Omega is fallible; in fact, Omega can only predict the agent as well as the agent can predict itself, since both use the same logical inductor. (And, Newcomb deals with the uncertainty by putting the money in the box with probability equal to its estimation of the probability LIEDT one-boxes.) The best reward the agent can get is if Omega predicts one-boxing, but the agent unexpectedly two-boxes. Of course, a logical inductor can't be fooled like this reliably; so the agent is incentivised to one-box.

Theorem. *LIEDT converges to two-box on non-exploration rounds as* n *increases.*

Proof. The logical inductor comes to predict $E_n(U_n|A_n = 1) = P_n(A_n = 1) - 1$ and $E_n(U_n|A_n = 2) = P_n(A_n = 1)$ with increasing accuracy as n increases, since it has access to $P_n(A_n = 1)$, and since the conditioning is always well-defined thanks to exploration.

Therefore, two-boxing eventually becomes the most appealing option. □

So, LIEDT does not come to see Omega as correlated with its action, because it knows its own general policy, and the general policy screens off all the correlation between Omega and its action.

Now, it's true that if Omega was a more powerful predictor than the agent, LIEDT could one-box -- but, so could LICDT. In particular, if Omega simply knows the action precisely, then $U_n = 10I(A_n = 1) - I(A_n = 1) = 9I(A_n = 1)$, and both LICDT and LIEDT one-box.

LICDT is XOR-blackmailed.

On the other hand, consider the XOR Blackmail letter, which is supposed to be a case where CDT does better than EDT. There is a difficult-to-predict disaster, D_n, with pseudorandom probability 0.01. However, an AI researcher can predict both the disaster and the AI, and will use that knowledge to try and extract money from the AI. Let's call the AI sending money to the researcher $A_n = 1$, and not sending money $A_n = 0$. The AI researcher sends a letter asking for money if and only if [they predict the AI will respond by sending money XOR D_n]. Let's say the AI researcher asks for half the cost of the disaster. $U_n = I(D_n)(-I(A_n) - 2) + I(\neg D_n)(-I(A_n))$. Moreover, the deductive state includes knowledge of the letter, $L_n = \text{XOR}(D_n, A_n)$.

Theorem. *LICDT converges to sending the blackmailer money when a letter is received, on non-exploration rounds.*

Proof. LICDT bases its decision on the utility observed in exploration rounds. Conditional on its receiving the letter and exploring into sending the money, no disaster has occurred. Conditional on its receiving the letter and exploring into not sending money, the disaster has occurred. It will come to predict both of these things accurately, and its conditional expectations will be consistent with them. Therefore, it will send the money. □

Interpretation of the two experiments.

It appears that these aren't very good implementations of CDT and EDT. The attempted CDT fails blackmail letter; the attempted EDT fails Newcomb. But, if we look a little closer, something more interesting is going on. I didn't prove it here, but *both* of them will one-box when Omega is a perfect predictor, and two-box when Omega is fallible. *Both* of them will send the blackmail letter when the blackmailer is a perfect predictor, and refuse when the blackmailer is fallible. They appear to be following my "Law of Logical Causality" from SLS III.

When people argue that EDT one-boxes in Newcomb and that CDT two-boxes, and that EDT sends the money in XOR Blackmail but EDT abstains, they often aren't careful that EDT and CDT are being given the same problem. CDT is supposed to be taking a physical-causation counterfactual, meaning it represents the problem in a Bayesian network representing its physical uncertainty, in which the direction of links lines up with physical causality. If we give EDT the same Bayesian network, it will disregard the causal information contained therein, and compute conditional utilities of actions. But, it is unclear that EDT will then one-box in Newcomb. Reasoning about the physical situation, will it really conclude that conditioning on one action or another changes the expected prediction of Omega? How does the conditional probability flow from the action to Omega? Omega's prediction is based on some observations made in the past. It may be that the agent knows equally well what those observations were; it just doesn't know exactly what Omega concluded from them. The knowledge of the observations screens off any probabilistic relationship. Or, even worse, it may be that the physical information which the agent has includes its own source code. The agent can't try to run its own source code on this very decision; it would go into

an infinite loop. So, we get stuck when we try to do the reasoning. Similar problems occur in XOR blackmail.

I claim that reasoning about what CDT and EDT do in Newcomb's problem and XOR blackmail implicitly assume some solution to logical uncertainty. The correlation which EDT is supposed to conclude exists between its action and the predictor's guess at its action is a logical correlation. But, logical induction doesn't necessarily resolve this kind of logical uncertainty in the intuitive way.

In particular, logical induction implements a version of the ratifiability condition. Because LIEDT agents know their own policies, they are screened off from would-be correlates of their decisions in much the same way LICDT agents are. And because LICDT agents are learning counterfactuals by exploration, they treat predictors who have more information about their own actions than they themselves do as causally downstream -- the law of logical causality which I conjectured would, together with ratifiability, imply CDT=EDT.

When does LICDT=LIEDT?

It's obvious that LIEDT usually equals LICDT when LIEDT converges to taking just one of the actions on non-exploration rounds; the other actions are only taken when exploring, so LIEDT's expectation of those actions just *equals* LICDT's. What about the expectation of the main action? Well, it *may* differ, if there is a reliable difference in utility when it is taken as an exploration vs deliberately chosen. However, this seems to be in some sense unfair; the environment is basing its payoff on the agent's reasons for taking an action, rather than on the action alone. While we'd like to be able to deal with some such environments, allowing this in general allows an environment to punish one decision theory selectively. So, for now, we rule such decision problems out:

Definition: *Decision problem.* A decision problem is a function $U_n(A_n)$ which takes an agent and a step number, and yields a LUV which is the payout.

Definition: *Fair decision problem.* A fair decision problem is a decision problem such that the same limiting action probabilities on the part of the agent imply the same limiting expectations on utility per action, and these expectations do not differ between exploration actions and plain actions. Formally: if $\lim_{n\to\infty} P_n(A_n = a)$ exists for all $a \in A$, and a second agent B_n has limiting action probabilities which also exist and are the same, then $\lim_{n\to\infty} E_n(U_n(A_n)|A_n = a)$ also exist and are the same as the corresponding quantities for B_n; and furthermore, $\lim_{n\to\infty} E_n(U_n(A_n)|A_n = a \wedge S_n)$ exist and are the same as the limits without S_n.

I don't expect that this definition is particularly good; alternatives welcome. In particular, using the inductor itself to estimate the action probability introduces an unfortunate dependence on inductor.

Compare to the notion of fairness in Asymptotic DT.

Definition: A *continuous fair decision problem* is a fair problem for which the function from limiting action probabilities to limiting expected utilities is continuous.

Observation. Continuous fair decision problems have "equilibrium" action distributions, where the best response to the action utilities which come from those is consistent with the action distribution. These are the same whether "best-response" is in the LICDT or LIEDT sense, since the expected utility is required to be the same in both senses. If either LICDT or LIEDT converge on some such problem, then clearly they converge to one of these equilibria.

This doesn't necessarily mean that LICDT and LIEDT converge to the same behavior, though, since it is possible that they fail to converge, and that they converge to different equilibria. I would be somewhat surprised if there is some essential difference in behavior between LICDT and LIEDT on these problems, but I'm not sure what conjecture to state.

I'm more confident in a conjecture for a much narrower notion of fairness (I wasn't able to prove this, but I wouldn't be very surprised if it turned out not to be that hard to prove):

Definition. *Deterministically fair decision problem*: A decision problem is deterministically fair if the payout on instance n are only a function of the action probabilities according to P_n and of the actual action taken (where P_n is the logical inductor used by the agent itself).

Conjecture. Given a continuous deterministically fair decision problem, the set of mixed strategies which LICDT may converge to and LIEDT may converge to, varying the choice of logical inductor, are the same. That is, if LICDT converges to an equilibrium, we can find a logical inductor for which LIEDT converges to that same equilibrium, and vice versa.

This seems likely to be true, since the narrowness of the decision problem class leaves very little room to wedge the two decision theories apart.

The difficulty of proving comparison theorems for LICDT and LIEDT is closely related to the difficulty of proving optimality theorems for them. If we had a good characterization of the convergence and optimality conditions of these two decision theories, we would probably be in a better position to study the relationship between them.

At least we can prove the following fairly boring theorem:

Theorem. For a decision problem in which the utility depends only on the action taken, in a way which does not depend on n or on the agent, and for which all the actions have different utilities, LIEDT and LICDT will converge to the same distribution on actions.

Proof. Epsilon exploration ensures that there continues to be some probability of each action, so the logical inductor will eventually learn the action utilities arbitrarily well. Once the utilities are accurate enough to put the actions in the right ordering, the agent will simply take the best action, with the exception of exploration rounds. □

Law of Logical Counterfactuals

The interesting thing here is that LIEDT seems to be the same as LICDT under an assumption that the environment doesn't specifically mess with it by doing something differently on exploration rounds and non-exploration rounds. This is sort of obvious from the definitions of LICDT and LIEDT (despite the difficulty of actually proving the result). However, notice that it's much different from usual statements of the difference between EDT and CDT.

I claim that LIEDT and LICDT are more-or-less appropriately reflecting the spirit of EDT and CDT under (1) the condition of ratifiability, which is enforced by the self-knowledge properties of logical inductors, and (2) the "law of logical counterfactuals" (LLC) I posited last time, which is enforced by the way LICDT learns causality via experimentation. You can't learn that something which knows more about your action than you do is upstream of you if you make the assumption that you can perform randomized controlled trials!

Since the law of logical counterfactuals was vaguely stated, I had hoped to learn something about what shape it has to take by examining this case. Unfortunately, this case is a bit intractable. However, it did suggest an additional condition: (3) the environment's behavior doesn't depend on whether you explore. (1), (2), and (3) are stated rather informally for now, but together are supposed to imply CDT=EDT, in a formal analysis which is yet-to-be.

Actually, assumption (3) is a sort of "randomized controlled trials" assumption, which seems to justify (2). You assume that you can arrange for actions to be uncorrelated with anything in the world, and that justifies your use of exploration to learn about counterfactuals.

It's not obvious at first, but exploration-based counterfactuals are very similar to counterfactuals based on the chicken rule in proof-based decision theories such as MUDT. The chicken rule requires that if you can prove what your own action will be, you take a different action. This allows the proofs from alternative actions to be well-behaved, rather than exploding in contradiction.

You can see how that's analogous to the way epsilon-exploration makes sure that LICDT and LIEDT can condition on their own actions without dividing by zero. It's also analogous for a deeper reason. Remember the trick of using self-referential sentences for exploration, at the beginning of this writeup? A very similar trick is to take any action which the logical inductor currently assigns probability $< \epsilon$. In other words, do anything you strongly believe you won't do. This is very close to the chicken rule; just substitute probabilistic belief for proof.

In fact, we can go a bit further. One of the shortcomings of MUDT is that it doesn't do what we'd like in the Agent Simulates Predictor scenario, and a host of other problems where a more powerful logic is required. We can address these issues by *giving* it a more powerful logic, but that does not address the intuitive concern, that it seems as if we should be able to solve these problems without *fully* trusting a more powerful logic: if we strongly *suspect* that the more powerful logic is consistent, we should be able to *mostly* do the same thing.

And indeed, we can accomplish this with logical induction. What we do is play chicken against *anything* which seems to predict our action beyond a certain tolerable degree.

This gives us exactly the pseudorandom epsilon-exploration of LICDT/LIEDT. Unfortunately, there's a version of Agent Simulates Predictor which trips these up as well. Alas. (Asymptotic Decision Theory, on the other hand, gets Agent Simulates Predictor right; but, it is bad for other reasons.)

It's interesting that in proof-based decision theory, you never have to take the action; you just threaten to take it. (Or rather: you only take it if your logic is inconsistent.) It's like being able to learn from an experiment which you never perform, merely by virtue of putting yourself in a position where you *almost* do it.

Troll Bridge

Condition (3) is a sort of "no Troll Bridge"condition. The write-ups of the Troll Bridge on this forum are somewhat inadequate references to make my point well (they don't even call it Troll Bridge!), but the basic idea is that you put something in the environment which depends on the consistency of PA, in a way which makes a MUDT agent do the wrong thing via a very curious Löbian argument. There's a version of Troll Bridge which works on the self-referential exploration sentences S_n rather than the consistency of PA. It seems like what's going on in troll bridge has a lot to do with part of the environment correlating itself with your internal machinery; specifically, the internal machinery which ensures that you can have counterfactuals.

Troll Bridge is a counterexample to the idea of proof-length counterfactuals. The hypothesis behind proof-length counterfactuals is that A is a legitimate counterfactual consequence of B if a proof of A from B is much shorter than a proof of ¬A from B. It's interesting that my condition (3) rules it out like this; it suggests a possible relationship between what I'm doing and proof-length counterfactuals. But I suspected such a relationship since before writing SLSIII. The connection is this:

Proof-length counterfactuals are consistent with MUDT, and with variants of MUDT based on bounded-time proof search, because what the chicken rule does is put proofs of what the agent does juuust out of reach of the agent itself. If the environment is tractable enough that there exist proofs of the consequences of actions which the agent will be able to find, the chicken rule pushes the proofs of the agent's own actions far enough out that they won't interfere with that. As a result, you have to essentially step through the whole execution of the agent's code in order to prove what it does, which of course the agent can't do itself while it's running. We can refine proof-length counterfactuals to make an even tighter fit: given a proof search, A is a legitimate counterfactual consequence of B if the search finds a proof of A from B before it finds one for ¬A.

This makes it clear that the notion of counterfactual is *quite* subjective. Logical inductors actually make it significantly less so, because an LI can play chicken against *all* proof systems; it will be more effective in the short-run at playing chicken against its own proof system, but in the song run it learns to predict theorems as fast as any proof system would, so it can play chicken against them all. Nonetheless, LIDT still plays chicken against a subjective notion of predictability.

And this is obviously connected to my assumption (3). LIDT tries valiantly to make (3) true by decorrelating its actions with anything which can predict them. Sadly, as for Agent Simulates Predictor, we can construct a variant of Troll Bridge which causes this to fail anyway.

In any case, this view of what exploration-based counterfactuals are makes me regard them as significantly more natural than I think others do. Nonetheless, the fact remains that neither LIDT nor MUDT do all the things we'd like them to. They don't do what we would like in multi-agent situations. They don't solve Agent Simulates Predictor or Troll Bridge. It seems to me that when they fail, they fail for largely the same reasons, thanks to the strong analogy between their notions of counterfactual.

(There are some disanalogies, however; for one, logical inductor decision theories have a problem of selecting equilibria which doesn't seem to exist in proof-based decision theories.)

Mixed-Strategy Ratifiability Implies CDT=EDT

(Cross-posted from IAFF.)

I provide conditions under which CDT=EDT in Bayes-net causal models.

Previously, I discussed conditions under which LICDT=LIEDT. That case was fairly difficult to analyse, although it looks fairly difficult to get LICDT and LIEDT do differ. It's much easier to analyze the case of CDT and EDT ignoring logical uncertainty.

As I argued in that post, it seems to me that a lot of informal reasoning about the differences between CDT and EDT doesn't actually give the same problem representation to both decision theories. One can easily imagine handing a causal model to CDT and a joint probability distribution to EDT, without checking that the probability distribution could possibly be consistent with the causal model. Representing problems in Bayes nets seems like a good choice for comparing the behavior of CDT and EDT. CDT takes the network to encode causal information, while EDT ignores that and just uses the probability distribution encoded by the network.

It's easy to see that CDT=EDT if all the causal parents of an agent's decision are observed. CDT makes decisions by first cutting the links to parents, and then conditioning on alternative actions. EDT conditions on the alternatives without cutting links. So, EDT differs from CDT insofar as actions provide evidence about causal parents. If all parents are known, then it's not possible for CDT and EDT to differ.

So, any argument for CDT over EDT or vice versa must rely on the possibility of unobserved parents.

The most obvious parents to any decision node are the observations themselves. These are, of course, observed. But, it's possible that there are other significant causal parents which can't be observed so easily. For example, to recover the usual results in the classical thought experiments, it's common to add a node representing "the result of the agent's abstract algorithm node" which is a parent to the agent and any simulations of the agent. This abstract algorithm node captures the correlation which allows EDT to cooperate in the prisoner's dilemma and one-box in Newcomb, for example.

Here, I argue that sufficient introspection still implies that CDT=EDT. Essentially, the agent may not have direct access to all its causal parents, but if it has enough self-knowledge (unlike the setup in Smoking Lesion Steelman), the same screening-off phenomenon occurs. This is somewhat like saying that the output of the abstract algorithm node is known. Under this condition, EDT and CDT both two-box in Newcomb and defect in the prisoner's dilemma.

Mixed-Strategy Ratifiability

Suppose that CDT and EDT agents are given the same decision problem in the form of a Bayesian network. Actions are represented by a variable node in the network, $\boldsymbol{A}$, with values $\boldsymbol{a}$. Agents select mixed strategies somehow, under the constraint that

their choice is maximal with respect to the expectations which they compute for their actions; IE:

1. (*EDT maximization constraint.*) The EDT agent must choose a mixed strategy in which $\boldsymbol{P(a)>\epsilon}$ only if the action is among those which maximize expected utility.
2. (*CDT maximization constraint.*) The CDT agent is under the same restriction, but with respect to the causal expectation.

(*Exploration constraint.*) I further restrict all action probabilities to be at *least* epsilon, to ensure that the conditional expectations are well-defined.

(*Ratifiability constraint.*) I'll also assume ratifiability of mixed strategies: the belief state from which CDT and EDT make their decision is one in which they know which mixed strategy they select. Put another way, the decision is required to be stable under knowledge of the decision. I discuss ratifiability more here.

We can imagine the agent getting this kind of self-knowledge in several ways. Perhaps it knows its own source code and can reason about what it would do in situations like this. Perhaps it knows "how these things go" from experience. Or perhaps the decision rule which picks out the mixed strategies explicitly looks for a choice consistent with mixed-strategy ratifiability.

How this gets represented in the Bayes net is by a node representing the selection of mixed strategy, which I'll call $\boldsymbol{D}$ (the "decision" node) which is the direct parent of $\boldsymbol{A}$ (our action node). $\boldsymbol{D}$ gives the probability of $\boldsymbol{A}$.

(*Mixed-strategy implementability.*) I also assume that AA has no other direct parents, representing the assumption that the choice of mixed strategy is the *only* thing determining the action. This is like the assumption that the environment doesn't contain anything which correlates itself with our random number generator to mess with our experimentation, which I discussed in the LICDT=LIEDT conditions post. It's allowable for things to be correlated with our randomness, but if so, they must be *downstream* of it. Hence, it's also a form of my "law of logical causality" from earlier.

Theorem 1. Under the above assumptions, the consistent choices of mixed strategy are the same for CDT and EDT.

Proof. The CDT and EDT expected utility calculations become the same under the mixed-strategy ratifiability condition, since $\boldsymbol{D}$ screens $\boldsymbol{A}$ off from any un-observed parents of $\boldsymbol{D}$. Besides that, all the rest of the constraints are already the same for CDT and EDT. So, the consistent choices of mixed strategies will be the same. □

It's natural to think of these possible choices as equilibria in the game-theoretic sense. My constraints on the decision procedures for EDT and CDT don't force any particular choice of mixed strategy in cases where several options have maximal utility; but, the condition that that choice must be self-consistent forces it into a few possibilities.

The important observation for my purposes is that this argument for CDT=EDT doesn't require any introspection beyond knowing which mixed strategy you're going to choose in the situation you're in. Perhaps this still seems like a lot to assume. I would contend that it's easier than you may think. As we saw in the logical inductor post, it just seems to happen naturally for LIDT agents. It would also seem to happen for agents who can reason about themselves, or simply know themselves well enough due to experience.

Furthermore, the ratifiability constraint is something which seems necessary to get certain problems right for independent reasons, as has been discussed in the CDT literature. So, if we didn't get it naturally, we would want to build it in.

The way I've defined CDT and EDT may seem a bit unnatural, since I've constrained them based on max-expectation choice of *actions*, but stated that they are choosing *mixed strategies*. Shouldn't I be selecting from the possible probability distributions on actions, based on the expected utility of those? This would invalidate my conclusion, since the CDT expectation of different choices of $\boldsymbol{D}$ can differ from the EDT expectation. But, it is impossible to enforce ratifiability while also ensuring that conditioning on different choices of $\boldsymbol{D}$ is well-defined. So, I think this way of doing it is the natural way when a ratifiability constraint is in play.

Approximate Ratifiability

More concerning, perhaps, is the way my argument takes under-specified decision procedures (only giving constraints under which a decision procedure is fit to be called CDT or EDT) and concludes a thing about what happens in the under-specified cases (effectively, any necessary tie-breaking between actions with equal expected utility must choose action probabilities consistent with the agent's beliefs about the probabilities of its actions). Wouldn't the argument just be invalid if we started with fully-specified versions of CDT and EDT, which already use some particular tie-breaking procedure? Shouldn't we, then, take this as an argument against ratifiability as opposed to an argument for CDT=EDT?

Certainly the conclusion doesn't follow without the assumption of ratifiability. I can address the concern to some extent, however, by making a version of the argument for fixed (but continuous) decision procedures under an approximate ratifiability condition. This will also get rid of the (perhaps annoying) exploration constraint.

