Logical decision theories

Main 2 | Intro Economists 3 | Intro Computer Sci... 6 | Intro Philosophers 4 | Loose Intro Every... 6

The work on this page is just getting started. Its content is still in a rough state.

Say what? Decision Theory domain Go faster

Introduction to Logical Decision Theory for Computer Scientists

Relies on: Ability to read algebra

Teaches: Logical decision theories, Causal decision theories, Evidential decision theories

read until "evidential vs counterfactual reasoning"

Summary

Almost-all decision theories agree on using the notion of 'expected utility' as the foundation of agent definitions and rational choice. Decision theories differ on exactly how to calculate the expectation--the probability of an outcome, conditional on an action.

This foundational difference bubbles up to real-life questions about whether to vote in elections, or accept a lowball offer at the negotiating table. When you're thinking about what happens if you don't vote in an election, should you calculate the expected outcome as if *only your* vote changes, or as if *all the people sufficiently similar to you* would also decide not to vote?

Questions like these belong to a larger class of problems, Newcomblike decision problems, in which some other agent is similar to us or reasoning about what we will do in the future. The central principle of 'logical decision theories', several families of which will be introduced, is that we ought to choose as if we are controlling the logical output of our abstract decision algorithm.

Newcomblike considerations--which might initially seem like unusual special cases--become more prominent as agents can get higher-quality information about what algorithms or policies other agents use: Public commitments, machine agents with known code, smart contracts running on Ethereum. Newcomblike considerations also become more important as we deal with agents that are very similar to one another; or with large groups of agents that are likely to contain high-similarity subgroups; or with problems where even small correlations are enough to swing the decision.

In philosophy, the debate over decision theories is seen as a debate over *the principle of rational choice.* Do 'rational' agents refrain from voting in elections, because their one vote is very unlikely to change anything? Do we need to go beyond 'rationality', into 'social rationality' or 'superrationality' or something along those lines, in order to describe agents that could possibly make up a functional society? Do rational agents sometimes wistfully wish they were irrational? Some decision theories have been accused of going into infinite loops on Newcomblike problems; can the rational choice ever be undefined? (Logical decision theorists reply 'No'.) If you were building a decision theory into a machine intelligence, which decision theory would you use and what would be the consequences of building an agent like that?

This overview first covers the Prisoner's Dilemma, one of the classic Newcomblike problems that is basic to game theory and to coordination problems in economics. It talks about the Prisoner's Dilemma game between two agents that know each other's source code; in order to motivate the next section, which overviews the motivation for the most currently accepted decision theory, and introduces the new notion of 'logical decision theory' or agents that choose as if controlling the logical outputs of their algorithms. This overview then covers some of the more philosophical arguments that have been brought to bear on what we should consider to be the principle of rational choice, and quickly summarizes some of the pragmatic consequences. At the end are paths for further reading.

The overall academic status of logical decision theory is 'a new challenger to the currently dominant decision theory, but most people have never heard of it, although this is starting to change'.

The Prisoner's Dilemma

a chance to betray the other (Defect); someone who Defects gets one year off their own prison sentence, but adds two years onto the other person's prison sentence. Alternatively, you can Cooperate with the other prisoner by remaining silent.

So if (o_1, o_2) is the outcome for Player 1 and Player 2 respectively, the outcome matrix for the classical Prisoner's Dilemma is:

	Player 2 Defects:	**Player 2 Cooperates:**
Player 1 Defects:	**(2 years, 2 years)**	**(0 years, 3 years)**
Player 1 Cooperates:	**(3 years, 0 years)**	**(1 year, 1 year)**

Rewriting this as a game with moves D and C, and positive payoffs where $X denotes "X utility":

	D_2	C_2
D_1	$(\$1, \$1)$	$(\$3, \$0)$
C_1	$(\$0, \$3)$	$(\$2, \$2)$

On currently standard decision theory, it is said to be 'rational' (we'll go into the debate over this term shortly) for a selfish player to play Defect in the Prisoner's Dilemma. This is bothersome because if two 'rational' players play Defect, it leads to the outcome ($1, $1) which is Pareto dominated by the outcome ($2, $2) which both players would prefer. It seems like there is a way for both parties to do better, which, on the standard theory, two rational agents cannot manage to obtain for themselves.