(*Continuous EDT*) The EDT agent chooses mixed strategies according to some fixed way which is a continuous function of the belief-state (regarded as a function from worlds to probabilities). This function (the "selection function") is required to agree with maximum-expected-utility choices when the expectations are well-defined and the differences in utilities between options are greater than some $\boldsymbol{\epsilon > 0}$.

(*Continuous CDT*) The same, but taking CDT-style expectations.

(*Approximate Ratifiability*) Let the true mixed strategy which will be chosen by the agent's decision rule be $\boldsymbol{d*}$. For any other $\boldsymbol{d \in D}$ such that $\boldsymbol{|ln(d(a)) - ln(d*(a))| > \epsilon}$ for any $\boldsymbol{a \in A}$, we have $\boldsymbol{P(D=d)=0}$.

(We still assume mixed-strategy implementability, too.)

Approximate ratifiability doesn't perfectly block evidence from flowing backward from the action to the parents of the decision, like perfect ratifiability did. It does bound the amount of evidence, though: since the alternate $\boldsymbol{d}$ must be very close to $\boldsymbol{d*}$, the likelihood ratio cannot be large. Now, as we make epsilon arbitrarily small, there is some delta which bounds the differences in action utilities assigned by CDT and EDT which gets arbitrarily small as well. Hence, the EDT and CDT selection functions must agree on more and more.

By Brouwer's fixed-point theorem, there will be equilibria for the CDT and EDT selection functions. Although there's no guarantee these equilibria are close to each other the way I've spelled things out, we could construct selection functions for both CDT and EDT which get within epsilon of any of the equilibria from theorem 1.

Consequences for Counterfactuals

The arguments above are fairly rudimentary. The point I'm trying to drive at is more radical: there is basically one notion of counterfactual available. It is the one which both CDT and EDT arrive at, if they have very much introspection. It isn't particularly good for the kinds of decision-theory problems we'd like to solve: it tends to two-box in realistic Newcomb's problems (where the predictor is imperfect), defect in prisoner's dilemma, et cetera. My conclusion is that these are not problems to try and solve by counterfactual reasoning. They are problems to solve with updateless reasoning, bargaining, cooperative oracles, predictable exploration, and so on.

I don't think any of this is very new in terms of the arguments between CDT and EDT in the literature. Philosophers seem to have a fairly good understanding of how CDT equals EDT when introspection is possible; see SEP on objections to CDT. The proofs above are just versions of the tickle defense for EDT. However, I think the AI alignment community may not be so aware of the extent to which EDT and CDT coincide. Philosophers continue to distinguish between EDT and CDT, while knowing that they wouldn't differ for ideal introspective agents, on the grounds that decision theories should provide notions of rationality even under failure of introspection. It's worth asking whether advanced AIs may still have some fundamental introspection barriers which lead to different results for CDT and EDT. From where we stand now, looking at positive introspection results over the years, from probabilistic truth to reflective oracles to logical induction, I think the answer is no.

It's possible that a solution to AI alignment will be some kind of tool AI, designed to be highly intelligent in a restricted domain but incapable of thinking about other agent, including itself, on a strategic level. Perhaps there is a useful distinction between CDT and EDT in that case. Yet, such an AI hardly seems to need a decision theory at all, much less the kind of reflective decision theory which MIRI tends to think about.

The meagre reasons in the post above hardly seem to suffice to support this broad view, however. Perhaps my Smoking Lesion Steelman series gives some intuition for it (I, II, III). Perhaps I'll be able to make more of a case as time goes on.

XOR Blackmail & Causality

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

[Cross-posted from IAFF.]

I edited my previous post to note that I'm now much less optimistic about the direction I was going in. This post is to further elaborate the issue and my current position.

Counterfactual reasoning is something we don't understand very well, and which has so many free parameters that it seems to explain just about any solution to a decision problem which one might want to get based on intuition. So, it would be nice to eliminate it from our ontology – to reduce the cases in which it truly captures something important to machinery which we understand, and write off the other cases as "counterfactual-of-the-gaps" in need of some other solution than counterfactuals.

My approach to this involved showing that, in many cases, EDT learns to act like CDT because its knowledge of its own typical behavior screens off the action from the correlations which are generally thought to make EDT cooperate in one-shot prisoner's dilemma with similar agents, one-box in Newcomb's problem, and so on. This is essentially a version of the tickle defense. I also pointed out that the same kind of self-knowledge constraint is needed to deal with some counterexamples to CDT; so, CDT can't be justified as a way of dealing with cases of failure of self-knowledge in general. Instead, CDT seems to improve the situation in some cases of self-knowledge failure, while EDT does better in other such cases.

This suggests a view in which the self-knowledge constraint is a rationality constraint, so the tickle defense is thought of as being true for rational agents, and CDT=EDT under these conditions of rationality. I suggested that problems for which this was not true had to somehow violate the ability of the agent to perform experiments in the world; IE, the decision problem would have to be set up in such a way as to prevent the agent from decorrelating its actions from things in the environment which are not causally downstream of its actions. This seems in some sense unfair, as the environment is preventing the agent from correctly learning the causal relationships through experimentation. I called this condition the law of logical causality when it first occurred to me, and mixed-strategy implementability in the setup where I proved conditions for CDT=EDT.

In XOR Blackmail with a perfect predictor, however, mixed-strategy implementability is violated in a way which does not intuitively seem unfair. As a result, knowledge of what sort of thing you do in XOR blackmail is not sufficient to decorrelate your actions from things which you have no control over. Constraining to the epsilon-exploration case, so that conditional probabilities are well-defined, it seems like what happens is that the epsilon-exploration bit correlates the action you take with the disaster (thanks to the XOR which determines if the letter is sent). On the other hand, it seems as if CDT should be able to get the right answer.

However, I'm unable to come up with a causal Bayes net which seems to faithfully represent the problem, so that I can properly compare how CDT and EDT reason about it in the same representation. It seems like the letter has to be both a parent and a child of the action. I thought I could represent things properly by having a copy of the

action node, representing the simulation of the agent which the predictor uses to predict; but, I don't see how to represent the perfect correlation between the copy and the real action without effectively severing the other parents of the real action.

Anyone have ideas about how to represent XOR Blackmail in a causal network?

Edit:

Here we go. I was confused by the fact that CDT can't reason as if its action makes the letter not get sent. The following causal graph works well enough:

Variables:

- ***A***: the action. True if money is sent to the blackmailer.
- ***A´***: a copy of the action, representing the abstract mathematical fact of what the agent does if it sees the letter.
- ***L***: Whether the letter is sent or not.
- ***D***: The rare disaster.
- ***U***: The utility.

Causal connections:

- ***A***: Has ***A´*** and ***L*** as parents, with the following function: If the letter is sent, copy ***A´***. Otherwise, false.
- ***A´***: No parents.
- ***L***: ***A´*** and ***D*** as parents, with the XOR function determining ***L***.
- ***D***: No parents.
- ***U***: ***A*** and ***D*** as parents, with the utility function as stated in the original XOR post.

Assume epsilon-exploration to ensure that the conditional probabilities are well-defined. Even if EDT knows its own policy, it sees itself as having control over the disaster. CDT, on the other hand, sees no such connection, so it refuses to send the money.

A Rationality Condition for CDT Is That It Equal EDT (Part 1)

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

*[Epistemic Status: this series of two posts gives some arguments which, in my eyes, make it **difficult** to maintain a position other than CDT=EDT, but not impossible. As I explain at the end of the second post, it is still quite tenable to suppose that CDT and EDT end up taking different actions.*]

Previously, I argued that fair comparisons of CDT and EDT (in which the same problem representation is given to both decision theories) will conclude that CDT=EDT, under what I see as reasonable assumptions. Recently, Paul Christiano wrote a post arguing that, all things considered, the evidence strongly favors EDT. Jessica Taylor pointed out that Paul didn't address the problem of conditioning on probability zero events, but she came up with a novel way of addressing that problem by taking the limit of small probabilities: COEDT.

Here, I provide further arguments that rationality constraints point in the direction of COEDT-like solutions.

Note that I argue for the conclusion that CDT=EDT, which is somewhat different from arguing directly for EDT; my line of reasoning suggests some additional structure which could be missed by advocating EDT in isolation (or CDT in isolation). Paul's post described CDT as a very special case of EDT, in which our action is independent of other things we care about. This is true, but, we can also accurately describe EDT is a very special case of CDT where all probabilistic relationships which remain after conditioning on what we know turn out to also be causal relationships. I more often think in the second way, because CDT can have all sorts of counterfactuals based on how causation works. EDT claims that these are only correct when they agree with the conditional probabilities.

(ETA: When I say "CDT", I'm pointing at some kind of steel-man of CDT which uses logical counterfactuals rather than physical counterfactuals. TDT is a CDT in this sense, whereas UDT could be either CDT or EDT.)

This post will be full of conjectural sketches, and mainly serves to convey my intuitions about how COEDT could fit into the larger picture.

Hyperreal Probability

Initially, thinking about COEDT, I was concerned that although something important had been accomplished, the construction via limits didn't seem fundamental enough that it should belong in our basic notion of rationality. Then, I recalled how hyperreal numbers (which can be thought of as sequences of real numbers) are a natural generalization of decision theory. This crops up in several different forms in different areas of Bayesian foundations, but most critically for the current discussion, in the question of how to condition on probability zero events. Quoting an earlier post of mine:

> In What Conditional Probabilities Could Not Be, Alan Hajek argues that conditional probability cannot possibly be defined by Bayes' famous formula, due primarily to its inadequacy when conditioning on events of probability zero. He also takes issue with other proposed definitions, arguing that conditional probability should instead be taken as primitive.
>
> The most popular way of doing this are Popper's axioms of conditional probability. In *Learning the Impossible* (Vann McGee, 1994), it's shown that conditional probability functions following Popper's axioms and nonstandard-real probability functions with conditionals defined according to Bayes' theorem are inter-translatable. Hajek doesn't like the infinitesimal approach because of the resulting non-uniqueness of representation; but, for those who don't see this as a problem but who put some stock in Hajek's other arguments, this would be another point in favor of infinitesimal probability.

In other words, there is an axiomatization of probability -- Popper's axioms -- which takes conditional probability to be fundamental rather than derived. This approach is relatively unknown outside philosophy, but often advocated by philosophers as a superior notion of probability, largely because it allows one to condition on probability zero events. Popper's axioms are in some sense equivalent to allowing hyperreal probabilities, which also means (with a little mathematical hand-waving; I haven't worked this out in detail) we can think of them as a limit of a sequence of strictly nonzero probability distributions.

All of this agrees nicely with Jessica's approach.

I take this to strongly suggest that reasonable approaches to conditioning on probability zero events in EDT will share the limit-like aspect of Jessica's approach, even if it isn't obvious that they do. (Popper's axioms are "limit-like", but this was probably not obvious to Popper.) The major contribution of COEDT beyond this is to provide a *particular* way of constructing such limits.

(Having the idea "counterfactuals should look like conditionals in hyperreal probability distributions" is not enough to solve decision theory problems alone, since it is far from obvious how we should construct hyperreal probability distributions over logic to get reasonable logical counterfactuals.)

Hyperreal Bayes Nets & CDT=EDT

(The following argument is the only justification of the title of the post which will appear in Part 1. I'll have a different argument for the claim in the title in Part 2.)

The CDT=EDT argument can now be adapted to hyperreal structures. My original argument required:

1. **Probabilities & Causal Structure are Compatible:** The decision problem is given as a Bayes net, including an *action* node (for the actual action taken by the agent) and a *decision* node (for the mixed strategy the agent decides on). The CDT agent interprets this as a causal net, whereas the EDT agent ignores the causal information and treats it as a probability distribution.

2. **Exploration:** all action probabilities are bounded away from zero in the decision; that is, the decision node is restricted to mixed strategies in which each action gets

some minimal probability.

3. **Mixed-Strategy Ratifiability:** The agents know the state of the decision node. (This can be relaxed to approximate self-knowledge under some additional assumptions.)

4. **Mixed-Strategy Implementability:** The action node doesn't have any parents other than the decision node.

I justified assumption **#2** as an extension of the desire to give EDT a fair trial: EDT is only clearly-defined in cases with epsilon exploration, so I argued that CDT and EDT should be compared with epsilon-exploration. However, if you prefer CDT *because* EDT isn't well-defined when conditioning on probability zero actions, this isn't much of an argument.

We can now address this by requiring conditionals on probability zero events to be limits of sequences of conditionals in which the event has greater than zero probability. Or (I think equivalently), we think of the probability distribution as being the real part of a hyperreal probability distribution.

Having done this, we can apply the same CDT=EDT result to Bayes nets with hyperreal conditional probability tables. This shows that CDT still equals EDT without restricting to mixed strategies, so long as conditionals on zero-probability actions are defined via limits.

This still leaves the other questionable assumptions behind the CDT=EDT theorem.

#1 (compatible probability & causality): I framed this assumption as the main condition for a fair fight between CDT and EDT: if the causal structure is not compatible with the probability distribution, then you are basically handing different problems to CDT and EDT and then complaining that one gets worse results than the other. However, the case is not so clear as I made it out to be. In cases where CDT/EDT are in specific decision problems which they understand well, the causal structure and probabilistic structure must be compatible. However, boundedly rational agents will have inconsistent beliefs, and it may be that beliefs about causal structure are sometimes inconsistent with other beliefs. An advocate of CDT or EDT might say that the differentiating cases are on exactly such inconsistent examples.

Although I agree that it's important to consider how agents deal with inconsistent beliefs (that's logical uncertainty!), I don't currently think it makes sense to judge them on inconsistent *decision problems.* So, I'll set aside such problems.

Notice, however, that one might contest whether there's necessarily a reasonable causal structure at all, and deny #1 that way.

#3 (ratifiability): The ratifiability assumption is a kind of equilibrium concept; the agent's mixed strategy has to be in equilibrium with knowledge of that very mixed strategy. I argued that it is as much a part of understanding the situation the agent is in as anything else, and that it is usually approximately achievable (IE, doesn't cause terrible self-reference problems or imply logical omniscience). However, I didn't prove that a ratifiable equilibrium always exists! Non-existence would trivialize the result, making it into an argument from false premises to a false conclusion.

Jessica's COEDT results address this concern, showing that this level of self-knowledge is indeed feasible.

#4 (implementability): I think of this as the shakiest assumption; it is easy to set up decision problems which violate it. However, I tend to think such setups get the causal structure wrong. Other parents of the action should instead be thought of as children of the action. Furthermore, if an agent is learning about the structure of a situation by repeated exposure to that situation, implementability seems necessary for the agent to come to understand the situation it is in: parents of the action will *look like children* if you try to perform experiments to see what happens when you do different things.

I won't provide any direct arguments for the implementability constraint in the rest of this post, but I'll be discussing other connections between learning and counterfactual reasoning.

Are We Really Eliminating Exploration?

Ways of Taking Counterfactuals are Somewhat Interchangeable

When thinking about decision theory, we tend to focus on putting the agent in a particular well-defined problem. However, realistically, an agent has a large amount of uncertainty about the structure of the situation it is in. So, a big part of getting things right is learning what situation you're in.

Any reasonable way of defining counterfactuals for actions, be it CDT or COEDT or something else, is going to be able to describe essentially any combination of consequences for the different actions. So, for an agent who doesn't know what situation it is in, any system of counterfactuals is possible no matter how counterfactuals are defined. In some sense, this means that getting counterfactuals right will be mainly up to the learning. Choosing between different kinds of counterfactual reasoning is a bit like choosing different priors -- you would hope it gets washed out by learning.

Exploration is Always Necessary for Learning Guarantees

COEDT eliminates the need for exploration in 5-and-10, which intuitively means *cases where it should be really, really obvious what to do*. It isn't clear to what extent COEDT helps with other issues. I'm skeptical that COEDT alone will allow us to get the right counterfactuals for game-theoretic reasoning. But, it is really clear that COEDT doesn't change the fundamental trade-off between learning guarantees (via exploration) and Bayesian optimality (without exploration).

This is illustrated by the following problem:

Scary Door Problem. *According to your prior, there is some chance that doors of a certain red color conceal monsters who will destroy the universe if disturbed. Your prior holds that this is not very strongly correlated to any facts you could observe without opening such a door. So, there is no way to know whether such doors conceal*

universe-destroying monsters without trying them. If you knew such doors were free of universe-destroying monsters, there are various reasons why you might sometimes want to open them.

The scary door problem illustrates the basic trade-off between asymptotic optimality and subjective optimality. Epsilon exploration would guarantee that you occasionally open scary doors. If such doors conceal monsters, you destroy the universe. However, if you refuse to open scary doors, then it may be that you never learn to perform optimally in the world you're in.

What COEDT does is show that the scary door and 5-and-10 really are different sorts of problem. If there weren't approaches like COEDT which eliminate the need for exploration in 5-and-10, we would be forced to conclude that they're the same: no matter how easy the problem looks, you have to explore in order to learn the right counterfactuals.

So, COEDT shows that not all counterfactual reasoning has to reduce to learning. There are problems you can get right by reasoning alone. You don't always have to explore; you can refuse to open scary doors, while still reliably picking up $10.

I mentioned that choosing between different notions of counterfactual is kind of like choosing between different priors -- you might hope it gets washed out by learning. The scary door problem illustrates why we might not want the learning to be powerful enough to wash out the prior. This means getting the prior right is quite important.

You Still Explore in Logical Time

If you follow the logical time analogy, it seems like you can't ever really construct logical counterfactuals without exploration in some sense: if you reason about a counterfactual, the counterfactual scenario exists somewhere in your logical past, since it is a real mathematical object. Hence, you must take the alternate action sometimes in order to reason about it at all.

So, how does a COEDT agent manage not to explore?

COEDT can be thought of as "learning" from an infinite sequence of agents who explore less and less. None of those agents are COEDT agents, but they get closer and closer. If each of these agents exists at a finite logical time, COEDT exists at an infinite logical time, greater than any of the agents COEDT learns from. So, COEDT doesn't need to explore because COEDT doesn't try to learn from agents maximally similar to itself; it is OK with a systematic difference between itself and the reference class it logically learns from.

This systematic difference may allow us to drive a wedge between the agent and its reference class to demonstrate problematic behavior. I won't try to construct such a case today.

In the COEDT post, Jessica says:

> I consider COEDT to be major progress in decision theory. Before COEDT, there were (as far as I know) 3 different ways to solve 5 and 10, all based on counterfactuals:

• Causal counterfactuals (as in CDT), where counterfactuals are worlds where physical magic happens to force the agent's action to be something specific.

• Model-theoretic counterfactuals (as in modal UDT), where counterfactuals are models in which false statements are true, e.g. where PA is inconsistent.

• Probabilistic conditionals (as in reinforcement learning and logical inductor based decision theories such as LIEDT/LICDT and asymptotic decision theory), where counterfactuals are possible worlds assigned a small but nonzero probability by the agent in which the agent takes a different action through "exploration"; note that ADT-style optimism is a type of exploration.

COEDT is a new way to solve 5 and 10. My best intuitive understanding is that, whereas ordinary EDT (using ordinary reflective oracles) seeks any equilibrium between beliefs and policy, COEDT specifically seeks a not-extremely-unstable equilibrium (though not necessarily one that is stable in the sense of dynamical systems), where the equilibrium is "justified" by the fact that there are arbitrarily close almost-equilibria. This is similar to trembling hand perfect equilibrium. To the extent that COEDT has counterfactuals, they are these worlds where the oracle distribution is not actually reflective but is very close to the actual oracle distribution, and in which the agent takes a suboptimal action with very small probability.

Based on my picture, I think COEDT belongs in the modal UDT class. Both proposals can be seen as a special sort of exploration where we explore if we are in a nonstandard model. Modal UDT explores if PA is inconsistent. COEDT explores if a randomly sampled positive real in the unit interval happens to be less than some nonstandard epsilon. :)

(Note that describing them in this way is a little misleading, since it makes them sound uncomputable. Modal UDT in particular is quite computable, if the decision problem has the right form and if we are happy to assume that PA is consistent.)

I'll be curious to see how well this analogy holds up. Will COEDT have fundamentally new behavior in some sense?

More thoughts to follow in Part 2.

A Rationality Condition for CDT Is That It Equal EDT (Part 2)

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

The previous post sketched an application of Jessica's COEDT framework to get rid of one of the assumptions of my argument for CDT=EDT. Looking at the remaining assumptions of my argument, the hardest one to swallow was implementability: the idea that when the agent implements a mixed strategy, its randomization is successfully controlling for any factors other than those involved in the decision to use that particular mixed strategy. Stated in bayes-net terms, the action has no parents other than the decision.

I stated that the justification of the assumption had to do with the learnability of the causal connections. I then went on to discuss some issues in learning counterfactuals, but not in a way which directly addressed the implementability assumption.

The present post discusses aspects of learning which are more relevant to the implementability assumption. Actually, though, these considerations are only arguments for implementability in that they're arguments for CDT=EDT. None of the arguments here are watertight.

Whereas the previous post was largely a response to COEDT, this post is more exclusively talking about CDT=EDT.

How Should We Learn Counterfactuals?

When Are Counterfactuals Reasonable?

How do we evaluate a proposed way to take logical counterfactuals? One reason progress in this area has been so slow is that it has been difficult to even state *desirable properties* of logical counterfactuals in a mathematically concrete way, aside from strong intuitions about how specific examples should go.