There have been many objections to the setup that yields this bothersome conclusion. For example: "In the real world, you would have some fellow-feeling for your fellow conspirator, and negatively value their prison sentence; real people aren't entirely selfish." Or, "Defecting in the Prisoner's Dilemma would ruin your reputation and maybe lead to reprisals after you got out of prison." We can try to construct a True Prisoner's Dilemma by modifying the circumstances so as to address these objections: for example, two charities which each firmly believe their mission to be more important, which both know some detrimental facts about the other charity, and which are both in private, secret meetings with a philanthropist deciding how to divide funding between them.

In the Iterated Prisoner's Dilemma (IPD), we play the Prisoner's Dilemma against the same agent 100 times in a row. On the IPD a famous strategy, Tit for Tat, does whatever the other player did on the previous round, and begins by Cooperating. This strategy usually does extremely well in tournaments of bots, even when playing against much more complicated programs.

Suppose two 'rational' agents (that is, rational according to currently standard decision theory) play the Iterated Prisoner's Dilemma with a known time horizon - both agents know the game stops after the 100th round. Then on the 100th round, both agents will Defect, since this move has no future consequences. On the 99th round, both agents, knowing the other will Defect regardless on the 100th round, will also Defect. By induction, two agents with definite common knowledge that both agents are 'rational' (as opposed to the other agent possibly being Tit for Tat instead), will Defect against each other on all rounds.

Even accepting the basic setup and its outcome matrix, a number of philosophers and computer scientists have had trouble accepting that rational agents must defect in the Prisoner's Dilemma--let alone that agents with common knowledge of each other's rationality must spend 100 rounds Defecting against each other on the IPD. Douglas Hofstadter suggested that 'superrationality' would include taking into account that the other agent was reasoning in a situation very similar to yours, and that the two of you were likely to arrive at similar rational answers, whatever the rational answer was.

Prisoner's Dilemma with common knowledge of code

Imagine a Prisoner's Dilemma tournament in which submitted bots compete against each other in the *one-shot* Prisoner's Dilemma, but with *knowledge of the other bot's code.* [1] This is the "program equilibrium" problem posed by Tennenholtz in 2010.

At minimum (as Tennenholtz observed) you can do better in this tournament than by defecting *all* the time: you can notice when the other bot's code is an exact copy of your own code, and if so, cooperate. But is it possible to do even better?

Suppose you find yourself facing a bot like this:

```
def FairBot1(otherAgent):
   if (otherAgent(FairBot1) == Cooperate):
     return Cooperate
   else:
     return Defect
```

FairBot1 itself, despite the resemblance to Tit for Tat, is definitely not an optimal player in the Prisoner's Dilemma tournament we just defined. One reason is that FairBot1 cooperates with CooperateBot.

```
def CooperateBot(otherAgent):
  return Cooperate
```

(It may help to visualize CooperateBot as a stone with the word "cooperate" written on it, to help pump the intuition that you ought to defect when there's only a stone on the other side.)

One also observes that when FairBot1 plays against a copy of itself, both agents go into an infinite loop and never return a value.

The following algorithm for fixing the infinite-loop problem may sound unrealistic at first, but it can run in a much shorter time than one might expect.

First, pick a simple proof system, such as first-order arithmetic. We'll dub this proof system $\mathcal{T}$ for 'theory'.

Denote "$\mathcal{T}$ proves the quoted sentence S" by writing $\mathbf{Prov}(\mathcal{T}, \ulcorner S \urcorner)$.

Then rewrite Fairbot2 to say:

```
def Fairbot2(otherAgent):
  if (Prov(T, "otherAgent(Fairbot2) == Cooperate")):
    return Cooperate
  else:
    return Defect
```

...that is, "I cooperate, if I prove the other agent cooperates."

Surprisingly, this leads to no infinite regress! You might think that it seems equally consistent to suppose either "Two Fairbot2s both defect, and therefore both fail to prove the other cooperates" or "Both FairBot2s cooperate, and therefore both prove the other cooperates" with the first equilibrium seeming more likely because of the chicken-and-egg problem. Actually, a strange little rule called Löb's theorem implies that both agents prove Cooperation and therefore both Cooperate.

Even more surprisingly, when we have complex systems of "A is true if B is provable and C is not provable" plus "C is true if it's provable that the consistency of $\mathcal{T}$ implies B and D" etcetera, we can compute exactly what is and isn't provable in polynomial time.

The upshot is that if we have complicated systems of agents that do X if they can prove other agents would do Y if they were playing against Z, etcetera, we can evaluate almost immediately what actually happens.

One of the early papers in logical decision theory, "Robust Cooperation in the Prisoner's Dilemma: Program Equilibrium Through Provability Logic" exhibited a proof (with running code) that there existed a simple agent PrudentBot which:

- Cooperated with another PrudentBot.
- Cooperated with FairBot2.
- Defected against CooperateBot (aka a rock with the word "Cooperate" written on it).
- Was not exploitable (PrudentBot never plays Cooperate when the other player plays Defect).

...and it was later shown that PrudentBot was a special case of something that looks like a more general, unified decision-making rule, Proof based decision theory, which in turn is one variant of a Logical decision theory.

(This intro will not dive into the detail of why Löb's theorem is true or how to evaluate Modal agents, but see the intro on Proof based decision theory referenced at bottom.)

Different calculations of expected utility

You may have heard it asserted that it is not 'rational' to vote in elections, since your own vote has only a very tiny probability of changing the election outcome.

When people talk about 'rationality' in this sense, they're appealing to a decision-theory formulation called 'causal decision theory' (aka CDT), the closest thing there is to a currently standard formulation of rationality. CDT says that the principle of rational choice is to calculate expected utility according to the causal consequences of your physical act (in a formal sense we'll define shortly). Unless your voting *causes* many other people to vote, as opposed to all of you just being in *similar situations,* CDT says that your vote is unlikely to cause a change in the election outcome.

Elections are one of the cases where LDT makes a big difference in what appears rational or normative, compared to classical CDT. Populations voting in elections are large enough that there may be many other people with decision algorithms similar to yours - not everyone, but a logically correlated cohort large enough that it might matter. You can reasonably believe that everyone in your cohort will probably decide to vote (or not vote) in something like unison, and ask whether the probable benefit of the whole cohort voting is worth the cost of the whole cohort voting.

The debate over Newcomblike decision problems turns out to revolve around the question of how to formally define "The probability of an outcome, *conditioned on* choice X" within calculations of expected utility.

Evidential versus counterfactual conditioning

Almost everyone in the field of decision theory agrees that *some* version of expected utility is central to rationality; there are whole rafts of coherence theorems showing that various forms of bizarre vengeance are visited upon agents whose behavior *can't* be viewed as consistent with some set of probabilistic beliefs and a utility function.

If you've read a textbook discussing the notion of 'expected utility', it probably defined it as something like:

$$\mathbb{E}[\mathcal{U}|a_x] = \sum_{o_i \in \mathcal{O}} \mathcal{U}(o_i) \cdot \mathbb{P}(o_i|a_x)$$

where

- $\mathbb{E}[\mathcal{U}|a_x]$ is our average expectation of utility, if action a_x is chosen;
- $\mathcal{O}$ is the set of possible outcomes;
- $\mathcal{U}$ is our utility function, mapping outcomes onto real numbers;
- $\mathbb{P}(o_i|a_x)$ is the conditional probability of outcome o_i if a_x is chosen.

This formula is almost universally agreed to be wrong.

The problem is the use of evidential conditioning in $\mathbb{P}(o_i|a_x)$. On this formula we are behaving as if we're asking, "What would be my revised probability for $\mathbb{P}(o_i)$, if I was *told the news* or *observed the evidence* that my action had been a_x?"

Causal decision theory (the current standard) says we should instead use the *counterfactual conditional* $\mathbb{P}(a_x \mathbin{\square\!\!\rightarrow} o_i)$.

The difference between evidential and counterfactual conditioning is standardly contrasted by these two sentences:

- If Lee Harvey Oswald didn't shoot John F. Kennedy, somebody else did.
- If Lee Harvey Oswald hadn't shot John F. Kennedy, somebody else would have.

In the first sentence, we're being told as news that Oswald didn't shoot Kennedy, and updating our beliefs accordingly to match the world we already saw. In the second world, we're imagining how a counterfactual world would have played out if Oswald had acted differently.