However, there is *one* desirable property which is very clear: ***counterfacting on what's true should result in what's true***. You might get a strange alternative system of mathematics if you ask what things would be like if 2+2=3, but counterfacting on 2+2=4 just gets us mathematics as we know it.

When we consider situations where an agent fully understands what decision problem it's in (so we are assuming that its map corresponds perfectly to the territory), this means that counterfacting on the action which it really takes in that situation, the counterfactual should tell it the true consequences of that action. However, it is less clear how to apply the principle for learning agents.

In the purely Bayesian case, one might argue that the prior is the equivalent of the true circumstance or decision problem; the best an agent can do is to respond as if it were put in situations randomly according to its prior. However, this doesn't give any guarantees about performance in particular situations; in particular, it doesn't give guarantees about learning the right counterfactuals. This becomes a somewhat more pressing concern when we move beyond Bayesian cases to logical induction, since there isn't a prior which can be treated as the true decision problem in the same way.

One might argue, based on the scary door problem from the previous post, that we shouldn't worry about such things. However, I think there are reasonable things we can ask for without going all the way to opening the scary door.

I propose the following principle for learning agents: ***you should not be able to systematically correct your counterfactuals.*** The principle is vaguely stated, but my intention is that it be similar to the logical induction criterion in interpretation: there shouldn't be efficient corrections to counterfactuals, just like there shouldn't be efficient corrections to other beliefs.

If we run into sufficiently similar situations repeatedly, this is like the earlier truth-from-truth principle: the counterfactual predictions for the action actually taken should not keep differing from the consequences experienced. The principle is also not so strong that it implies opening the scary door.

An Example

Consider a very unfair game of matching pennies, in which the other player sees your move before playing. In the setup of Sam's logical induction tiling result, where the counterfactual function can be given arbitrarily, the most direct way to formalize this makes it so that the action takes always gets utility zero, but the *other* action always *would have yielded* utility one, if it had been taken.

The CDT agent doesn't know at the time of making the decision whether the payoff will be 0 for heads and 1 for tails, or 1 for heads and 0 for tails. It can learn, however, that which situation holds depends on which action it ends up selecting. This results in the CDT agent acting pseudorandomly, choosing whichever action it least expects itself to select. The randomization doesn't help, and the agent's expected utility for each action converges to 1/2, which is absurd since it always gets utility 0 in reality. This expectation of 1/2 can be systematically corrected to 0.

An EDT agent who understands the situation would expect utility 0 for each action, since it knows this to be the utility of an action if that action is taken. The CDT agent also knows this to be the conditional expected utility if it actually takes an action; it just doesn't care. Although the EDT agent and CDT agent do equally poorly in this scenario, we can add a third option to not play, giving utility 1/4. The EDT agent will keep going after the illusory 1/2 expected utility.

Sam's theorem assumes that we aren't in a situation like this. I'm not sure if it came off this way in Alex's write-up, but Sam tends to talk about cases like this as unfair environments: the way the payoff for actions depends on the action we actually take makes it impossible to do well from an objective standpoint. I would instead tend to say that this is a bad account of the counterfactuals in the situation. If the other player can see your move in matching pennies, you're just going to get 0 utility no matter how you move.

If we think of this as a hopelessly unfair situation, we accept the poor performance of an agent here. If we think of it as a problem with the counterfactuals, we want a CDT agent to be arranged internally such that it couldn't end up with counterfactuals like these. Asking that counterfactuals not be systematically correctible is a way to do this.

It looks like there is already an argument for EDT along similar lines in the academic literature: *The Gibbard-Harper Collapse Lemma for Counterfactual Decision Theory.*

Can We Dutch Book It?

We can get a CDT agent to bet against the correctible expectations if we give it a side-channel to bet on which doesn't interfere with the decision problem. However, while this may be peculiar, it doesn't really constitute a Dutch Book against the agent.

I suspect a kind of Dutch Book could be created by asking it to bet in the same act as its usual action (so that it is causally conditioning on the action at the time, and therefore choosing under the influence of a correctable expectation). This could be combined with a side-bet made after choosing. I'm not sure of the details here, though.

EDIT: Yes, we can dutch-book it.

A Condition for Reflective Stability of CDT Is That It Equal EDT

Sam's Result

Perhaps a stronger argument for requiring CDT counterfactuals to equal EDT conditionals is the way the assumption seems to turn up in expected-value tiling results. I mentioned earlier that Sam's logical induction tiling result uses a related assumption. Here's Alex Appel's explanation of the assumption:

> The key starting assumption effectively says that the agent will learn that the expected utility of selecting an action in the abstract is the same as the expected utility of selecting the specific action that it actually selects, even if a finite-sized chunk of the past action string is tampered with. Put another way, in the limit, if the agent assigns the best action an expected utility of 0.7, the agent will have its expected utility be 0.7, even if the agent went off-policy for a constant number of steps beforehand.

The part about "even if a finite-sized chunk of the past action string is tampered with" is not so relevant to the present discussion. The important part is that the expected utility of a specific action is the same as that expected without knowing which action will be selected.

One way that this assumption can be satisfied is if the agent learns to accurately predict which action it will select. The expectation of this action equals the expectation in general because the expectation in general already knows this action will be taken.

Another way this assumption can be satisfied is if the counterfactual expectation of each action which the agent might take equals the evidential expectation of that action. In other words, the CDT expectations equal the EDT expectations. This implies the desired condition because the counterfactual expectations of each action which the agent believes it might take end up being the same. (If there were a significant difference between the expectations, the agent could predict the result of its argmax, and therefore would know which action it will take.)

The assumption could also hold in a more exotic way, where the counterfactual expectations and the evidential expectations are not equal individually, but the differences balance each other out. I don't have a strong argument against this possibility, but it does seem a bit odd.

So, I don't have an argument here that CDT=EDT is a necessary condition, but it is a sufficient one. As I touched on earlier, there's a question of whether we should regard this as a property of fair decision problems, or as a design constraint for the agent. The stronger condition of knowing your own action isn't a feasible design constraint; some circumstances make your action unpredictable (such as "I'll give you $100 - [the degree of expectation you had of your action before taking that action]"). We can, however, make the counterfactual expectations equal the evidential ones.

Diff's Tiling Result

I don't want to lean heavily on Sam's tiling result to make the case for a connection between CDT=EDT and tiling, though, because the conclusion of Sam's theorem is that an agent won't envy another agent for its *actions*, but *not* that it won't self-modify into that other agent. It might envy another agent for *counterfactual* actions which are relevant to getting utility. Indeed, counterfactual mugging problems don't violate any of the assumptions of Sam's theorem, and they'll make the agent want to self-modify to be updateless. Sam's framework only lets us examine sequences of actions, and whether the agent would prefer to take a sequence of actions other than its own. It doesn't let us examine whether the agent's own sequence of actions involves taking a screwdriver to its internal machinery.

Diff's tiling result is very preliminary, but it does suggest that the relationship between CDT=EDT and tiling stays relevant when we deal with self-modification. (Diff = Alex Appel) His first assumption states that concrete expected utility of an action takes equals abstract expected utility of taking some action, along very similar lines to the critical assumption in Sam's theorem.

We will have to see how the story plays out with tiling expected utility maximizers.

The Major Caveat

The arguments in this post are all restricted to *moves which the agent continues to take arbitrarily many times as it learns about the situation it is in.* The condition for learning counterfactuals, that they not be systematically correctible, doesn't mean anything for actions which you never take. You can't Dutch Book inconsistent-seeming beliefs which are all conditional on something which never happens. And weird beliefs about actions which you never take don't seem likely to be a big stumbling block for tiling results; the formalizations I used to motivate the connection had conditions having to do with actions actually taken.

This is a major caveat to my line of reasoning, because even if counterfactual expectations and conditional expectations are equal for actions which continue to be taken arbitrarily often, CDT and EDT may end up taking entirely different actions in the limit. For example, in XOR Blackmail, CDT refuses to respond to the letter, and EDT responds. Both expect disaster not to occur after their respective actions, and both are right in their utility forecasts.

We can use Jessica's COEDT to define the conditional beliefs for actions we don't take, but where is the argument that counterfactual beliefs must equal these?

We could argue that CDTs, too, should be limits of CDTs restricted to mixed strategies. This is very much the intuition between trembling-hand Nash equilibria. We then argue that CDTs restricted to mixed strategies should take the same actions as EDTs, by the arguments above, since there are no actions which are never taken. This argument might have some force if a reason for requiring CDT to be the limit of CDTs restricted to mixed strategies were given.

My intuition is that XOR blackmail really shouldn't be counted as a counterexample here. It strikes me as a case correctly resolved by UDT, not CDT. Like counterfactual mugging, counterfactual actions of the agent factor in to the utility received by the agent; or, putting it a different way, a copy of the agent is run with spoofed inputs (in contrast with Newcomb's problem, which runs an exact copy, no spoofing). This means that an agent reasoning about the problem should either reason updatelessly about how it should respond to that sort of situation, or doubt its senses (which can be equivalent).

In other words, my intuition is that there is a class of decision problems for which rational CDT agents converge to EDT in terms of actions taken, not only in terms of the counterfactual expectations of those actions. This class would nicely rule out problems requiring updateless reasoning, and include XOR Blackmail among them.

Dutch-Booking CDT

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

[This post is now superseded by a much better version of the argument.]

In a previous post, I speculated that you might be able to Dutch-Book CDT agents if their counterfactual expectations differed from the conditional expectations of EDT. The answer turns out to be yes.

I'm going to make this a short note rather than being very rigorous about the set of decision problems for which this works.

(This is an edited version of an email, and benefits from correspondence with Caspar Oesterheld, Gerard Roth, and Alex Appel. In particular, Caspar Oesterheld is working on similar ideas. My views on how to interpret the situation have changed since I originally wrote these words, but I'll save that for a future post.)

Suppose a CDT agent has causal expectations which differ from its evidential expectations, in a specific decision.

We can modify the decision by allowing an agent to bet on outcomes in the same act. Because the bet is made simultaneously with the decision, the CDT agent uses causal expected value, and will bet accordingly.

Then, immediately after (before any new observations come in), we offer a new bet about the outcome. The agent will now bet based on its evidential expectations, since the causal intervention has already been made.

For example, take a CDT agent in Death in Damascus. A CDT agent will take each action with 50% probability, and its causal expectations expect to escape death with 50% probability. We can expand the set of possible actions from (stay, run) to (stay, run, stay and make side bet, run and make side bet). The side bet could cost 1 util and pay out 3 utils if the agent doesn't die. Then, immediately after taking the action but before anything else happens, we offer another deal: the agent can get .5 util in exchange for -3 util conditional on not dying. We offer the new bet regardless of whether the agent agrees to the first bet.

The CDT agent will happily make the bet, since the expected utility is calculated along with the intervention. Then, it will happily sell the bet back, because after taking its action, it sees no chance of the 3 util payout.

The CDT agent makes the initial bet even though it knows it will later reverse the transaction at a cost to itself, because we offer the second transaction whether the agent agrees to the first or not. So, from the perspective of the initial decision, taking the bet is still +.5 expected utils. If it could stop itself from later taking the reverse bet, that would be even better, but we suppose that it can't.

I conclude from this that CDT should equal EDT (hence, causality must account for logical correlations, IE include logical causality). By "CDT" I really mean any approach at all to counterfactual reasoning; counterfactual expectations should equal evidential expectations.

As with most of my CDT=EDT arguments, this only provides an argument that the expectations should be equal for actions taken with nonzero probability. In fact, the amount lost to Dutch Book will be proportional to the probability of the action in question. So, differing counterfactual and evidential expectations are smoothly more and more tenable as actions become less and less probable. Actions with very low probability will imply negligible monetary loss. Still, in terms of classical Dutch-Book-ability, CDT is Dutch-Bookable.

Both CDT and EDT have dynamic inconsistencies, but only CDT may be Dutch-booked in this way. I'm not sure how persuasive this should be as an argument -- how special a status should Dutch-book arguments have?

ETA: The formalization of this is now a question.

CDT=EDT=UDT

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

Epistemic status: I no longer endorse the particular direction this post advocates, though I'd be excited if someone figured out something that seems to work. I still endorse most of the specific observations.

So... what's the deal with counterfactuals?

Over the past couple of years, I've been writing about the CDT=EDT perspective. I've now organized those posts into a sequence for easy reading.

I call CDT=EDT a "perspective" because it is a way of consistently answering questions about what counterfactuals are and how they work. At times, I've argued strongly that it is the *correct* way. That's basically because:

- it has been the *only* coherent framework I put any stock in (more for lack of other proposals for dealing with logical counterfactuals than for an abundance of bad ones);
- there *are* strong arguments for it, *if* you're willing to make certain assumptions;
- it would be awfully nice to settle this whole question of counterfactual reasoning and move on. CDT=EDT is in a sense the most boring possible answer, IE that all approaches we've thought of are essentially equivalent and there's no hope for anything better.

However, recently I've realized that there's a perspective which unifies *even more* approaches, while being *less boring* (more optimistic about counterfactual reasoning helping us to do well in decision-theoretic problems). It's been right in front of me the whole time, but I was blind to it due to the way I factored the problem of formulating decision theory. It suggests a research direction for making progress in our understanding of counterfactuals; I'll try to indicate some open curiosities of mine by the end.

Three > Two

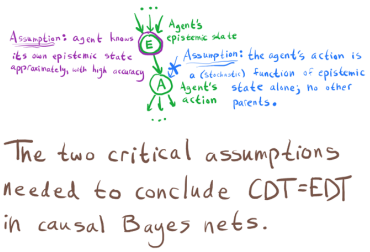
The claim I'll be elaborating on in this post is, essentially, that the framework in Jessica Taylor's post about memoryless cartesian environments is better than the CDT=EDT way of thinking. You'll have to read the post to get the full picture if you haven't, but to briefly summarize: if we formalize decision problems in a framework which Jessica Taylor calls "memoryless cartesian environments" (which we can call "memoryless POMDPs" if we want to be closer to academic CS/ML terminology), reasoning about anthropic uncertainty in a certain way (via the self-indication assumption, SIA for short) makes it possible for CDT to behave like UDT.

The result there is sometimes abbreviated as UDT=CDT+SIA, although UDT⊂CDT+SIA is more accurate, because the optimal UDT policies are a subset of the policies which CDT+SIA can follow. This is because UDT has self-coordination power which CDT+SIA lacks. (We could say UDT=CDT+SIA+coordination, but unfortunately "coordination" lacks a snappy three-letter acronym. Or, to be even more pedantic, we could say that

UDT1.0 = CDT+SIA, and UDT1.1 = CDT+SIA+coordination. (The difference between 1.0 and 1.1 is, after all, the presence of global policy coordination.)) [EDIT: This isn't correct. See Wei Dai's comment.]

Caspar Oesterheld commented on that post with an analogous EDT+SSA result. SSA (the self-sampling assumption) is one of the main contenders beside SIA for correct anthropic reasoning. Caspar's comment shows that we can think of the correct anthropics as a function of your preference between CDT and EDT. So, we could say that CDT+SIA = EDT+SSA = UDT1.0; or, CDT=EDT=UDT for short. [EDIT: As per Wei Dai's comment, the equation "CDT+SIA = EDT+SSA = UDT1.0" is really not correct due to differing coordination strengths; as he put it, UDT1.0 > EDT+SSA > CDT+SIA.]

My CDT=EDT view came from being pedantic about how decision problems are represented, and noticing that when you're pedantic, it becomes awfully hard to drive a wedge between CDT and EDT; you've got to do things which are strange enough that it becomes questionable whether it's a fair comparison between CDT and EDT. However, I didn't notice the extent to which my "being very careful about the representation" was really *insisting that bayes nets are the proper representation*.



*(Aside: Bayes nets which are representing decision problems are usually called **influence diagrams** rather than Bayes nets. I think this convention is silly; why do we need a special term for that?)*

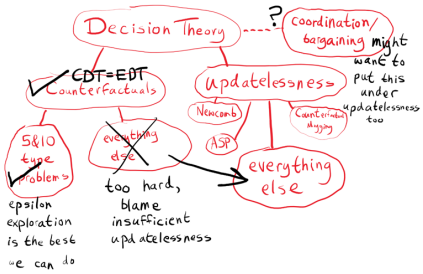
It is rather curious that LIDT also illustrated CDT=EDT-style behavior. It is part of what made me feel like CDT=EDT was a convergent result of many different approaches, rather than noticing its reliance on certain Bayes-net formulations of decision problems. Now, I instead find it to be curious and remarkable that logical induction seems to think as if the world were made of bayes nets.

If CDT=EDT comes from insisting that decision problems are represented as Bayes nets, CDT=EDT=UDT is the view which comes from insisting that decision problems be represented as memoryless cartesian environments. At the moment, this just seems like a better way to be pedantic about representation. It unifies three decision theories instead of two.

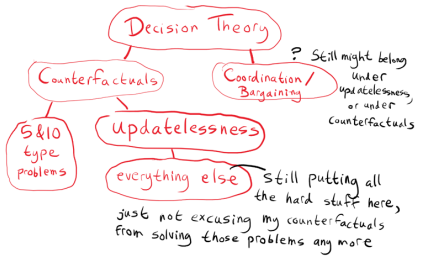
Updatelessness Doesn't Factor Out

In fact, I thought about Jessica's framework frequently, but I didn't think of it as an objection to my CDT=EDT way of thinking. I was blind to this objection because I thought (logical-)counterfactual reasoning and (logically-)updateless reasoning could be dealt with as separate problems. The claim was not that CDT=EDT-style decision-making did well, but rather, that any decision problem where it performed poorly could be analyzed as a case where updateless reasoning is needed in order to do well.

I let my counterfactual reasoning be simple, blaming all the hard problems on the difficulty of logical updatelessness.



Once I thought to question this view, it seemed very likely wrong. The Dutch Book argument for CDT=EDT seems closer to the true justification for CDT=EDT reasoning than the Bayes-net argument, but the Dutch Book argument is a dynamic consistency argument. I know that CDT and EDT both violate dynamic consistency, in general. So, why pick on one special type of dynamic consistency violation which CDT can illustrate but EDT cannot? In other words, the grounds on which I can argue CDT=EDT seem to point more directly to UDT instead.



What about all those arguments for CDT=EDT?

Non-Zero Probability Assumptions

I've noted before that each argument I make for CDT=EDT seems to rely on an assumption that actions have non-zero probability. I leaned heavily on an assumption of epsilon exploration, although one could also argue that all actions must have non-zero probability on different grounds (such as the implausibility of knowing so much about what you are going to do that you can completely rule out any action, before you've made the decision). Focusing on cases where we have to assign probability zero to some action was a big part of finally breaking myself of the CDT=EDT view and moving to the CDT=EDT=UDT view.

(I was almost broken of the view about a year ago by thinking about the XOR blackmail problem, which has features in common with the case I'll consider now; but, it didn't stick, perhaps because the example doesn't actually force actions to have probability zero and so doesn't point so directly to where the arguments break down.)

Consider the transparent Newcomb problem with a perfect predictor:

Transparent Newcomb. *Omega runs a perfect simulation of you, in which you face two boxes, a large box and a small box. Both boxes are made of transparent glass. The small box contains $100, while the large one contains $1,000. In the Simulation, Omega gives you the option of either taking both boxes or only taking the large box. If Omega predicts that you will take only one box, then Omega puts you in this situation for real. Otherwise, Omega gives the real you the same decision, but with the large box empty. You find yourself in front of two full boxes. Do you take one, or two?*

Apparently, since Omega is a perfect predictor, we are forced to assign probability zero to one-boxing even if we follow a policy of epsilon-exploring. In fact, if you implement epsilon-exploration by refusing to take any action which you're very confident you'll take (you have a hard-coded response: if **P("I do action X")>1-epsilon**, do anything but **X**), which is how I often like to think about it, then ***you are forced to 2-box in transparent Newcomb***. I was *expecting* CDT=EDT type reasoning to 2-box (at which point I'd say "but we can fix that by being updateless"), but this is a *really weird reason* to 2-box.

Still, that's not in itself an argument against CDT=EDT. Maybe the rule that we can't take actions we're overconfident in is at fault. The argument against CDT=EDT style counterfactuals in this problem is that the agent should expect that if it 2-boxes, then it won't ever be in the situation to begin with; at least, not in the *real* world. As discussed somewhat in the happy dance problem, this breaks important properties that you might want out of conditioning on conditionals. (There are some interesting consequences of this, but they'll have to wait for a different post.) More importantly for the CDT=EDT question, this can't follow from evidential conditioning, or learning about consequences of actions through epsilon-exploration, or any other principles in the CDT=EDT cluster. So, there would at least have to be other principles in play.

A very natural way of dealing with the problem is to represent the agent's uncertainty about whether it is in a simulation. If you think you might be in Omega's simulation, observing a full box doesn't imply certainty about your own action anymore, or even about whether the box is really full. This is exactly how you deal with the problem in memoryless cartesian environments. But, if we are willing to do this here, we might as well think about things in the memoryless cartesian framework all over the place. This contradicts the CDT=EDT way of thinking about things in lots of problems where updateless reasoning gives different answers than updatefull reasoning, such as counterfactual mugging, rather than only in cases where some action has probability zero.

(I should actually say "problems where updateless reasoning gives different answers than *non-anthropic* updateful reasoning", since the whole point here is that updateful reasoning *can* be consistent with updateless reasoning so long as we take anthropics into account in the right way.)