If K denotes the proposition that somebody else shot Kennedy and O denotes the proposition that Oswald shot him, then the first sentence and second sentence are respectively talking about:

- $\mathbb{P}(K|\neg O)$
- $\mathbb{P}(\neg O \mathbin{\square\!\!\rightarrow} K)$

Calculating expected utility using evidential conditioning is widely agreed to lead to an irrational policy of 'managing the news'. For example, suppose that toxoplasmosis, a parasitic infection carried by cats, can cause toxoplasmosis-infected humans to become fonder of cats. [2]

You are now faced with a cute cat that has been checked by a veterinarian who says this cat definitely does *not* have toxoplasmosis. If you decide to pet the cat, an impartial observer watching you will conclude that you are 10% more likely to have toxoplasmosis, which can be a fairly detrimental infection. If you don't pet the cat, you'll miss out on the hedonic enjoyment of petting it. Do you pet the cat?

Most decision theorists agree that in this case you should pet the cat. Either you already have toxoplasmosis or you don't. Petting the cat can't *cause* you to acquire toxoplasmosis. You'd just be missing out on the pleasant sensation of cat-petting.

Afterwards you may update your beliefs based on observing your own decision, and realize that you had toxoplasmosis all along. But when you're considering the consequences of actions, you should reason that *if counterfactually* you had not pet the cat, you *still* would have had toxoplasmosis *and* missed out on petting the cat. (Just like, if Oswald *hadn't* shot Kennedy, nobody else would have.)

theory. Evidential decision theory answers the central question "How do I condition on my choices?" by replying "Condition on your choices as if observing them as evidence" or "Take the action that you would consider best if you heard it as news."

(For a more severe criticism of evidential decision theory showing how more clever agents can pump money out of evidential decision agents, see the Termite Dilemma.)

Causal counterfactuals

Causal decision theory, the current academic standard, says that the expected utility formula should be written:

$$\mathbb{E}[\mathcal{U}|a_x] = \sum_{o_i \in \mathcal{O}} \mathcal{U}(o_i) \cdot \mathbb{P}(a_x \mathrel{\square\!\!\rightarrow} o_i)$$

This leads into the question of how we compute $\mathbb{P}(a_x \mathrel{\square\!\!\rightarrow} o_i)$.

In the philosophical literature, it's often assumed that we intuitively know what the counterfactual results must be. (E.g., we're just taking for granted that you somehow know that if Oswald hadn't shot Kennedy, nobody else would have; this is intuitively obvious.) This is formalized by having a conditional distribution $\mathbb{P}(\bullet \mid\mid \bullet)$ which is treated as heaven-sent and includes all counterfactual conditionals.

If we're not happy to leave it at that (and we shouldn't be), we can import the theory of causal models developed by Judea Pearl et. al. The theory of causal models formally states how to perform counterfactual surgery on graphical models of causal processes.

Formally, we have a directed acyclic graph such as:

- X_1 -> {X_2, X_3} -> X_4 -> X_5

One of Judea Pearl's examples of such a causal graph is:

- SEASON -> {RAINING, SPRINKLER} -> {SIDEWALK} -> {SLIPPERY}

This says, e.g.:

- That the current SEASON affects the probability that it's RAINING, and separately affects the probability of the SPRINKLER turning on. (But RAINING and SPRINKLER don't affect each other; if we know the current SEASON, we don't need to know whether it's RAINING to figure out the probability the SPRINKLER is on.)
- RAINING and SPRINKLER can both cause the SIDEWALK to become wet. (So if we did observe that the sidewalk was wet, then even already knowing the SEASON, we would estimate a different probability that it was RAINING depending on whether the SPRINKLER was on. The SPRINKLER being on would 'explain away' the SIDEWALK's observed wetness without any need to postulate RAIN.)
- Whether the SIDEWALK is wet is the sole determining factor for whether the SIDEWALK is SLIPPERY. (So that if we *know* whether the SIDEWALK is wet, we learn nothing more about the probability that the path is SLIPPERY by being told that the SEASON is summer. But if we didn't already know whether the SIDEWALK was wet, whether the SEASON was summer or fall might be very relevant for guessing whether the path was SLIPPERY!)