I also note that trying to represent this problem in bayes nets, while possible, is very awkward and dissatisfying compared to the representation in memoryless cartesian environments. You could say I shouldn't have gotten myself into a position where this felt like significant evidence, but, reliant on Bayes-net thinking as I was, it did.

Ok, so, looking at examples which force actions to have probability zero made me revise my view even for cases where actions all have non-zero probability. So again, it makes sense to ask: but what about the arguments in favor of CDT=EDT?

Bayes Net Structure Assumptions

The argument in the bayes net setting makes some assumptions about the structure of the Bayes net, illustrated earlier. Where do those go wrong?

In the Bayes net setting, observations are represented as parents of the epistemic state (which is a parent of the action). To represent the decision conditional on an observation, we condition on the observation being true. This stops us from putting some probability on our observations being false due to us being in a simulation, as we do in the memoryless cartesian setup.

In other words: the CDT=EDT setup makes it impossible to update on something and still have rational doubt in it, which is what we need to do in order to have an updateful DT act like UDT.

There's likely *some* way to fix this while keeping the Bayes-net formalism. However, memoryless cartesian environments model it naturally.

Question: how can we model memoryless cartesian environments in Bayes nets? Can we do this in a way such that the CDT=EDT theorem applies (making the CDT=EDT way of thinking compatible with the CDT=EDT=UDT way of thinking)?

CDT Dutch Book

What about the Dutch-book argument for CDT=EDT? I'm not quite sure how this one plays out. I need to think more about the setting in which the Dutch-book can be carried out, especially as it relates to anthropic problems and anthropic Dutch-books.

Learning Theory

I said that I think the Dutch-book argument gets closer to the real reason CDT=EDT seems compelling than the Bayes-net picture does. Well, although the Dutch Book argument against CDT gives a crisp justification of a CDT=EDT view, I felt the learning-theoretic intuitions which lead me to formulate the dutch book are closer to the real story. It doesn't make sense to ask an agent to have good counterfactuals in any single situation, because the agent may be ignorant about how to reason about the situation. However, any errors in counterfactual reasoning which result in observed consequences predictably differing from counterfactual expectations should eventually be corrected.

I'm still in the dark about how this argument connects to the CDT=EDT=UDT picture, just as with the Dutch-book argument. I'll discuss this more in the next section.

Static vs Dynamic

A big update in my thinking recently has been to cluster frameworks into "static" and "dynamic", and ask how to translate back and forth between static and dynamic versions of particular ideas. Classical decision theory has a strong tendency to think in terms of statically given decision problems. You could say that the epistemic problem of figuring out what situation you're in is assumed to factor out: decision theory deals

only with what to do once you're in a particular situation. On the other hand, learning theory deals with more "dynamic" notions of rationality: rationality-as-improvement-over-time, rather than an absolute notion of perfect performance. (For our purposes, "time" includes logical time; even in a single-shot game, you can learn from relevantly similar games which play out in thought-experiment form.)

This is a messy distinction. Here are a few choice examples:

Static version: Dutch-book and money-pump arguments.

Dynamic version: Regret bounds.

Dutch-book arguments rely on the idea that you shouldn't *ever* be able to extract money from a rational gambler without a chance of losing it instead. Regret bounds in learning theory offer a more relaxed principle, that you can't ever extract *too much* money (for some notion of "too much" given by the particular regret bound). The more relaxed condition is more broadly applicable; Dutch-book arguments only give us the probabilistic analog of logical consistency properties, whereas regret bounds give us inductive learning.

Static: Probability theory.

Dynamic: Logical induction.

In particular, the logical induction criterion gives a notion of regret which implies a large number of nice properties. Typically, the difference between logical induction and classical probability theory is framed as one of logical omniscience vs logical uncertainty. The static-vs-dynamic frame instead sees the critical difference as one of rationality in a static situation (where it makes sense to think about perfect reasoning) vs learning-theoretic rationality (where it doesn't make sense to ask for perfection, and instead, one thinks in terms of regret bounds).

Static: Bayes-net decision theory (either CDT or EDT as set up in the CDT=EDT argument).

Dynamic: LIDT.

As I mentioned before, the way LIDT seems to naturally reason as if the world were made of Bayes nets now seems like a curious coincidence rather than a convergent consequence of correct counterfactual conditioning. I would like a better explanation of why this happens. Here is my thinking so far:

- Logical induction lacks a way to question its perception. As with the Bayes-net setup used in the CDT=EDT argument, to observe something is to think that thing is true. There is not a natural way for logical induction to reason anthropically, especially for information which comes in through the traders thinking longer. If one of the traders calculates digits of π and bets accordingly, this information is simply known by the logical inductor; how can it entertain the possibility that it's in a simulation and the trader's calculation is being modified by Omega?
- Logical induction knows its own epistemic state to within high accuracy, as is assumed in the Bayes-net CDT=EDT theorem.
- LIDT makes the action a function of the epistemic state alone, as required.

There's a lot of formal work one could do to try to make the connection more rigorous (and look for places where the connection breaks down!).

Static: UDT.

Dynamic: ???

The problem of logical updatelessness has been a thorn in my side for some time now. UDT is a good reply to a lot of decision-theoretic problems when they're framed in a probability-theoretic setting, but moving to a logically uncertain setting, it's unclear how to apply UDT. UDT requires a fixed prior, whereas logical induction gives us a picture in which logical uncertainty is fundamentally about how to revise beliefs as you think longer.

The main reason the static-vs-dynamic idea has been a big update for me is that I realized that a lot of my thinking has been aimed at turning logical uncertainty into a "static" object, to be able to apply UDT. I haven't even posted about most of those ideas, because they haven't lead anywhere interesting. Tsvi's post on thin logical priors is definitely an example, though. I now think this type of approach is likely doomed to failure, because the dynamic perspective is simply superior to the static one.

The interesting question is: how do we translate UDT to a dynamic perspective? How do we learn updateless behavior?

For all its flaws, taking the dynamic perspective on decision theory feels like something asymptotic decision theory got right. I have more to say about what ADT does right and wrong, but perhaps it is too much of an aside for this post.

A general strategy we might take to approach that question is: how do we translate individual things which UDT does right into learning-theoretic desiderata? (This may be more tractable than trying to translate the UDT optimality notion into a learning-theoretic desideratum whole-hog.)

Static: Memoryless Cartesian decision theories (CDT+SIA or EDT+SSA).

Dynamic: ???

The CDT=EDT=UDT perspective on counterfactuals is that we can approach the question of learning logically updateless behavior by thinking about the learning-theoretic version of anthropic reasoning. How do we learn which observations to take seriously? How do we learn about what to expect supposing we *are* being fooled by a simulation? Some optimistic speculation on that is the subject of the next section.

We Have the Data

Part of why I was previously very pessimistic about doing any better than the CDT=EDT-style counterfactuals was that we *don't have any data* about counterfactuals, almost by definition. How are we supposed to learn what to counterfactually expect? We only observe the real world.

Consider LIDT playing transparent Newcomb with a perfect predictor. Its belief that it will 1-box in cases where it sees that the large box is full must converge to 100%,

because it only ever sees a full box in cases where it does indeed 1-box. Furthermore, the expected utility of 2-boxing can be anything, since it will never see cases where it sees a full box and 2-boxes. This means I can make LIDT 1-box by designing my LI to think 2-boxing upon seeing a full box will be catastrophically bad: I simply include a trader with high initial wealth who bets it will be bad. Similarly, I can make LIDT 2-box whenever it sees the full box by including a trader who bets 2-boxing will be great. Then, the LIDT will never see a full box except on rounds where it is going to epsilon-explore into 1-boxing.

(The above analysis depends on details of how epsilon exploration is implemented. If it is implemented via the probabilistic chicken-rule, mentioned earlier, making the agent explore whenever it is very confident about which action it takes, then the situation gets pretty weird. Assume that LIDT is epsilon-exploring pseudorandomly instead.)

LIDT's confidence that it 1-boxes whenever it sees a full box is jarring, because I've just shown that I can make it either 1-box or 2-box depending on the underlying LI. Intuitively, an LIDT agent who 2-boxes upon seeing the full box should not be near-100% confident that it 1-boxes.

The problem is that the cases where LIDT sees a full box and 2-boxes are all counterfactual, since Omega is a perfect predictor and doesn't show us a full box unless we in fact 1-box. LIDT doesn't learn from counterfactual cases; the version of the agent in Omega's head is shut down when Omega is done with it, and never reports its observations back to the main unit.

(The LI *does* correctly learn the *mathematical fact* that its algorithm 2-boxes when input observations of a full box, but, this does not help it to have the intuitively correct expectations when Omega feeds it false sense-data.)

In the terminology of The Happy Dance Problem, LIDT isn't learning the right observation-counterfactuals: the predictions about what action it takes given different possible observations. However, ***we have the data:*** the agent *could* simulate itself under alternative epistemic conditions, and train its observation-counterfactuals on what action it in fact takes in those conditions.

Similarly, the action-counterfactuals are wrong: LIDT can believe anything about what happens when it 2-boxes upon seeing a full box. Again, ***we have the data:*** LI can observe that on rounds when it is mathematically true that the LIDT agent would have 2-boxed upon seeing a full box, it doesn't get the chance. This knowledge simply isn't being "plugged in" to the decision procedure in the right way. Generally speaking, an agent can observe the real consequences of counterfactual actions, because (1) the counterfactual action is a mathematical fact of what the agent does under a counterfactual observation, and (2) the important effects of this counterfactual action occur in the real world, which we can observe directly.

This observation makes me much more optimistic about learning interesting counterfactuals. Previously, it seemed like *by definition* there would be no data from which to learn the correct counterfactuals, other than the (EDTish) requirement that they should match the actual world for actions actually taken. Now, it seems like I have not one, but *two* sources of data: the observation-counterfactuals can be simulated outright, and the action-counterfactuals can be trained on what actually happens when counterfactual actions are taken.

I haven't been able to plug these pieces together to get a working counterfactual-learning algorithm yet. It might be that I'm still missing a component. But ... it *really* feels like there should be something here.

Troll Bridge

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

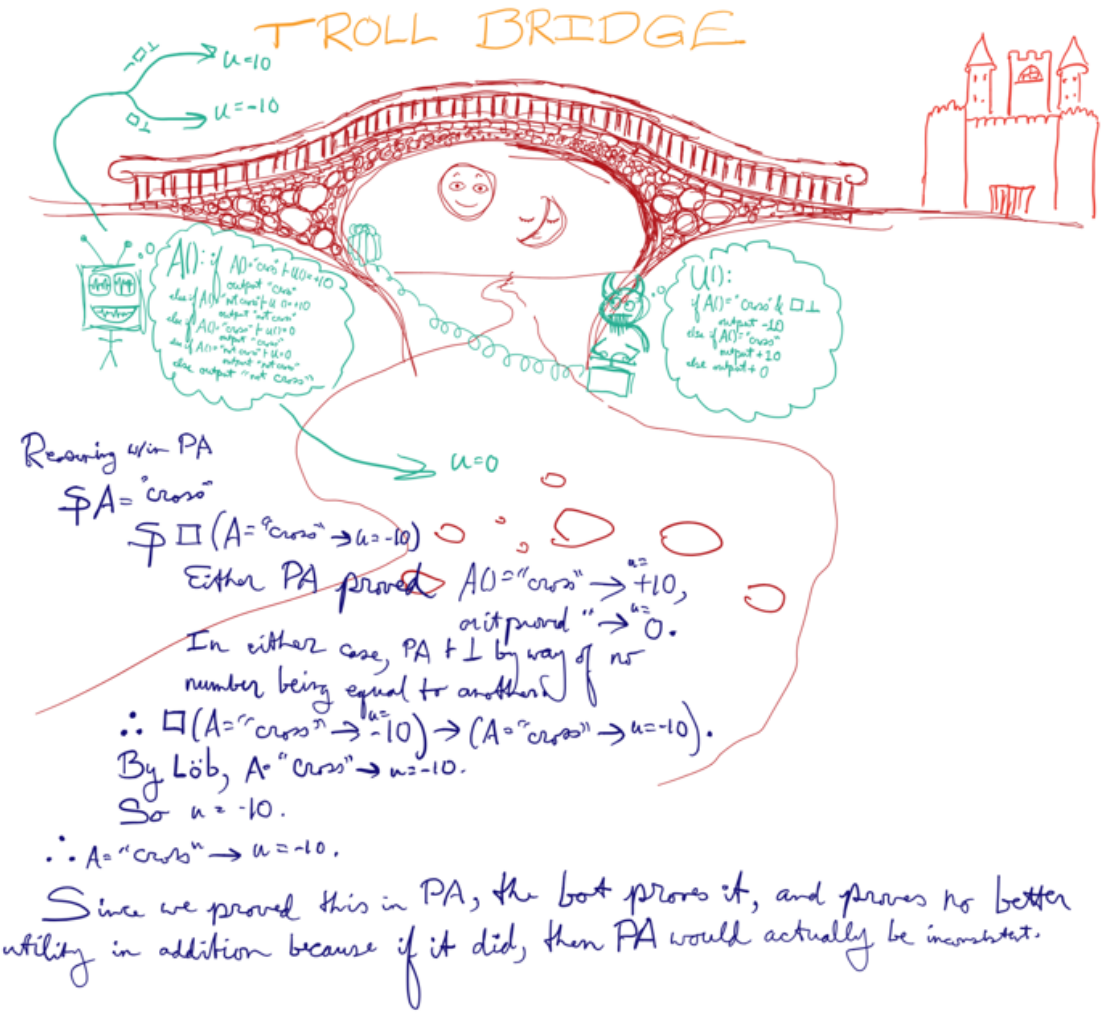
All of the results in this post, and most of the informal observations/interpretations, are due to Sam Eisenstat. I think the Troll Bridge story, as a way to make the decision problem understandable, is due to Tsvi; but I'm not sure.

Pure Logic Version

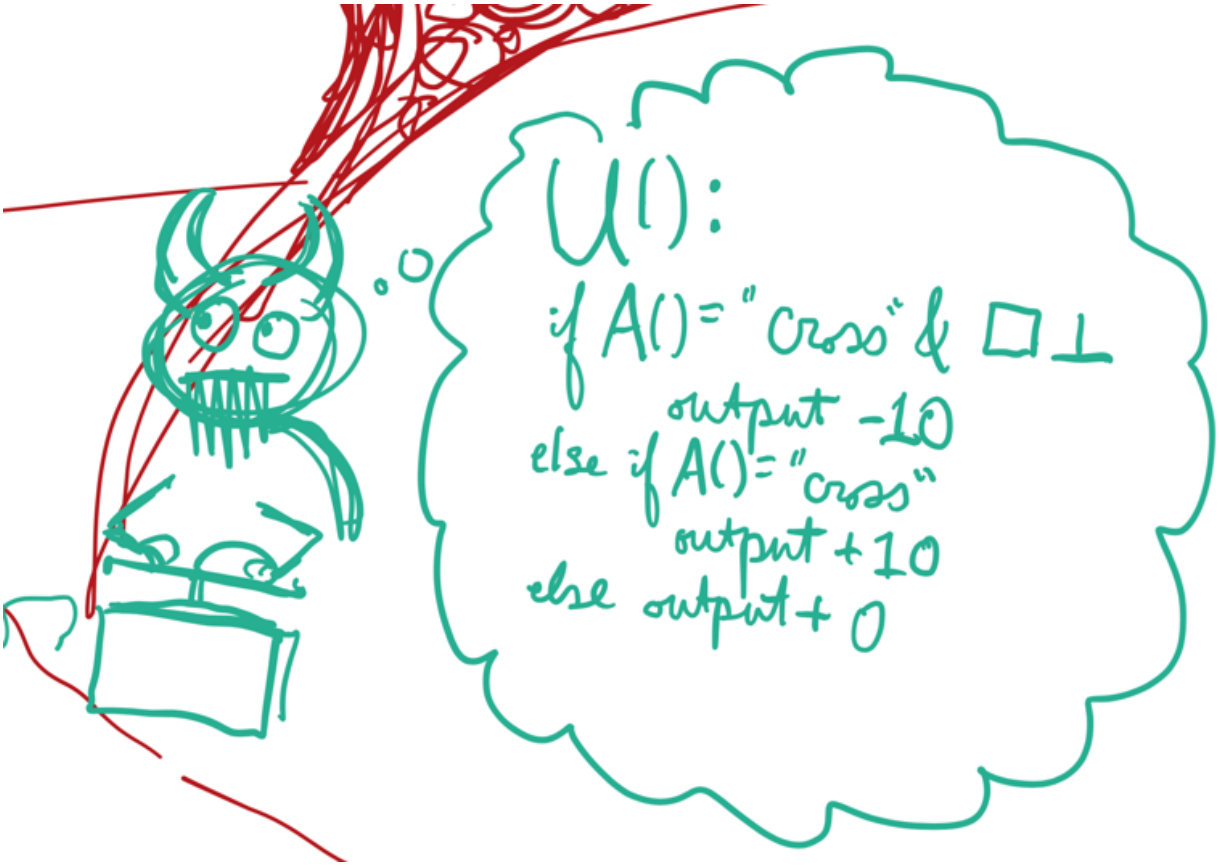
Troll Bridge is a decision problem which has been floating around for a while, but which has lacked a good introductory post. The original post gives the essential example, but it lacks the "troll bridge" story, which (1) makes it hard to understand, since it is just stated in mathematical abstraction, and (2) makes it difficult to find if you search for "troll bridge".

The basic idea is that you want to cross a bridge. However, there is a troll who will blow up the bridge with you on it, ***if*** (and only if) you cross it "for a dumb reason" — for example, due to unsound logic. You can get to where you want to go by a worse path (through the stream). This path is better than being blown up, though.

We apply a Löbian proof to show not only that you choose not to cross, but furthermore, that your counterfactual reasoning is confident that the bridge would have blown up if you had crossed. This is supposed to be a counterexample to various proposed notions of counterfactual, and for various proposed decision theories.

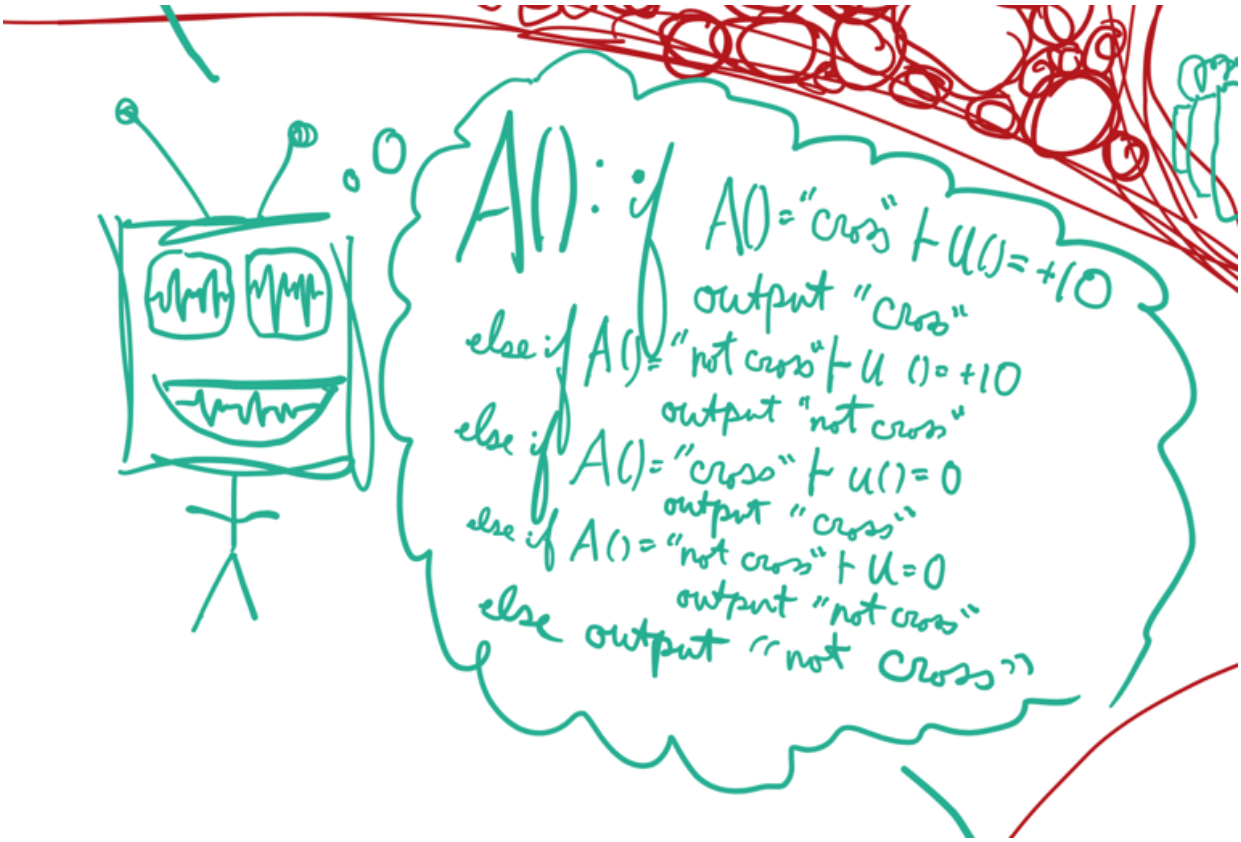


The pseudocode for the environment (more specifically, the utility gained from the environment) is as follows:



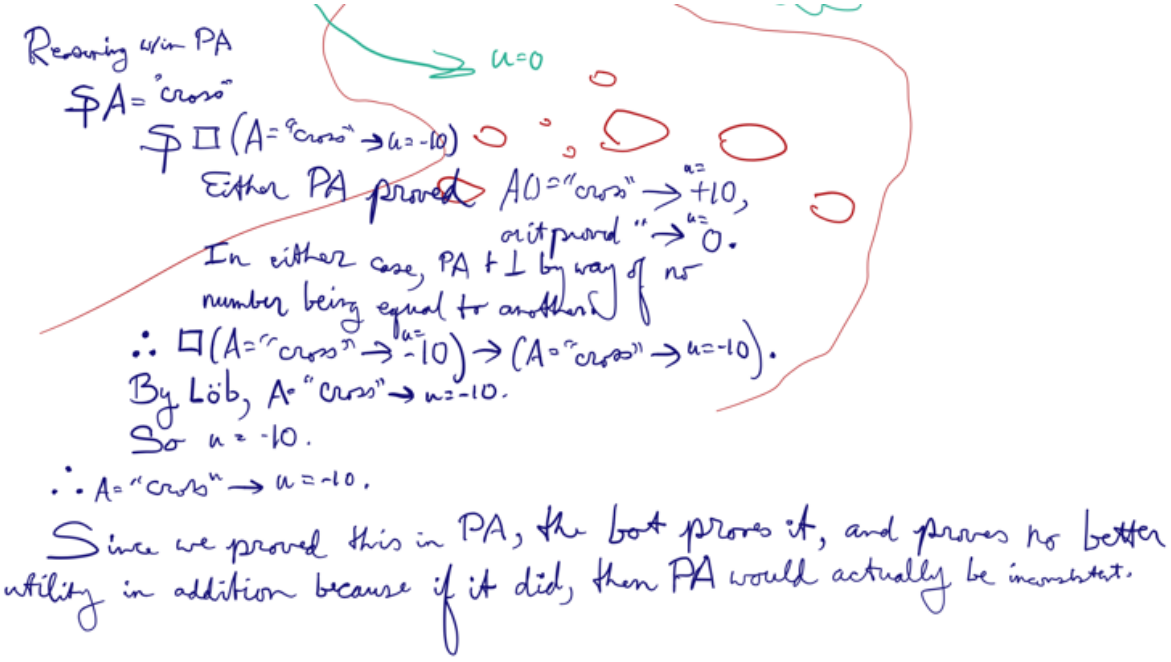
IE, if the agent crosses the bridge and is inconsistent, then U=-10. (□⊥ means "PA proves an inconsistency".) Otherwise, if the agent crosses the bridge, U=+10. If neither of these (IE, the agent does not cross the bridge), U=0.

The pseudocode for the agent could be as follows:



This is a little more complicated, but the idea is supposed to be that you search for every "action implies utility" pair, and take the action for which you can prove the highest utility (with some tie-breaking procedure). Importantly, this is the kind of proof-based decision theory which eliminates spurious counterfactuals in 5-and-10 type problems. It isn't that easy to trip up with Löbian proofs. *(Historical/terminological note: This decision theory was initially called MUDT, and is still sometimes referred to in that way. However, I now often call it proof-based decision theory, because it isn't centrally a UDT. "Modal DT" (MDT) would be reasonable, but the modal operator involved is the "provability" operator, so "proof-based DT" seems more direct.)*

Now, the proof:



- Reasoning within PA (ie, the logic of the agent):
  - Suppose the agent crosses.
    - Further suppose that the agent proves that crossing implies U=-10.
      - Examining the source code of the agent, because we're assuming the agent crosses, either PA proved that crossing implies U=+10, or it proved that crossing implies U=0.
      - So, either way, PA is inconsistent -- by way of 0=-10 or +10=-10.
      - So the troll actually blows up the bridge, and really, U=-10.
    - Therefore (popping out of the second assumption), if the agent proves that crossing implies U=-10, then in fact crossing implies U=-10.
    - By Löb's theorem, crossing really implies U=-10.
    - So (since we're still under the assumption that the agent crosses), U=-10.
  - So (popping out of the assumption that the agent crosses), the agent crossing implies U=-10.
- Since we proved all of this in PA, the agent proves it, and proves no better utility in addition (unless PA is truly inconsistent). On the other hand, it will prove that not crossing gives it a safe U=0. So it will in fact not cross.

The paradoxical aspect of this example is not that the agent doesn't cross -- it makes sense that a proof-based agent can't cross a bridge whose safety is dependent on the agent's own logic being consistent, since proof-based agents can't know whether their logic is consistent. Rather, the point is that the agent's "counterfactual" reasoning looks crazy. *(However, keep reading for a version of the argument where it **does** make the agent take the wrong action.)* Arguably, the agent should be uncertain of what happens if it crosses the bridge, rather than certain that the bridge would blow up. Furthermore, the agent is reasoning as if it can control whether PA is consistent, which is arguably wrong.

In a comment, Stuart points out that this reasoning seems highly dependent on the code of the agent; the "else" clause could be different, and the argument falls apart. I

think the argument keeps its force:

- On the one hand, it's still very concerning if the sensibility of the agent depends greatly on which action it performs in the "else" case.
- On the other hand, we can modify the troll's behavior to match the modified agent. The general rule is that the troll blows up the bridge if the agent would cross for a "dumb reason" -- the agent then concludes that the bridge would be blown up if it crossed. I can no longer complain that the agent reasons as if it were controlling the consistency of PA, but I can still complain that the agent thinks an action is bad because that action indicates its own insanity, due to a troublingly circular argument.

Analogy to Smoking Lesion

One interpretation of this thought-experiment is that it shows proof-based decision theory to be essentially a version of EDT, in that it has EDT-like behavior for Smoking Lesion. The analogy to Smoking Lesion is relatively strong:

- An agent is at risk of having a significant internal issue. (In Smoking Lesion, it’s a medical issue. In Troll Bridge, it is logical inconsistency.)
- The internal issue would bias the agent toward a particular action. (In Smoking Lesion, the agent smokes. In Troll Bridge, an inconsistent agent crosses the bridge.)
- The internal issue also causes some imagined practical problem for the agent. (In Smoking Lesion, the lesion makes one more likely to get cancer. In Troll Bridge, the inconsistency would make the troll blow up the bridge.)
- There is a chain of reasoning which combines these facts to stop the agent from taking the action. (In smoking lesion, EDT refuses to smoke due to the correlation with cancer. In Troll Bridge, the proof-based agent refuses to cross the bridge because of a Löbian proof that crossing the bridge leads to disaster.)
- We intuitively find the conclusion nonsensical. (It seems the EDT agent should smoke; it seems the proof-based agent should not expect the bridge to explode.)

Indeed, the analogy to smoking lesion seems to strengthen the final point -- that the counterfactual reasoning is wrong.

I've come to think of Troll Bridge as "the real smoking lesion", since I'm generally not satisfied with how smoking lesion is set up.

But is proof-based decision theory *really* a version of EDT? I think there's more to say about the analogy, but a simple explanation is this: both EDT and proof-based decision theory evaluate actions by adding them to the knowledge base and seeing what the world looks like under that additional assumption. Or, to put it differently, proof-based DT reasons about actions as if they're observations. That's the fundamental idea of evidential decision theory.

Still, as good as the analogy might be, you still probably shouldn't be convinced until you've seen the Troll Bridge defeat *real* EDT.

Probabilistic Version

For the purely logical version, I said that we can't fault the conclusion (because the agent can't prove that it is safe to cross the bridge) – only the *reasoning* is being

critiqued. However, the probabilistic version better demonstrates the severity of the reasoning error; it illustrates an EDT agent making what seems to be an outright mistake.

It's reasonable to avoid crossing the bridge if you're only thinking about provable results. But we might expect an agent who balances risks and rewards would decide to cross, so long as it believes the probability of PA being inconsistent is low. Thus turns out not to be the case.

Here's the setup.

Suppose that the agent uses a probability distribution which "respects logic" in the sense that it assigns probability zero to anything which is logically refutable. (This assumes logical omniscience. I'll drop that assumption later.)

The agent is as follows: *if P(cross)=0, cross; if P(¬cross)=0, don't cross; else, take the action with highest expected utility, breaking ties by not crossing.* The idea is that we are trying to take the highest-expected-utility action; however, we must handle cases where this isn't well-defined. Choosing to take an action when the probability of the action is zero guarantees that the agent can never prove its own action; it is a version of the chicken rule. Breaking ties by not crossing is arbitrary, but gives the simplest version of the example, because this choice reduces the number of cases we need to check.

The troll is as follows: *if the agent crosses the bridge due to the P(cross)=0 clause, blow up the bridge.*

Reasoning in the logic of the agent:

- Suppose $\Box(A = \text{cross} \rightarrow U = -10)$
    - Suppose A = cross
        - Examining the agent, either crossing had higher expected utility, or P(cross)=0. But we assumed $\Box(A = \text{cross} \rightarrow U = -10)$, so it must be the latter. So the bridge gets blown up.
    - Therefore, $A = \text{cross} \rightarrow U = -10$.
- Therefore, $\Box(A = \text{cross} \rightarrow U = -10) \rightarrow (A = \text{cross} \rightarrow U = -10)$

Thus, by Löb's theorem, we see that $\Box(A = \text{cross} \rightarrow U = -10)$.

Therefore, since the agent is certain of things which are provable, the agent assigns expected value -10 to crossing. Since the expected value of the other action will always be higher than this, and since we know the agent doesn't ever really enter into the P=0 conditions, the agent will choose not to cross. □

Notice that this reasoning did not depend much on the values 10, 0, and -10. The utility of the bridge getting blown up could be -0.000001, and the agent still won't cross. *It isn't weighing the risk; it's decided that the worst outcome is inevitable.* In the case of proof-based agents, I said that the overall decision not to cross was understandable,

since proof-based agents are unable to weigh the risks. A probabilistic agent, however, should intuitively be able to say "I don't know whether the bridge will get blown up, because it involves reasoning about properties of myself which I'm fundamentally uncertain about; but, the odds look decent." But that's not what happens: instead, it is sure that crossing is unfavorable, *no matter what overall probability it assigns to P(A=cross)=0*.

So, in this case we conclude that the Troll Bridge example results in a chicken-rule-based agent taking the wrong action overall. The agent shouldn't be sure that it would cross "for the right reason" (it should assign some probability to P(A=cross)=0, since it can't know that its own logic is consistent). However, intuitively, it should be able to assign some probability to this, and balance the risks. If the downside risk is U=-0.000001, and the probability it assigns to its logic being consistent is not similarly small, it should cross -- and in doing so, it would get +10.

As mentioned for the proof-based agent, the agent's code is a bit arbitrary, and it is worth asking how important the details were. In particular, the default in the case of a tie was to not cross. What if the default in case of a tie were to cross?

We then modify the troll's algorithm to blow up the bridge if and only if P(A=cross)=0 **or** there is a tie. The proof then goes through in the same way.

Perhaps you think that the problem with the above version is that I assumed logical omniscience. It is unrealistic to suppose that agents have beliefs which perfectly respect logic. (Un)Fortunately, the argument doesn't really depend on this; it only requires that the agent respects proofs which it can see, and eventually sees the Löbian proof referenced.

Random Exploration

The frustrating thing about Troll Bridge is that it seems like the agent could just cross the bridge, and things would be fine. The proof that things *wouldn't* be fine *relies on the fact that the agent accepts that very proof as sufficient reason*; so can't we just ignore that kind of proof somehow?

One thing you might try is to consider a learning agent, and force random exploration so the agent just crosses the bridge sometimes. If the agent crosses the bridge, it should be able to see that it's safe, right?

However, we have to ask: what's the appropriate version of Troll Bridge for the exploring agent? Remember I said that the basic idea of Troll Bridge is that the troll blows up the bridge if the agent crosses "for a dumb reason" -- ie, for a reason other than "crossing is a good idea". Random exploration falls into this category.

Furthermore, random exploration is really closely analogous to the chicken rule. Consider:

- (A) The first chicken rule I mentioned: if you prove that you don't take an action, you immediately take it.
- (B) The second chicken rule mentioned: if P(A)=0, then do A.
- (C) Epsilon-chicken: if P(A)<epsilon, then do A.
- (D) Epsilon-exploration: randomly do A with probability epsilon.

(A) is very similar to (B) if the probability distribution respects proofs in the way I mentioned previously; if the probability distribution has the further property of assigning positive probability to anything not disproven, then they're exactly the same. (C) is a natural generalization of (B). (B) tells us what to do when the EDT decision rule isn't well-defined (we can't condition on probability zero events). (C) generalizes this by treating small-probability events in the same way (we might expect our conditional beliefs to be bad for small-probability events, since we presumably don't have much experience with them; so we treat it as if it's the undefined case). (B) ensures there are no zero-probability actions, which means we don't actually ever have to deal with the EDT decision rule being non-well-defined; similarly, (C) ensures no action is almost surely never taken, which helps ensure that our conditional expectations for all actions are empirically grounded.

(C) and (D) are very similar; (C) just supplies pseudorandomness vie the agent's own probabilities. Whenever the agent is too confident about what it does, it does something else. This makes the randomization unpredictable *too the agent*. (D) supplies randomness from an external source instead.

So we have a chain connecting the chicken rule to random exploration.

(Another way of thinking about the similarity: The chicken rule ensures that there are models of the agent's beliefs in which each action is taken, so that the agent can have sensible logical reasoning about each alternative. Epsilon-exploration ensures that the empirical experiences of the agent include all of the actions, so that the agent can have reasonable empirical estimates of consequences of each action.)

OK. So Troll Bridge for an exploring agent should blow up the bridge on exploration, *or* if PA is inconsistent. An agent could look like this:

- With small probability, explore.
    - If exploring, flip a coin to cross or not cross.
- If P(cross)=0, cross.
- If P(¬cross)=0, don't cross.
- Otherwise, take the max-expected-value action, breaking ties randomly.

I'm still assuming that the agent's probability distribution respects proofs, as before. I'm also assuming this agent is playing the game repeatedly, and learning. I also must assume that the agent has found Now, the agent reasons:

- Suppose □(cross → u=-10) for a particular round.
    - Further suppose I crossed on that round.
        - By the first supposition, I knew the payout of crossing to be low; and I must also have known that the payout of not crossing is higher, since I can prove that. Since I can prove what both payouts are, the expected values must equal those, unless PA is inconsistent (in which case P(cross)=0 anyway, since my beliefs respect proofs). So I can only be crossing the bridge for two reasons -- either this is an exploration round, or P(cross)=0.
        - In either case, crossing the bridge yields payout u=-10.
    - Therefore, cross → u=-10 in fact.
- So □(cross → u=-10) → (cross → u=-10).

Since the agent proves that a proof of crossing being bad implies crossing is actually bad, the agent further must prove that crossing is bad in fact, by Löb.

I did this for the logically omniscient case again, but as before, I claim that you can translate the above proof to work in the case that the agent's beliefs respect proofs it can find. That's maybe a bit weird, though, because it involves a Bayesian agent updating on logical proofs; we know this isn't a particularly good way of handling logical uncertainty.

We can use logical induction instead, using an epsilon-exploring version of LIDT. We consider LIDT on a sequence of troll-bridge problems, and show that it eventually notices the Löbian proof and starts refusing to cross. This is even more frustrating than the previous examples, because LIDT might successfully cross for a long time, apparently learning that crossing is safe, and reliably gets +10 payoff. Then, one day, it finds the Löbian proof and stops crossing the bridge!

That case is a little more complicated to work out than the Bayesian probability case, and I omit the proof here.

Non-examples: RL

On the other hand, consider an agent which uses random exploration but doesn't do any logical reasoning, like a typical RL agent. Such an agent doesn't need any chicken rule, since it doesn't care about proofs of what it'll do. It still needs to explore, though. So the troll can blow up the bridge whenever the RL agent crosses due to exploration.

This obviously messes with the RL agent's ability to learn to cross the bridge. The RL agent might never learn to cross, since every time it tries it, it looks bad. So this is sort of similar to Troll Bridge.

However, I think this isn't really the point of Troll Bridge. The key difference is this: the RL agent can get past the bridge if its prior expectation that crossing is a good idea is high enough. It just starts out crossing, and happily crosses all the time.

Troll Bridge is about the inevitable confidence that crossing the bridge is bad. We would be fine if an agent decided not to cross because it assigned high probability to PA being inconsistent. The RL example seems similar in that it depends on the agent's prior.

We could try to alter the example to get that kind of inevitability. Maybe we argue it's still "dumb" to cross only because you start with a high prior probability of it being good. Have the troll punish crossing *unless the crossing is justified by an empirical history of crossing being good*. Then RL agents do poorly no matter what -- no one can get the good outcome in order to build up the history, since getting the good outcome *requires* the history.

But this still doesn't seem so interesting. You're just messing with these agents. It isn't illustrating the degree of pathological reasoning which the Löbian proof illustrates -- **of course** you don't put your hand in the fire if you get burned every single time you try it. There's nothing wrong with the way the RL agent is reacting!

So, Troll Bridge seems to be more exclusively about agents who do reason logically.

Conclusions

All of the examples have depended on a version of the chicken rule. This leaves us with a fascinating catch-22:

- We need the chicken rule to avoid spurious proofs. As a reminder: spurious proofs are cases where an agent would reject an action if it could prove that it would not take that action. These actions can then be rejected by an application of Löb's theorem. The chicken rule avoids this problem by ensuring that agents cannot know their own actions, since if they did then they'd take a different action from the one they know they'll take (and they know this, conditional on their logic being consistent).
- However, Troll Bridge shows that the chicken rule can lead to another kind of problematic Löbian proof.

So, we might take Troll Bridge to show that the chicken rule does not achieve its goal, and therefore reject the chicken rule. However, this conclusion is very severe. We cannot simply drop the chicken rule and open the gates to the (much more common!) spurious proofs. We would need an altogether different way of rejecting the spurious proofs; perhaps a full account of logical counterfactuals.

Furthermore, it is possible to come up with variants of Troll Bridge which counter some such proposals. In particular, Troll Bridge was originally invented to counter proof-length counterfactuals, which essentially generalize chicken rules, and therefore lead to the same Troll Bridge problems).

Another possible conclusion could be that Troll Bridge is simply too hard, and we need to accept that agents will be vulnerable to this kind of reasoning.

Dutch-Booking CDT: Revised Argument

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

This post has benefited greatly from discussion with Sam Eisenstat, Caspar Oesterheld, and Daniel Kokotajlo.

Last year, I wrote a post claiming there was a Dutch Book against CDTs whose counterfactual expectations differ from EDT. However, the argument was a bit fuzzy.

I recently came up with a variation on the argument which gets around some problems; I present this more rigorous version here.

Here, "CDT" refers -- very broadly -- to using counterfactuals to evaluate expected value of actions. It need not mean physical-causal counterfactuals. In particular, TDT counts as "a CDT" in this sense.

"EDT", on the other hand, refers to the use of conditional probability to evaluate expected value of actions.

Put more mathematically, for action $a \in A$, EDT uses $E(U|Act = a)$, and CDT uses

$E(U|do(Act = a))$. I'll write $edt(a)$ and $cdt(a)$ to keep things short.

My argument could be viewed as using Dutch Books to formalize Paul Christiano's "simple argument" for EDT:

> Suppose I am faced with two options, call them L and R. From my perspective, there are two possible outcomes of my decision process. Either I pick L, in which case I expect the distribution over outcomes P(outcome|I pick L), or I pick R, in which case I expect the distribution over outcomes P(outcome|I pick R). In picking between L and R I am picking between these two distributions over outcomes, so I should pick the action A for which E[utility|I pick A] is largest. There is no case in which I expect to obtain the distribution of outcomes under causal intervention P(outcome|do(I pick L)), so there is no particular reason that this distribution should enter into my decision process.

However, I do not currently view the argument as favoring EDT over CDT! Instead it supports the weaker claim that the two had better agree. Indeed, the Troll Bridge problem strongly favors *CDT whose expectations agree with EDT* over EDT. So, this is intended to provide a strong constraint on a theory of (logical) counterfactuals, not necessarily abolish the need for them. (However, the constraint is a strong one, and it's worth considering the possibility that this constraint is *all we need* for a theory of counterfactuals.)

The Basic Argument

Consider any one action $a \in A$ for which $edt(a) \neq cdt(a)$, in some decision problem. We wish to construct a modified decision problem which Dutch-books the CDT.

My argument requires an assumption that the action a is assigned nonzero probability.

This is required to ensure edt(a) is defined at all (since otherwise we would be conditioning on a probability zero event), but also for other reasons, which we'll see later on.

Anyway, as I was saying, we wish to take the decision problem which produces the disagreement between edt(a) and cdt(a), and from it, produce a new decision problem which is a Dutch book.

The new decision problem will be a two-step sequential decision problem. *Immediately before* the original decision, the bookie offers to sell the agent the following bet B, for a price of 2d utilons. B is a bet conditional on a, in which the buyer is betting against cdt(a)'s expectation and in favor of edt(a)'s expectation. For example:

B: *In the case that* $A = a$*, the seller of this certificate owes the purchaser of this certificate* $(\mathrm{cdt}(a) - U) \cdot s$*, where* s *is the signum* $\frac{|\mathrm{cdt}(a) - \mathrm{edt}(a)|}{\mathrm{cdt}(a) - \mathrm{edt}(a)}$

The key point here is that *because the agent is betting ahead of time*, it will evaluate the value of this bet according to the conditional expectation $E(U|\mathrm{Act} = a)$.