A causal model goes beyond the graph by including specific probability functions $\mathbb{P}(X_i|\mathbf{pa}_i)$ for how to calculate the probability of each node X_i taking on the value x_i given the values $\mathbf{pa}_i$ of x_i's immediate ancestors. It is implicitly assumed that the causal model factorizes, so that the probability of any value assignment $\mathbf{x}$ to the whole graph can be calculated using the product:

$$\mathbb{P}(\mathbf{x}) = \prod_i \mathbb{P}(x_i|\mathbf{pa}_i)$$

Then the counterfactual conditional $\mathbb{P}(\mathbf{x}|\operatorname{do}(X_j = x_j))$ is calculated via:

$$\mathbb{P}(\mathbf{x}|\operatorname{do}(X_j = x_j)) = \prod_{i \neq j} \mathbb{P}(x_i|\mathbf{pa}_i)$$

(We assume that $\mathbf{x}$ has x_j equaling the $\mathbf{do}$-specified value of X_j; otherwise its conditioned probability is defined to be 0.)

This is actually pretty straightforward; it just says that when we set $\mathbf{do}(X_j = x_j)$ we ignore the ordinary parent nodes for X_j and just say that whatever the values of $\mathbf{pa}_j$, the probability of $X_j = x_j$ is 1.

same way that (ordinarily) we think our choices today affect how much money we have tomorrow, but not how much money we had yesterday.

Then expected utility should be calculated as:

$$\mathbb{E}[\mathcal{U}|\operatorname{do}(a_x)] = \sum_{o_i \in \mathcal{O}} \mathcal{U}(o_i) \cdot \mathbb{P}(o_i|\operatorname{do}(a_x))$$

Under this rule, we won't calculate that we can affect the probability of having toxoplasmosis by petting the cat, since our choice to pet the cat is causally downstream of whether we have toxoplasmosis.

Tickles and infinite loops

One class of *technical* problems with causal decision theory comes from how, in the case of the Toxoplasmosis Dilemma, we arrive at the correct qualitative decision but the wrong quantitative answer for expected utility.

Suppose the prior probability of having toxoplasmosis is 10%, and the posterior probability after being seen to pet the cat is 20%. Suppose that *not* having toxoplasmosis has $100 utility; that having toxoplasmosis has $0 utility; and that, given the amount you enjoy petting cats, petting the cat adds $1 of utility to your outcome.

Then the above formula for deciding whether to pet the cat suggests that petting leads to an expected utility of $91, and not petting leads to an expected utility of $90. This tells us to pet the cat, which is the correct decision, but it also tells us to expect $91 of expected utility after petting the cat, where we actually receive $81 in expectation. It seems like the intuitively "correct" answer is that we should calculate $81 of utility for petting the cat and $80 utility for not petting it.

You might initially be tempted to solve this problem by doing the calculation in phases:

- Phase 1: Calculate the decision based on prior beliefs.
- Phase 2: Update our beliefs based on having observed our emerging preference - we notice the 'tickle' of an impulse to decide a particular way.
- Phase 3: Recalculate the expected utilities based on the posterior beliefs, possibly picking a new action; then goto 2.

...and then wait for this algorithm to settle into a consistent state.

Besides lacking the efficiency and elegance of computing our decision in one swoop, it seems possible for an agent like this to go into an infinite loop. There have been proposals to break this infinite loop by randomizing actions. But then we again end up calculating the wrong outcome probabilities conditional on the action we actually take, and this can lead into other seemingly stupid decisions. (See here.)

Newcomb's Problem and Parfit's Hitchhiker

The academic debate on decision theory revolved mainly around *Newcomblike decision problems,* a broad class which includes the Prisoner's Dilemma, voting in elections, coordinating with other agents who are reasoning about you, and so on.

Roughly, we could describe Newcomblike problems as those where somebody similar to you, or trying to predict you, exists in the environment. In this case your decision can *correlate* with events outside you, without your action *physically causing* those events.

The original Newcomb's Problem was rather artificial, but is worth recounting for historical reasons:

An alien named Omega presents you with two boxes, a transparent box A containing $1,000, and an opaque Box B. Omega then flies away, leaving you with the choice of whether to take only Box B ('one-box') or to take Box A plus Box B ('two-box'). Omega has put $1,000,000 in Box B if and only if Omega predicted that you would take only one box; otherwise Box B is empty.

Omega has already departed, so Box B is already empty or already full.

Omega is an excellent predictor of human behavior; e.g., we can suppose Omega has been observed to run this experiment 73 times and to be right in its predictions every time. Newcomb's Problem originally stipulated that if you try to decide by flipping a coin, Omega leaves Box B empty; we could alternatively suppose that Omega can predict the coinflip.