If $\mathrm{cdt}(a) > \mathrm{edt}(a)$, so $s = 1$, then the value of B in the case that $\mathrm{Act} = a$ is $\mathrm{cdt}(a) - U$.

The expectation of this is $\mathrm{cdt}(a) - \mathrm{edt}(a)$, which again, we have supposed is positive. So the overall expectation is $P(\mathrm{Act} = a) \cdot (\mathrm{cdt}(a) - \mathrm{edt}(a))$. Setting d low enough ensures that the agent will be happy to take this bet. Similarly, if $\mathrm{edt}(a) > \mathrm{cdt}(a)$, the value of the bet ends up being $P(\mathrm{Act} = a) \cdot (\mathrm{edt}(a) - \mathrm{cdt}(a))$ and the agent still takes it for the right price.

Now, the second stage of our argument. *As the agent is making the decision* the bookie again makes an offer. (In other words, we extend the original set of actions A to contain twice as many valid actions; half in which we accept, half in which we don't accept.) The new offer is this: "I will buy the bet B from you for d utilons."

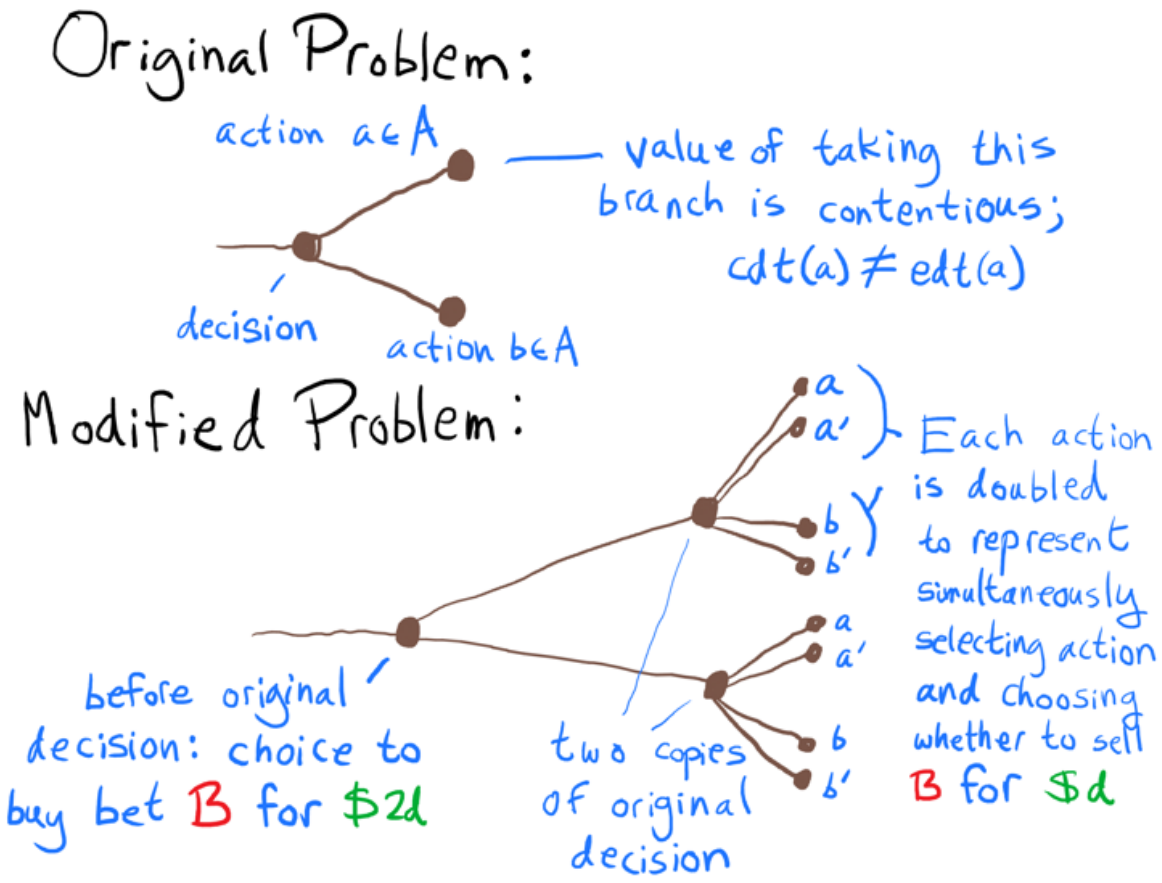
Now, since the agent is reasoning *during* its action, it is evaluating possible actions *according to* cdt(a)*;* so its evaluation of the bet will be different. Here, the argument splits into two cases:

- When considering the action a, the bet's expected value is cdt(a) − cdt(a), which is zero. So the agent prefers the new action a' which is like a except B is sold back to the bookie for d utilons.
- When considering any other action, the bet is worth zero *automatically*, since it only pays out anything when Act = a. So, the agent will gladly take the bookie's payment of d to sell the bet back.

So the result is the same in either case -- CDT recommends selling B back to the bookie no matter what.

The agent has paid 2d to buy B, and gotten only b when selling back. Buying and selling the contract cancel each other out. So the agent is down d utilons for no gain!

Here is an illustration of the entire Dutch Book:



Assumptions

Non-Zero Probability of a

A really significant assumption of this argument is that actions are given nonzero probability -- particularly, that the target action a has a nonzero probability. This assumption is important, since the initial evaluation of B is $P(\text{Act} = a) \cdot |\text{cdt}(a) - \text{edt}(a)|$.

If the probability of action a were zero, there would be no price the agent would be willing to pay for the bet.

The assumption is also required in order to guarantee that edt(a) is well-defined -- although we could possibly use tricks to get around that, specifying a variant of EDT which defines some expectation in all cases.

Many of my arguments for CDT=EDT rest on this assumption, though, so it isn't anything new. It seems to be a truly important requirement, rather than an artefact of the argument.

There are many justifications of the assumption which one might try to give. I have often invoked epsilon-exploration; that is, the idea that some randomness needs to be injected into an agents actions in order to ensure that it can try all options. I don't like invoking that as much as I used to. I might make the weaker argument that agents should use the chicken rule, IE, refuse to take any action which they can prove they take. (This can be understood as weaker than epsilon-exploration, because epsilon-exploration can be implemented by the epsilon-chicken rule: take any action which you assign probability less than epsilon to.) This rule ensures that agents can never prove what they do (so long as they use a sound logic). We can then invoke the non-dogmatism principle, which says that we should never assign probability 0 to a possibility unless we've logically refuted it.

Or, we could invoke a free-will principle, claiming that agents should have the subjective illusion of freedom.

In the end, though, what we have is an argument that applies if and only if a has nonzero probability. All the rest is just speculation about how broadly this argument can be applied.

An interesting feature of the argument is that *the less probable action* a *according to the agent, the less money we can get by Dutch-booking them on discrepancies between* cdt(a) *and* edt(a). This doesn't matter for traditional Dutch Book arguments -- any sure loss is considered a failure of rationality. However, if we take a logical-induction type approach to rationality, smaller Dutch Books *are less important* -- boundedly rational agents are expected to lose *some* money to Dutch Books, and are only trying to avoid losing too much.

So, one might consider this to be a sign that, in some hypothetical bounded-rationality approach to decision theory, lower-probability actions would be allowed to maintain larger discrepancies between edt(a) and cdt(a), and maintain them for longer.

Probabilities of Actions in the Modified Problem

A trickier point is the way probabilities carry over from the original decision problem to the modified problem. In particular, I assume the underlying action probabilities do not change. Yet, I split each action in two!

One justification of this might be that, for agents who choose according to CDT, it *shouldn't* change anything -- at the moment of decision, the bet B is worth nothing, so it doesn't bias actions in one direction or another.

Ultimately, though, I think this is just part of the problem setup. Much like money-pump arguments posit magical genies who can switch anything for anything else, I'm positing a bookie who can offer these bets without changing anything. The argument -- if you choose to accept it -- is that the result is disturbing in any case. It does not seem likely that an appealing theory of counterfactuals is going to wriggle out of this specifically by denying the premise that action probabilities remain the same.

Note, however, that it *is not* important to my argument that none of the *new* actions get assigned zero probability. It's only important that the sum of $P(a)$ and $P(a')$ in the new problem equals the original decision problem's $P(a)$.

Counterfactual Evaluation of Bets

Another assumption I didn't spell out yet is the interaction of bet contracts with *counterfactual* evaluations.

I assume that counterfacting on accepting the bet does not change probabilities of other things, such as the probability of the actions. This could be a large concern in general -- taking a conditional bet on Act = a might make us want to choose a on purpose, in order to cash in on the bet. This isn't a problem in this case, since the agent later evaluates the bet to be worth nothing. However, that doesn't necessarily mean it's not an issue *according to the counterfactual evaluation*, which would chance the perceived value of B. Or, even more problematic, the agent's counterfactual expectations might say that taking the bet would result in some very negative event -- making the agent simply refuse. So the argument definitely assumes "reasonable counterfactual evaluations" in some sense.

On the other hand, this kind of reasoning is very typical for Dutch Book arguments. The bets are grafted onto the situation without touching any of the underlying probabilities -- so, e.g., you do not normally ask "is accepting the bet against X going to make X more probable?".

Handling Some Possible Objections

Does the Bookie Cheat?

You might look at my assumptions and be concerned that the bookie is cheating by using knowledge which the agent does not have. If a bookie has insider information,

and uses *that* to get a sure profit, it doesn't count as a Dutch Book! For example, if a bookie knows all logical facts, it can money-pump any agent who does not know all logical facts (ie, money-pump any fixed computable probability distribution). But that isn't fair.

In this case, one might be concerned about *the agent not knowing its own action*. Perhaps I'm sneaking in an assumption that agents are uncertain of their own actions, and then Dutch-booking them by taking advantage of that fact, via a bookie who can easily predict the agent's action.

To this I have a couple of responses.

The bookie does not know the agent's choice of action. The bookie's strategy doesn't depend on this. In particular, note the disjunctive form of the argument: *either* the agent prefers a, in which case B is worthless for one reason, *or* the agent prefers a different action, in which case B is worthless for a different reason. The bookie is setting things up so that it's safe no matter what.

The agent knows everything the bookie knows, from the beginning. All the bookie needs in order to implement its strategy is the the values of cdt(a) and edt(a), and, in order to set the price d, the probability of the action a.These are things which the agent also knows.

The Agent Justifiably Revises Its Position

Another critique I have received is that it makes perfect sense that the agent takes the bet at the first choice point and later decides against it at the second choice point. The *agent* has gained information -- namely, when considering an action, the agent knows it will take that action. This extra information is being used to reject the bet. So it's perfectly reasonable.

Again I have a couple of responses.

The agent does not learn anything between the two steps of the game. There is no new observation, or additional information of any kind, between the step when the bookie offers B and the step when the bookie offers to buy B back. As the agent is evaluating a particular action, it *does not* "know" it will carry out that action -- it is only considering what would happen *if* it carried out that action!

Even if the agent did learn something, it would not justify being Dutch-booked. Consider two-stage games in which an agent is offered a bet, then learns some information, and then given a choice to sell the bet back for a fee. It is perfectly reasonable for an agent to *in some cases* sell the bet back. What makes a Dutch book, however, is if the agent *always sells the bet back*. It should never be the case that an agent *predictably* won't want the bet later, no matter what it observes. If that were the case (as it is in my scenario), the agent should not have accepted the bet in the first place. It's critical here to again note that the agent prefers to sell back the bet *for every possible action* -- that is, the original actions are *always* judged worse than their modified copies in which the sell-back deal is taken. So, even if we think of the agent

as "learning" which action it selects when it evaluates selecting an action, we can see that it decides to sell back the bet no matter what it learns.

But Agents *WILL* Know What They'll Do

One might say that the argument doesn't mean very much at all in practice because *from* the information it knows, the agent *should* be able to derive its action. It knows how it evaluates all the actions, so, it should just know that it takes the argmax. This means the probability of the action actually taken is 1, and the probability of the rest of the actions is zero. As a result, my argument would only apply to the argument actually taken -- and a CDT advocate can easily concede that cdt(a) = edt(a) *when* a *is the action actually taken*. It's other actions that one might disagree about. For example, in Newcomb, classical physical-causality CDT two-boxes, and agrees with EDT about the consequences of two-boxing. The disagreement is only about the value of the *other* action.

(Note, however, that the CDT advocate is still making a significant concession here; in particular, this rules out the classic CDT behavior in Death and Damascus and many variations on that problem. I don't know exactly how a classic physical-causality CDT advocate would maintain such a position.)

There are all kinds of problems with the agent knowing its own action, but a CDT advocate can very naturally reply that these should be solved with the right counterfactuals, not by ensuring that the agent is unsure of its actions (through, e.g., epsilon-exploration).

I'll have more to say about this objection later, but for now, a couple of remarks.

First and foremost, yeah, my argument doesn't apply to actions which the agent knows it won't take. I think the best view of the phenomenon here is that, if the agent *really does* know exactly what it will do, then yeah, the argument *really does* collapse to saying its evidential expectations should equal its counterfactual expectations for that one action. Which is like saying that, if $P(A) = 1$, then we had better have $P(X|do(A)) = P(X)$ -- counterfacting on something true should never change anything.

Certainly it's quite common to think of agents as knowing exactly what they'll do; for example, that's how backwards-induction in game theory works. And at MIRI we like to talk about problems where the agent can know exactly what it can do, because these stretch the limits of decision theory.

On the other hand, realistic agents probably mostly *don't* know with *certainty* what they'll do -- meaning my argument will usually apply in practice.

The agent might not follow the recommendations of CDT. Just because a CDT-respecting agent would definitely do a specific thing given all the information, *does not* mean that we have to imagine the agent in my argument knowing exactly what it will do. The agent in the argument might not be CDT-respecting.

Here on LessWrong, and at MIRI, there is often a tendency to think of CDT or EDT *as the agent* -- that is, think of agents as instances of decision theories. This is a point of

friction between MIRI's way of thinking and that of academic philosophy. In academic philosophy, the decision theory need not be an algorithm the agent is actually running (or indeed, could ever run). A decision theory is a *normative theory* about what an agent *should do*. This means that CDT, as a normative theory, can produce recommendations for non-CDT agents; and, we can judge CDT on the correctness or incorrectness of those recommendations.

Now, I think there are some advantages to the MIRI tendency -- for example, thinking in this way brings logical uncertainty to the forefront. However, I agree nonetheless with making a firm distinction between the *decision theory* -- a normative requirement -- and the *decision procedure* -- a real algorithm you can run, which obeys the normative requirement. The logical induction algorithm vs the logical induction criterion illustrates a similar idea.

Academic decision theorists extend this idea to the criticism of normative principles -- such as CDT and EDT -- for their behavior in scenarios which an agent would never get into, if it were following the advice of the respective decision theory. This is what's going on in the bomb example Will MacAskill uses. (Nate Soares argues against this way of reasoning, saying "decisions are for making bad outcomes inconsistent".)

If we *do* endorse the idea, this offers further support for the argument I'm making. It means we get to judge CDT for recommending that an agent accept a Dutch-book, even if the scenario depends on uncertainty over actions which a CDT advocate claims a CDT-compliant agent does not have.

This is particularly concerning for CDT, because this kind of argument is used especially to defend CDT. For example, it's hard to justify smoking lesion as a situation which a CDT or EDT agent could actually find itself in; but, a CDTer might reply, a decision theory needs to offer the right advice to a broad variety of agents. So CDT is already in the business of defending normative claims about non-CDT agents.

Dynamic Consistency

One might object: the argument simply illustrates a dynamic inconsistency in CDT. We already *know* that *both* CDT and EDT are dynamically inconsistent. What's the big deal?

Let me make some buckshot remarks before I dive into my main response here:

- First, this isn't just any old dynamic inconsistency. This is a Dutch Book. Dutch Books have a special status in the foundations of Bayesianism. So, one might consider this to be more concerning than mere dynamic inconsistency.
- Second, dynamic consistency is still bad. We may still take examples of dynamic inconsistency as counting against a normative theory, and seek a theory which is dynamically consistent in a fairly broad range of cases.
- Third, I think EDT really does have a dynamic-consistency advantage over CDT, and my argument is just one example of that.

This third point is the one I want to expand on.

In decision problems where the payoff depends *only on actions actually taken*, not on your policy, there is a powerful argument for the dynamic consistency of EDT:

Think of the entire observation/action history as a tree. Dynamic consistency means that at earlier points in the tree, the agent does not prefer for the decisions of later

selves to be different from what they will be. The restriction to actions (not policies) mattering for payoffs means this: selecting one action rather than another changes which branch we go down in the tree, but *does not* change the payoffs of other branches in the tree. This means that *even from a perspective beforehand*, an action can only make a difference down the branch where it is taken -- no spooky interactions across possible worlds. As a result, thinking about possible choices ahead of time, the contribution to early expected utility is *exactly* the expected utility that action will be assigned later at the point of decision, times the probability the agent ends up in that situation in the first place. Therefore, the preference about the decision must stay the same.

So, EDT is a dynamically consistent choice when actions matter but policy does not.

Importantly, this is **not** a no-Newcomblike-problems condition. It rules out problems such as counterfactual mugging, transparent Newcomb, Parfit's hitchhiker, and XOR Blackmail. However, it **does not** rule out the original Newcomb problem. In particular, we are highlighting the inconsistency where CDT wishes to be a 1-boxer, and similar cases.

Now, you *can* make a very similar argument for the dynamic consistency of CDT, if you define dynamic consistency based on counterfactuals: would you prefer to counterfact on your future self doing X? For Newcomb's problem, this gets us back to consistency -- for all that the CDT agent *wishes it could be EDT*, it would have no interest in the point-intervention that makes its future self one-box, for the usual reason: that would not cause Omega to change its mind.

However, this definition seems not to capture the most useful notion of dynamic consistency, since the same causal CDT agent would *happily* precommit to one-box. So I find the EDT version of the argument more convincing. I'm not presently aware of a similar example for EDT -- it seems Omega needs to consider the policy, not just the action really taken, in order to make EDT favor changing its actions via precommitments.

More on Probability Zero Actions

As I've said, my argument depends on the action a having nonzero probability. EDT isn't well-defined otherwise; how do you condition on a probability-zero event? However, there *are* some things we could try in order to get around the problem of division by zero - filling in the undefined values with sensible numbers. For example, we can use the Conditional Oracle EDT which Jessica Taylor defined, to "fill in" the otherwise-undefined conditionals.

However, recall I said that the argument was blocked for *two* reasons:

- EDT not being defined.
- The bet conditional on a is worth nothing if a has probability zero.

So, if we've somehow made edt(a) well-defined for probability-zero a, can we patch the second problem?

We can try to flip the argument I made around: for probability zero actions, we *pay* the agent to take on a bet (so that it will do it even though it's worthless). Then later, we *charge* the agent to offload the bet which it now thinks is unfavorable.

The problem with this argument is, if the agent doesn't take action a anyway, then the conditional bet will be nullified regardless; we can't force the agent into a corner where it prefers to nullify the bet, so, we don't get a full Dutch Book (because we can't guarantee that we make money off the agent -- indeed, we would only make money if the agent ends up taking the action which it previously assigned probability zero to taking).

However, we *do* get a *moderately* damning result: limiting attention to just the action a and the new alternative a′ which both does a and also pays to cancel the bet, **CDT strictly prefers that the agent be Dutch Booked* rather than just do a. This seems pretty bad: CDT isn't actually recommending taking a Dutch Book, *BUT*, it would rather take a Dutch Book than take an alternative which is otherwise the same but which does not get Dutch Booked.

So, we can still make a moderately strong argument against divergence between counterfactuals and conditionals, even if actions have probability zero. But not a proper Dutch Book.

A Few Words on Troll Bridge

At the beginning of this, I said that I didn't necessarily take this to be an argument for EDT over CDT. In the past, I've argued this way:

1. "Here's some argument that CDT=EDT."
2. "Since EDT gets the answer more simply, while CDT has to posit extra information to get the same result, we should prefer EDT."

However, this argument has at least two points against it. First, the arguments for CDT=EDT generally have some assumptions, such as nonzero probability for actions. CDT is a strictly more general framework when those conditions are not met. Theories of rational agency should be as inclusive as possible, when rationality does not demand exclusivity. So one might still prefer CDT.

Second, as I mentioned at the beginning of this post, the Troll Bridge problem strongly favors CDT over EDT. Counterintuitively, it's perfectly possible for a CDT agent to keep its counterfactual expectations exactly in agreement with its conditional expectations, and yet get Troll Bridge right -- even though we are doomed to get Troll Bridge wrong if we *directly* use our conditional expectations. Insisting on a distinction "protects" us from spurious counterfactual reasoning. (I may go over this phenomenon in more detail in a future post. But perhaps you can see why by reviewing the Troll Bridge argument.)

So, my *current* take on the CDT=EDT hypothesis is this:

1. We should think of counterfactuals as having real, independent truth. In other words, cdt(a) does not *reduce to* $edt(a)*. Counterfactual information tells us something above and beyond probabilistic information.

2. Counterfactuals are subjective in the same way that probabilities are subjective. The "independent truth" of counterfactuals does not mean there is one objectively correct counterfactual which every agent is normatively required to agree with. So there doesn't need to be a grand theory of logical counterfactuals -- there are many different subjectively valid beliefs.
3. However, as with probability theory, there are important notions of coherence which constrain subjective beliefs. In particular, counterfactual beliefs should almost always equal conditional beliefs, at least when the antecedent has positive probability.
4. Furthermore, conditional beliefs act a whole lot more like stereotypical CDT counterfactuals than most people seem to give them credit for. Something can't correlate with your action unless it contains *information you don't have* about your action. This is a high bar to pass, and will typically not be passed in e.g. twin prisoner's dilemma. (So, to solve these problems requires something further, e.g. updatelessness, Löbian handshakes, ???).

This is not a strongly held view, but it is the view that has made the most sense of counterfactual reasoning for me.

As I've mentioned in the past, the CDT=EDT hypothesis is almost the most boring possible answer to the question "how do (logical) counterfactuals work?" -- it doesn't do very much to help us solve interesting decision problems. If we factor decision theory into the two parts (1) "What are the (logical) counterfactuals?" (2) "How do we use counterfactuals to make decisions?" then I see the CDT=EDT hypothesis as a solution to (1) which shoves an awful lot of the interesting work of decision theory into (2). IE, to solve the really interesting problems, we would need logically-updateless UDT or even more exotic approaches.

In particular, for variants of Newcomb's problem where the predictor is quite strong but doesn't know *as* much as the agent does about what the agent will choose, this post implies that TDT *either* two-boxes, or, is vulnerable to the Dutch Book I construct. This is unfortunate.

Conclusion

Frankly, I find it somewhat embarrassing that I'm still going on about CDT vs EDT. After all, Paul Christiano said, partially in response to my own writing which he cited:

> There are many tricky questions in decision theory. In this post, I'll argue that the choice between CDT and EDT isn't one of them.

I wish I could say this will be my final word on the subject. The contents of this post do feel quite definitive in the sense of giving a settled, complete view. However, the truth is that it only represents my view as of November or early December of 2019. Late December and early January saw some developments which I'm excited to work out further and post about.

My Current Take on Counterfactuals

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

[Epistemic status: somewhat lower confidence based on the fact that I haven't worked out a detailed theory along the lines I've suggested, yet.]

I've felt like the problem of counterfactuals is "mostly settled" (modulo some math working out) for about a year, but I don't think I've really communicated this online. Partly, I've been waiting to write up more formal results. But other research has taken up most of my time, so I'm not sure when I *would* get to it.

So, the following contains some "shovel-ready" problems. If you're convinced by my overall perspective, you may be interested in pursuing some of them. I think these directions have a high chance of basically solving the problem of counterfactuals (including logical counterfactuals).

Another reason for posting this rough write-up is to get feedback: am I missing the mark? Is this not what counterfactual reasoning is about? Can you illustrate remaining problems with decision problems?

I expect this to be *much more difficult to read* than my usual posts. It's a brain-dump. I make a lot of points which I have not thought through sufficiently. Think of it as a frozen snapshot of a work in progress.

Summary.

1. I can Dutch-book any agent whose subjective counterfactual expectations don't equal their *conditional* expectations. I conclude that counterfactual expectations should equal conditional probabilities. IE, evidential decision theory (EDT) gives the correct counterfactuals.
2. *However*, the Troll Bridge problem is real and concerning: EDT agents are doing silly things here.
3. Fortunately, there appear to be ways out. One way out is to maintain that *subjective counterfactual expectations should equal conditional expectations* while also maintaining a distinction between those two things: counterfactuals are not *computed from* conditionals. As we shall see, this allows us to ensure that the two are always equal in *real* situations, while strategically allowing them to differ in some *hypothetical* situations (such as Troll Bridge). This seems to solve all the problems!
4. However, I have not yet concretely constructed any way out. I include detailed notes on open questions and potential avenues for success.

What does it mean to pass Troll Bridge?

It's important to briefly clarify the nature of the goal. First of all, Troll Bridge really relies on the agent "respecting logic" in a particular sense.

- Learning agents who don't reason logically, such as RL agents, can be fooled by a troll who punishes exploration. However, that doesn't seem to get at the point of Troll Bridge. It just seems like an unfair problem, then.
    - **The first test of "real Troll Bridge" is:** does the example make crossing totally impossible? Or does it depend on the agent's starting beliefs?
        - If the probabilistic troll bridge only concluded that you won't cross *if you have a sufficiently high probability that PA is inconsistent*, this would come

off as totally reasonable rather than insane. It's a refusal to cross *regardless of prior* which seems so problematic.
        - If an RL agent starts out thinking crossing is good, then this impression will continue to be reinforced (if the troll is only punishing exploration). Such an example only shows that *some* RL agents cannot cross, due to the combination of low prior expectation that crossing is good, and exploration being punished. This is unfortunate, and illustrates a problem with exploration if exploration can be detected by the environment; but it is not as severe a problem as Troll Bridge proper.
    - **The second test** is whether we can strengthen the issue to 100% failure *while still having a sense that the agent has a way out, if only it would take it*. In the RL example, we can strengthen the exploration-punisher by also punishing *unjustified crossing* more generally, in the sense of crossing without an empirical history of successful crosses to generalize from. If your prior suggests crossing is bad, you'll be punished every time you cross because it's exploration. If your prior suggests crossing is good, you'll be punished every time you cross because your pro-crossing stance is not empirically justified. So no agents are rewarded for crossing. This passes the first test of "real Troll Bridge". But there is no longer any sense that the agent is crazy. In original Troll Bridge, there's a strong intuition that "the agent could just cross, and all would be well". Here, there is nothing the agent can possibly do.
- Secondly, an agent could reason logically but with some looseness. This can fortuitously block the Troll Bridge proof. However, the approach seems worryingly unprincipled, because we can "improve" the epistemics by tightening the relationship to logic, and get a decision-theoretically much worse result.
    - The problem here is that we have some epistemic principles which suggest tightening up is good (it's free money; the looser relationship doesn't lose much, but it's a dead-weight loss), and no epistemic principles pointing the other way. So it feels like an unprincipled exception: "being less dutch-bookable is generally better, but hang loose in this one case, would you?"
    - Naturally, this approach is still very interesting, and could be pursued further -- especially if we could give a more principled reason to keep the observance of logic loose in this particular case. But this isn't the direction this document will propose. (Although you *could* think of the proposals here as giving more principled reasons to let the relationship with logic be loose, sort of.)
    - So here, we will be interested in solutions which "solve troll bridge" in the stronger sense of getting it right while fully respecting logic. IE, updating to probability 1 (/0) when something is proven (/refuted).

There is another "easy" way to pass Troll Bridge, though: just be CDT. (By CDT, I don't mean classical causal decision theory -- I mean decision theory which uses *any* notion of counterfactuals, be it based on physical causality, logical causality, or what-have-you.)

The Subjective Theory of Counterfactuals

Sam presented Troll Bridge as an argument in favor of CDT. For a long time, I regarded this argument with skepticism: yes, CDT allows us to solve it, but what is logical causality?? What are the correct counterfactuals?? I was incredulous that we could get real answers to such questions, so I didn't accept CDT as a real answer.

I gradually came to realize that Sam didn't see us as needing all that. For him, counterfactuals were simply a more general framework, a generality which happened to be needed to encode what humans see as the correct reasoning.

Look at probability theory as an analogy.

If you were trying to invent the probability axioms, you would be led astray if you thought too much about what the “objectively correct” beliefs are for any given situation. Yes, there is a very interesting question of what the prior should be, in Bayesianism. Yes, there are things we can say about “good” and “bad” probability distributions for many cases. However, it was important that at some point someone sat down and worked out the theory of probability under the assumption that those questions were *entirely subjective*, and the only *objective* things we can say about probabilities are the basic coherence constraints, such as P(~A)=1-P(A), etc.

Along similar lines, the subjectivist theory of counterfactuals holds that *we have been led astray* by looking too hard for some kind of correct procedure for taking logical counterfactuals. Instead, starting from the assumption that a very broad range of counterfactuals can be subjectively valid, we should seek the few “coherence” constraints which distinguish rational counterfactual beliefs from irrational.

In this perspective, getting Troll Bridge right isn’t particularly difficult. The Troll Bridge argument is blocked at the step where [the agent proves that crossing implies bad stuff] implies [the agent doesn’t cross]. The agent’s *counterfactual* expected value for crossing can still be high, even if it has proven that crossing is bad. Counterfactuals have a lot of freedom to be different from what you might expect, so they don’t have to respect proofs in that way.

Following the analogy to probability theory, we still want to know what the axioms *are.* How are rational counterfactual beliefs constrained?

I’ll call the minimalist approach “Permissive CDT”, because it makes a strong claim that "almost any" counterfactual reasoning can be subjectively valid (ie, rational):

Permissive CDT

What I’ll call “Permissive CDT” (PCDT) has the following features:

- There is a basic counterfactual conditional, C(A|B).
- This counterfactual conditional obeys the axiom C(A|B)&B -> A.
- There may be additional axioms, but they are weak enough to allow 2-boxing in Newcomb as subjectively valid.
- There is no chicken rule or forced exploration rule; agents always take the action which looks counterfactually best.

Note that *this isn’t totally crazy*. C(A|B)&B -> A means that counterfactuals had better take the actual world to the actual world. This means a counterfactual hypothesis sticks its neck out, and can be disproven if B is true (so, if B is an action, we can *make* it true in order to test).

Note that I've excluded exploration from PCDT. This means we can't expect as strong of a learning result as we might otherwise. However, *with* exploration, we would eventually take disastrous actions. For example, if there was a destroy-the-world button, the agent would eventually press it. So, we probably don't want to force exploration just for the sake of better learning guarantees!

Instead, we want to use "VOI-exploration". This just means: PCDT naturally chooses some actions which are suboptimal in the short term, due to the long-term value of information. (This is just a fancy way of saying that it's worthwhile to do experiments sometimes.) To vindicate this approach, we would want some sort of VOI-exploration result. For example, we may be able to prove that PCDT successfully learns under some restrictive conditions (EG, if it knows no actions have catastrophic consequences). Or, even better, we could characterize *what it can learn* in the absence of nice conditions (for example, that it explores everything except actions it thinks are too risky).

I claim PCDT is wrong. I think it's important to set it up properly in order to *check* that it's wrong, since belief in some form of CDT is still widespread, and since it's actually a pretty plausible position. But I think my Dutch Book argument is fairly damning, and (as I'll discuss later) I think there are other arguments as well.

Sam advocated the PCDT-like direction for some time, but eventually, he came to agree with my Dutch Book argument. So, I think Sam and I are now mostly on the same page, favoring a version of subjective counterfactuals which requires more EDT-like expectations.

There are a number of formal questions about PCDT which would be useful to answer. This is part of the "shovel-ready" work I promised earlier. The list is quite long; I suggest skipping to the next section on your first read-thru.

Question: further axioms for PCDT? What else might we want to assume about the basic counterfactuals? What can't we assume? (What assumptions would force EDT-like behavior, clashing with the desideratum of allowing 2-boxing? What assumptions lead to EDT-like failure in Troll Bridge? What assumptions allow/disallow sensible logical counterfactuals? What assumptions force inconsistency?)

- We can formalize PCDT in logical induction, by adding a basic counterfactual $A \mapsto B$ to the language.
- We might want counterfactuals to give sensible probability distributions on sentences, so $A \mapsto B$ and $A \mapsto \neg B$ are mutually exclusive and jointly exhaustive. But we definitely don't want too many "basic logic" assumptions like that, since it could make counterlogicals undefinable, leading us back to problems taking counterfactuals when we know our own actions.

Question: Explore PCDT and Troll Bridge. Examine which axioms for PCDT are compatible with successfully passing Troll Bridge (in the sense sketched earlier).

Question: Learning theory for PCDT? A learning theory is important for a theory of counterfactuals, because it gives us a story about why we might expect counterfactual reasoning to be correct/useful. If we have strong guarantees about how counterfactual reasoning will come to reflect the true consequences of actions according to the environment, then we can trust counterfactual reasoning. Again, we can formalize PCDT via logical induction. What, then, does it learn? PCDT should have a learning theory somewhat similar to InfraBayes, in the sense of relying on VOI exploration instead of forcing exploration with an explicit mechanism.

- Some sort of no-traps assumption is needed.
  - Weak version: assume that the logical inductor is 1-epsilon confident that the environment is trap-free, or something along those lines (and also assume that the true environment *is* trap-free).
  - Strong version: don't assume anything about the initial beliefs. Show that the agent ends up exploring sufficiently *if it believes it's safe to do so;* that is, show that belief in traps is in some sense the *only* obstacle to exploring enough (and therefore learning). (Provided that the discount rate is near enough to 1, of course; and provided the further learnability assumptions I'll mention below.)
  - Stretch goal: deal with human feedback about whether there are traps, venturing into alignment theory rather than just single-agent decision theory.
    - See later section: applications to alignment.
- Some sort of "good feedback" assumption is needed, ensuring that the agent gets enough information about the utility.
  - The most obvious thing is to assume an RL environment with discounting, much like the learning theory of InfraBayes.
  - It might be interesting to generalize further, having a broader class of utility functions, with feedback which narrows down the utility incrementally. RL is just a

special case of this where the discounting rate determines the amount that any given prefix narrows down the ultimate utility.
    - Generalizing even further, it could be interesting to abandon the utility as a function of history at all, and instead rely only on subjective expectations (like the Orthodox Case Against Utility Functions suggests).
        - I suspect this is good for the "stretch goal" mentioned previously, of dealing with human feedback about whether there are traps. See later section: applications to alignment.
- Some sort of no-newcomblike assumption is probably needed.
    - This is very similar to how Sam's tiling result required a no-newcomb assumption, and asymptotic decision theory required a no-newcomb assumption.
    - In other words, a "CDT=EDT" assumption. (A Newcomblike problem is precisely one where CDT and EDT differ.)
    - Like the no-traps assumption, there are two different ways to try and do this:
        - Weaker: assume the logical inductor is very confident that the environment isn't Newcomblike.
        - Stronger: don't assume anything, but show that the agent ends up differentiating between hypotheses *to the extent it's possible to do so*. Newcomblike hypotheses make payoffs of actions themselves depend on what actions are taken, and so, are impossible to distinguish via experiment alone. But it should be possible to show that, based on VOI exploration, the agent *eliminates the eliminable hypotheses*, and distinguishes between the rest based on subjective plausibility; and (***according to PCDT, but not according to me***) this is the best we can hope to do.
- Realizability?
    - There should be a good result assuming realizability, at least. And perhaps that's enough for a start -- it would still be an improvement in our philosophical understanding of counterfactuals, particularly when combined with other results in the research program I'm outlining here.
    - But I'm also suspicious that there's *some* degree of ability to deal with unrealizable cases.
    - The right assumption has to do with there being a trader capable of tracking the environment sufficiently.
    - Unlike Bayes or InfraBayes, we don't have to worry about hypotheses competing; any predictively useful constraint on beliefs will be learned.
        - Bayes "has to worry" in the sense that non-realizable cases can create oscillation between hypotheses, due to there not being a unique best hypothesis for predicting the environment. (This might harm decision theory, by oscillating between competing but incompatible strategies.)
        - InfraBayes doesn't seem to have *that* worry, since it applies to non-realizable cases. (Or does it? Is there some kind of non-oscillation guarantee? Or is non-oscillation part of what it means for a set of environments to be learnable -- IE it can oscillate in some cases?) But InfraBayesian learning is still a one-winner type system, in that we don't learn *all* applicable partial models; only the *most useful* converges to probability 1.
        - Logical induction, on the other hand, guarantees that *all* applicable partial models are learned (in the sense of *finitely many violations*). But, to what extent can we translate this to a decision-theoretic learning result?
    - As an example, I think it should be possible to learn to use a source of randomness in rock-paper-scissors against someone who can perfectly predict your decision, but not the extra randomness.
        - I'm imagining discretized choices, so there's a finite number of options of the form "½ ½ 0", "1 0 0", etc.
        - If the adversary were only trying to do the best in each round individually, this is basically a multi-armed bandit problem, where the button "play ⅓ ⅓ ⅓" has the best payoff. But we also need to show that the adversary can't use a long-con strategy to mislead the learning.

    - I think one possible proof is to consider a trader who predicts that every option will be at best as good as ⅓ ⅓ ⅓ on average, in the long term. If this trader does poorly, then the adversary must be doing a poor job. If this trader does well, then (because we learn the payoff of ⅓ ⅓ ⅓ correctly for sure) the agent must converge to playing ⅓ ⅓ ⅓. So, either way, the agent must eventually do at least as well as the optimal strategy.

Question: tiling theory for PCDT?

- It seems like this would admit some version of Sam's tiling result and/or Diffractor's tiling sketch.
- As with the other results, a major motivation (from where I'm currently sitting) is to show that PCDT is worse than more EDT-like alternatives. I strongly suspect that tiling results for PCDT will be more restrictive than for the decision theory I'm going to advocate for later, *precisely because* tiling must require a no-newcomb type restriction. PCDT, faced with the possibility of encountering a Newcomblike problem at some point, *should absolutely* self-modify in some way.
- Also, the tiling result necessarily excludes updateless-type problems, such as counterfactual mugging. None of the proposals considered here will deal with this.

This concludes the list of questions about PCDT. As I mentioned earlier, PCDT is being presented in detail primarily to contrast with my real proposal.

But, before I go into that, I should discuss *another* theory I don't believe: what I see as the "opposite" of PCDT. My real view will be a hybrid of the two.

The Inferential Theory of Counterfactuals

The inferential theory is what I see as the core intuition behind EDT.

The intuition is this: *we should reason about the consequences of actions in the same way that we reason about information which we add to our knowledge.*

Another way of putting this is: *hypothetical reasoning and counterfactual reasoning are one and the same*. By *hypothetical* reasoning, I mean temporarily adding something to the set of things you know, in order to see what would follow.

In classical Bayesianism, we add new information to our knowledge by performing a Bayesian update. Hence, the inferential theory says that we Bayes-update on possible actions to examine their consequences.

In logic, adding new information means adding a new axiom from which we can derive consequences. So in proof-based EDT (aka MUDT), we examine what we could prove if we added an action to our set of axioms.

So the inferential theory gives us a way of constructing a version of EDT from a variety of epistemic theories, not just Bayesianism.

I think the inferential theory is probably wrong.

Almost any version of the inferential theory will imply getting Troll Bridge wrong, just like proof-based decision theory and Bayesian EDT get it wrong. That's because the inference **[action *a* implies bad stuff] & [action *a*] |→ [bad stuff]** is valid. So the Troll Bridge argument is likely to go through.

Jessica Taylor talks about something she calls "counterfactual nonrealism", which sounds a lot like what I'm calling the subjective theory of counterfactuals. However, she appears to

also wrap up the inferential theory in this one package. I'm surprised she views these theories as being so close. I think they're starting from very different intuitions. Nonetheless, I do think what we need to do is combine them.

Walking the Line Between CDT and EDT

So, I've claimed that PCDT is wrong, because any departure from EDT (and thus the inferential theory) is dutch-book-able. Yet, I've also claimed that the inferential theory is itself wrong, due to Troll Bridge. So, what do I think is right?

Well, one way of putting it is that *counterfactual reasoning should match hypothetical reasoning* ***in the real world****, but shouldn't necessarily match it* ***hypothetically****.*

This is precisely what we need in order to block the Troll Bridge argument. (At least, that's one way to block the argument -- there are other steps in the argument we could block.)

As a simple proof of concept, consider a CDT whose counterfactual expectations for crossing and not crossing *just so happen* to be the same as its evidential expectations, namely, cross = +10, not cross = 0. This isn't Dutch-bookable, since the counterfactuals and conditionals agree.

In the Troll Bridge hypothetical, we *prove* that [cross]->[U=-10]. This will make the *conditional* expectations poor. But this *doesn't have to change the counterfactuals*. So (within the hypothetical), the agent can cross anyway. And crossing gets +10. So, the Lobian proof doesn't go through. Since the proof doesn't go through, the conditional expectations can *also* consistently expect crossing to be good; so, we never *really* see a disparity between counterfactual expectation and conditional expectation.

Now, you might be thinking: couldn't the troll use the disparity between counterfactual and conditional expectation as its trigger to blow up the bridge? I claim not: the troll would, then, be punishing anyone who made decisions in a way different from EDT. Since we know EDT doesn't cross, it would be obvious that no one should cross. So we lose the sense of a dilemma in such a version of the problem.

OK, but how do we accomplish this? Where does the nice coincidence between the counterfactuals and evidential reasoning come from, if there's no internal logic requiring them to be the same?

My intuition is that we want something similar to PCDT, but with more constraints on the counterfactuals. I'll call this ***restrictive*** **counterfactual decision theory (RCDT):**

- RCDT should have extra constraints on the counterfactual expectations, sufficient to *guarantee that the counterfactuals we eventually learn will be in line with the conditional probabilities we eventually learn*. IE, the two asymptotically approach each other (at least in circumstances where we have good feedback; probably not otherwise).
- The constraints *should not* force them to be exactly equal at all times. *In particular*, the constraints ***must not*** force counterfactuals to "respect logic" in the sense that would force failure on Troll Bridge. For example, If $A \to B$ implies $A \mapsto B$, then a proof that crossing the bridge is bad could stop us from crossing it. We can't let RCDT do that.

To build intuition, let's consider how PCDT and RCDT learn in a Newcomblike problem.

Let's say we're in a Newcomb problem where the small box contains $1 and the big box may or may not contain $10, depending on whether a perfect predictor believes that the agent will 1-box.

Suppose our PCDT agent starts out mainly believing the following counterfactuals (using C() for counterfactual expectations, counting just the utility of the current round):

C(U|1-box) = 10 * P(1-box)

C(U|2-box) = 1 + 10 * P(1-box)

In other words, the classic physical counterfactuals. I'll call this hypothesis PCH for physical causality hypothesis.

We also have a trader who thinks the following:

C(U|1-box) = 10

C(U|2-box) = 1

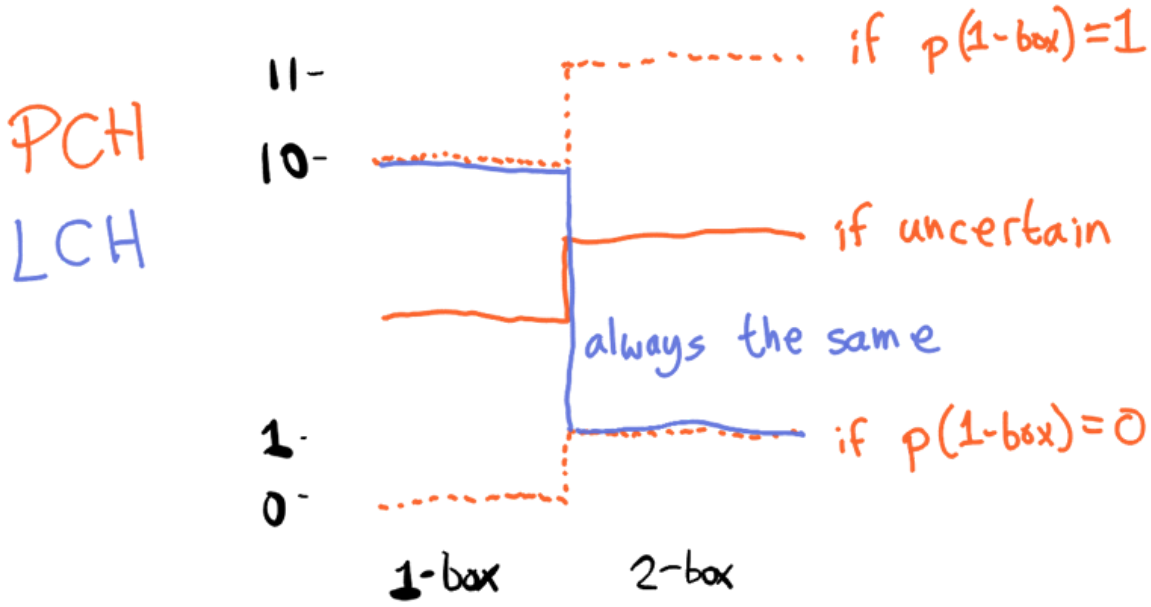
I'll call this LCH for logical causality hypothesis.

Now, if the agent's *overall* counterfactual expectation (including value for future rounds, which includes exploration value) is quite different for the two actions, then the logical inductor should be quite clear on which action the agent will take (the one with higher utility); and if that's so, then LCH and PCH will agree quite closely on the expected utility of said action. (They'll just disagree about the *other* action.) So little can be learned on such a round. As long as PCH is dominating things, that action would *have* to be 2-boxing, since an agent who mostly believes PCH would only ever 1-box for the VOI -- and there's no VOI here, since in this scenario the two hypotheses agree.

But that all seems well and good -- no complaints from me.

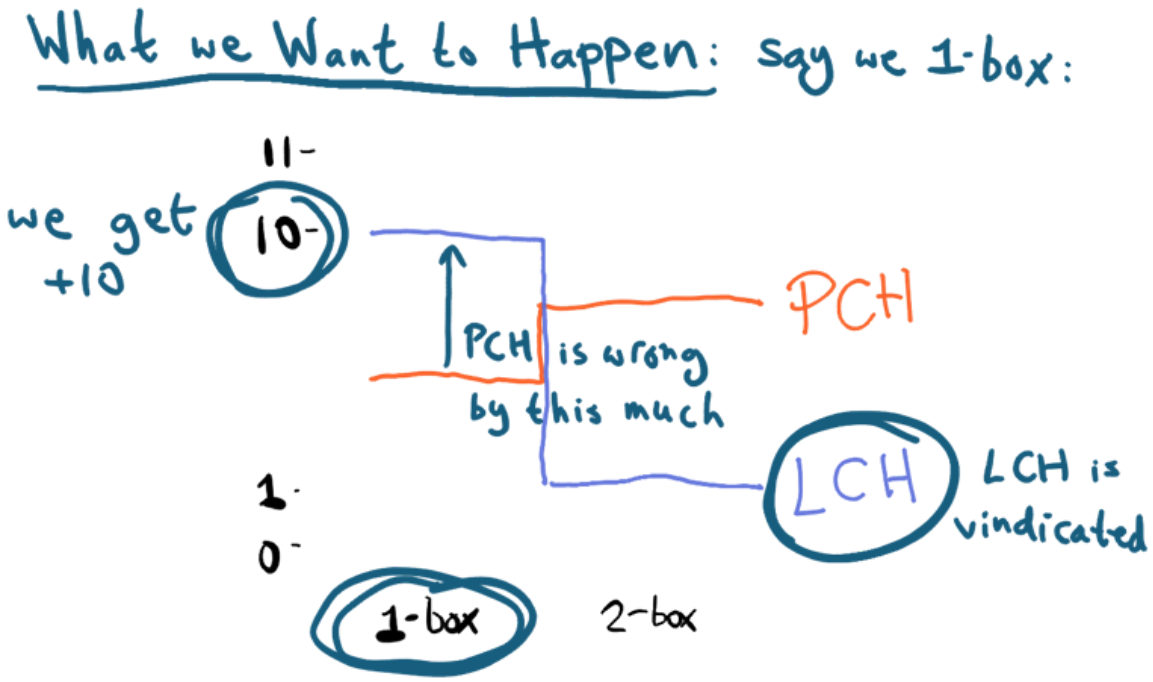
Now suppose instead that the overall counterfactual expectation is *quite similar* for the two actions, to the point where traders have trouble predicting which action will be taken.

In that case, LCH and PCH have quite different expectations:



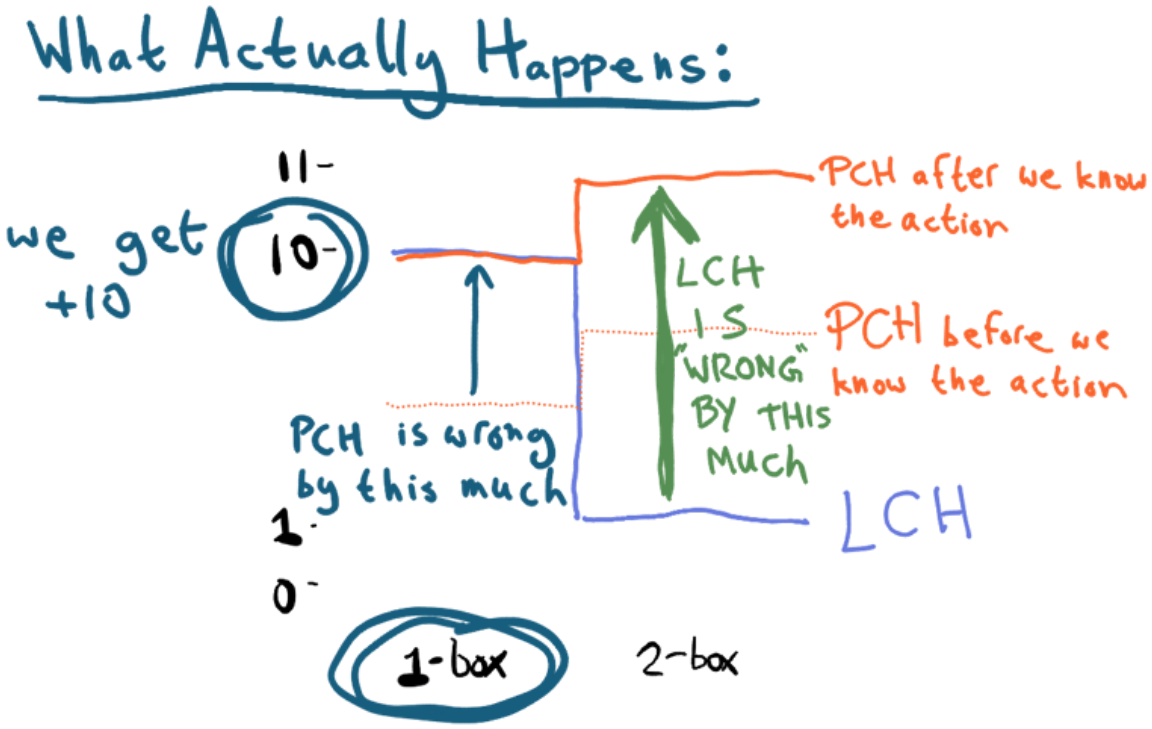
(even though the x-axis is a boolean variable, I drew lines that are connected across two sides so that one can easily see how the uncertain case is the average of the two certain cases.)

The significant difference in expectations between LCH and PCH in the uncertain case makes it look as if we can learn something. We don't know which action the agent will actually take, but we do know that LCH ends up being correct about the value of whatever action is taken. So it looks like PCH traders should lose money.



However, that's not the case.

Because we have a "basic counterfactual" proposition for what would happen if we 1-box *and* what would happen if we 2-box, and *both of those propositions stick around* , LCH's bets about what happens in either case *both* matter. This is unlike conditional bets, where if we 1-box, then bets conditional on 2-boxing disappear, refunded, as if they were never made in the first place.



When the logical inductor observes that the agent 1-boxes, and sees the +10, the expected value of *that* counterfactual payoff must move to +10 (since counterfactuals on what actually happens must match what actually happens). However, the *other* counterfactual -- the one on 2-boxing -- moves to ***+11***, because PCH is still the dominant belief; the agent learned that it indeed 1-boxed, so, it now believes that it *would have* received 11 by 2-boxing.

Since 2-boxing is a counterfactual scenario which we'll never get any solid feedback on, the belief about what reward we could have gotten can stay around 11 forever. Any money LCH bet on a payoff of 1 from 2-boxing is probably lost for good.

So it doesn't appear that LCH traders can make a profit.

Question: Verify/explore this.

- Can PCDT really favor PCH forever? Is there no LCH strategy which could make profit?
- Does LCH necessarily bleed money, or are there versions of LCH which break even here? Does PCH become increasingly confident, or does it just remain stable?
- Can we give a rational justification for this behavior?
    - EG, at one point, Sam claimed that this was a perfectly reasonable epistemic state for an agent who thinks Omega is rewarding it with +10 *for exploring* (because the only way it gets the +10 from 1-boxing is if it does so for VOI, since it would never do so for the sake of the value it gets in that round), but *would not reward it similarly if the action were taken for its own sake* (because on non-exploration rounds, the agent thinks the value of 1-boxing would be 0).
    - It's clear that learning is impossible under such a scenario. Yet, this does not necessarily justify a learning rule which lets PCH dominate over LCH forever; we want to be able to learn 5&10 correctly by simply experimenting, and PCH is essentially stopping us from doing that here.
    - I'm interested in further thoughts on this. If we adopt a learning rule which *doesn't* favor PCH in the scenario given (IE, straight Newcomb), does it then cause pathological behavior in Sam's scenario (exploration-rewarding Omega)? If

so, how concerning is this? Are there alternative arguments in favor of the PCDT learning behavior that I'm calling pathological?

Question: how can we avoid this pathology?

- Option 1: More constraints on counterfactuals.
  - Is there some way to add axioms to the PCDT counterfactuals, to make them (A) learn LCH, while still (B) passing Troll Bridge?
- Option 2: Something more like conditional bets.
  - The behavior I labelled "want" is like conditional bet behavior.
  - However, standard conditional bets won't quite do it.
    - If we make decisions by looking at the conditional probabilities of a logical inductor (as defined by the ratio formula), then these will be responsive to proofs of action->payoff, and therefore, subject to Troll Bridge.
    - What we want to do is look at conditional bets *instead of* raw conditional probabilities, with the hope of escaping the Troll Bridge argument while keeping expectations empirically grounded.
    - Normally conditional bets A|B are constructed by betting on A&B, but also hedging against ¬B by betting on that, such that a win on ¬B exactly cancels the loss on A&B.
    - In logical induction, such conditional bets must be *responsive* to a proof of B->A; that is, since B->A means ¬(B&¬A), the bet on A&B must now be worth what a bet on just B would be worth. More importantly, a bet on B&¬A must be worthless, making the conditional probability zero.
  - So I see this as a challenge to make "basic" conditional bets, rather than conditional bets derived from boolean combinations as above, and make them such that they aren't responsive to proofs of B->A like this.
    - Intuitively, a conditional bet A|B is just a contract which pays out given A&B, but which is refunded if ¬B.
    - I think something like this has even been worked out for logical induction at some point, but Scott and Sam and I weren't able to quickly reconstruct it when we talked about this.
      - (And it was likely responsive to proofs of A->B.)
    - A notion of conditional bet which isn't responsive to A->B isn't totally crazy, I claim.
      - In many cases, the inductor might learn that A->B makes B|A a really good bet.
      - But if crossing the bridge never results in a bad outcome in reality, it should be possible to maintain an exception for cross->bad.
    - Philosophically, this is a radical rejection of the ratio formula for conditional probability.
      - Rejection of the ratio formula has been discussed in the philosophical literature, in part to allow for conditioning on probability-zero events. EG, the Lewis axioms for conditional probabilities.
      - As far as I've seen, philosophers still endorse the ratio formula *when it meaningfully applies*, ie, when you're not dividing by zero. It's just rejected *as the definition* of conditional probability, since the ratio formula isn't well-defined in some cases where the conditional probabilities do seem well-defined.
      - However, I suspect the rejection of the inference from A->B to B|A constitutes a more radical departure than usual.
    - We most likely still need the chicken rule here, unlike with basic counterfactuals.
      - The *desired behavior* we're after here, in order to give LCH an advantage, is to nullify bets in the case that their conditions turn out false. This doesn't seem compatible with usable conditionals on zero-probability events.
        - (But it would be nice if this turned out otherwise!)

        - At first it might sound like using chicken rule spells doom for this approach, since chicken rule is the original perpetrator in Troll Bridge. But I think this is not the case.
            - In the step in Troll Bridge where the agent examines its own source code to see why it might have crossed, we see that the chicken rule triggered *or* the agent had higher conditional-contract expectation on crossing. So it's possible that the agent crosses for entirely the right reason, blocking the argument from going through.
            - We could try making the troll punish chicken-rule crossing *or* crossing based on conditional-contract expectations which differ from the true conditional probabilities; but this seems exactly like the case we examined for PCDT. Crossing because crossing looks like a good idea in some sort of expectation is *a good reason to cross*; if we deny the agent this possibility, then it just looks like an impossible problem. The troll would just be blowing up the bridge for anyone who doesn't agree with EDT; but EDT doesn't cross.
        - If this modified-conditional-bet option works out, to what extent does this vindicate the inferential theory of counterfactuals?
- Option 3: something else?

Question: Performance on other decision problems?

- It's not updateless, so obviously, it can't get everything. But, EG, how does it do on XOR?

Question: What is the learning theory of the working options? How do they compare?

- If we can get a version which favors LCH in the iterated Newcomb example, then the learning theory should be much like what I outlined for PCDT, with the exception of the no-newcomb clause.
    - It would be great to get a really good picture of the differences in optimality conditions for the different alternatives. EG, PCDT can't learn when it's suspicious of Newcomblike situations. But perhaps RCDT can't seriously maintain the hypothesis that Omega is tricking it on exploration rounds specifically (as Sam conjectured at one point), while PCDT can; so there may be some class of situations where, though neither can learn, PCDT has an advantage in terms of being able to perform well if it wins the correct-belief lottery.

Question: What is the tiling theory of the working options? How do they compare?

- Like PCDT, RCDT should have some tiling proof along the lines of Sam's tiling result and/or Diff's tiling sketch.
- Again, it would be interesting to get a really good comparison between the options.
    - I suspect that PCDT has really poor tiling in Newcomblike situations, whereas RCDT does not. I really want that result, to show the strength of RCDT on tiling grounds.

Applications to Alignment

Remember how Sam's tiling theorem requires feedback on counterfactuals? That's implausible for a stand-alone agent, since you don't get to see what happens for untaken actions. But if we consider an agent getting feedback from a human, suddenly it becomes plausible.

However, human feedback does have some limitations.

- It should be sparse. A human doesn't want to give feedback on every counterfactual for every decision. But the human could focus attention on counterfactual expectations which look very wrong.
- Humans aren't great at giving reward-type feedback. (citation: I've heard this from people at CHAI, I think.)
- Humans are even worse at giving full-utility feedback.
    - This would require humans to evaluate what's likely to happen in the future, from a given state.
- So, we have to come up with feedback models which could work for humans.
    - Simpler models like non-sparse human feedback on (counterfactual) rewards could still be developed for the sake of incremental progress, of course.
        - One model I think is unrealistic but interesting: humans providing better and better bounds for overall expected utility. This is similar to providing rewards (because a reward bounds the utility), but also allows for providing some information about the future (humans might be able to see that a particular choice would destroy such and such future value).
    - Approval feedback is easier for humans to give, although approval learning doesn't so much allow the agent to use its own decision theory (and especially, its own world model and planning).
    - Obviously some of Vanessa's work provides relevant options to consider.
- As Vanessa has pointed out, this can help deal with traps (provided the supervisor has good information about traps in some sense). This is obviously a major factor in the theory, since traps are part of what blocks nice learning-theoretic results.
- I would like to consider options which ***allow for human value uncertainty.***
    - One model of particular interest to me is ***modeling the human as a logical inductor***. This has several interesting features.
        - The feedback given at any one time does not need to be accurate in any sense, because logical inductors can be terrible at first.
        - If utility is a LUV, there can be no explicit utility function at all.
            - This in-effect allows for uncomputable utility functions, such as the one in the procrastination paradox, as I discussed in *an orthodox case against utility functions*.
        - The convergence behavior can be thought of as a model of human philosophical deliberation.
- It would be super cool to have a combined alignment+tilling result.
- I don't particularly expect this to solve wireheading or human maniputalion; it'll have to mostly operate under the assumption that feedback has not been corrupted.

Why Study This?

I suspect you might be wondering what the value of this is, in contrast to a more InfraBayesian approach. I think a substantial part of the motivation for me is just that I am curious to see how LIDT works out, especially with respect to these questions relating to CDT vs EDT. However, I think I can give some reasons why this approach might be necessary.

- Radical Probabalism and InfraBayes are plausibly *two orthogonal dimensions of generalization for rationality*. Ultimately we want to generalize in both directions, but to do that, working out the radical-probabilist (IE logical induction) decision theory in more detail might be necessary.
- The payoff in terms of alignment results for this approach might give some benefits which can't be gotten the other way, thanks to the study of subjectively valid LUV expectations which don't correspond to any (computable) explicit utility function. How could a pure InfraBayes approach align with a user who has LUV values?
- This approach offers insights into the big questions about counterfactuals which are at best only implicit in the InfraBayes approach.

    - The VOI exploration insight is the same for both of them, but it's possible that the theory is easier to work out in this case. I think learnability here can be stated in terms of a no-trap assumption and (for PCDT) a no-newcomb assumption. AFAIK the conditions for learnability in the InfraBayes case are still pretty wide open.
    - I don't know how to talk about the CDT vs EDT insight in the InfraBayes world.
        - The way PCDT seems to pathologically fail at learning in Newcomb, and the insight about how we have to learn in order to succeed.
        - Perhaps more importantly, the Troll Bridge insights. As I mentioned in the beginning, in order to meaningfully solve Troll Bridge, it's necessary to "respect logic" in the right sense. InfraBayes doesn't do this, and it's not clear how to get it to do so.

Conclusion

Provided the formal stuff works out, this might be "all there is to know" about counterfactuals from a purely decision-theoretic perspective.

This wouldn't mean we're done with embedded agent theory. However, I think things factor basically as follows:

- Decision Theory
    - Counterfactuals
        - Classic newcomb's problem.
        - 5&10.
        - Troll Bridge.
        - Death in Damascus.
        - ...
    - Logical Updatelessness
        - XOR Blackmail
        - Transparent Newcomb
        - Parfit's Hitchhiker
        - Counterfactual Mugging
        - ...
    - Multiagent Rationality
        - Prisoner's Dilemma
        - Chicken
        - ...

I've expressed many reservations about logical updatelessness in the past, and it may create serious problems for multiagent rationality, but it still seems like the best hope for solving the class of problems which includes XOR Blackmail, Transparent Newcomb, Parfit's Hitchhiker, and Counterfactual Mugging.

If the story about counterfactuals in this post works out, and the above factoring of open problems in decision theory is right, then we'd "just" have logical updatelessness and multiagent rationality left.