Do you take both boxes, or only Box B?

- Argument 1: People who take only Box B tend to walk away rich. People who two-box tend to walk away poor. It is better to be rich than poor.
- Argument 2: Omega has already made its prediction. Box B is already empty or already full. It would be irrational to leave

Box B is now *already empty*, and irrationally leaving Box A behind would just counterfactually result in your getting $0 instead of $1,000.

Newcomb's Problem is conventionally seen as an example that splits the verdict of evidential decision theory ("Taking only Box B is good news! Do that.") versus causal decision theory ("Taking both boxes does not *cause* Box B to be empty, it just adds $1,000 to the reward") in a way that initially seems more favorable to evidential decision agents (who walk away rich).

The setup in Newcomb's Problem may seem contrived, supporting the charge that Omega is merely rewarding people with irrational dispositions. But consider the following variant, Parfit's Hitchhiker:

You are lost in the desert, your water bottle almost exhausted, when somebody drives up in a lorry. This driver is (a) entirely selfish, and (b) very good at detecting lies. (Maybe the driver went through Paul Ekman's training for reading facial microexpressions.)

The driver says that they will drive you into town, but only if you promise to give them $1,000 on arrival.

If you value your life at $1,000,000 (pay $1,000 to avoid 0.1% risks of death) then this problem is nearly isomorphic to Gary Drescher's *transparent Newcomb's Problem,* in which Box B is transparent, and Omega has put a visible $1,000,000 into Box B iff Omega predicts that you one-box when seeing a full Box B. This makes Parfit's Hitchhiker a *Newcomblike problem,* but one in which the driver's behavior seems selfishly sensible, and not at all contrived as in the case of Newcomb's Omega.

On reaching the city in Parfit's Hitchhiker, you might be tempted to reason that the car has already driven you there, and so, when you *now* make the decision in your selfishness, you will reason that you are better off by $1,000 *now* if you refuse to pay, since your decision can't alter the past. Likewise in the transparent Newcomb's Problem; Box B already seems visibly full, so the money is right there and it can't vanish if you take both boxes, right? But if you are a sort of agent that reasons like this, Box B is already empty. Parfit's driver asks you a few hard questions and then drives off to let you die in the desert. (Since you know your own algorithm, your current beliefs about your future decision are likely to tightly correlate with your future decision; the driver, who again is very good at reading faces, also asks you tough questions like "Do you think you're fooling yourself?" and "Are you imagining in detail the scenario where you're already in the city?" or "Are you secretly planning to change your mind later while telling yourself you're not?")

Both causal decision agents and evidential decision agents will two-box on the transparent version of Newcomb's Problem, or be left to die in the desert on Parfit's Hitchhiker. A causal agent who sees a full Box B reasons "I cannot cause Box B to become empty by leaving behind Box A. An evidential agent reasons, "It wouldn't be good news about anything in particular to leave behind Box A; I already *know* Box B is full."

Logical decision theory

The most general form of logical decision theory, "functional decision theory", argues that a logical agent ought to calculate expected utility as follows:

$$\mathbf{Q}(s) = \left(\underset{\pi_x \in \mathbf{\Pi}}{argmax} \sum_{o_i \in \mathcal{O}} \mathcal{U}(o_i) \cdot \mathbb{P}(\ulcorner \mathbf{Q} = \pi_x \urcorner \triangleright o_i)\right)(s)$$

Where:

- $\mathbf{Q}$ is the agent's current decision algorithm - that is, the whole calculation presently running.
- s is the agent's sense data.
- $\pi_x \in \mathbf{\Pi}$ is the output of $\mathbf{Q}$, a *policy* that maps sense data to actions.
- $\ulcorner \mathbf{Q} = \pi_x \urcorner$ is the proposition that, as a logical fact, the output of algorithm $\mathbf{Q}$ is π_x.
- $\mathbb{P}(X \triangleright o_i)$ is the probability of o_i *conditioned* on the logical fact X.

Functional decision theory is formally incomplete because nobody has a full, formal specification for how to calculate the logical conditioning operator $X \triangleright Y$. The only complete logical decision theories are those which calculate special cases of $X \triangleright Y$:

- Proof based decision theory treats X as a premise introduced into a standard first-order logical system, and says that $X \triangleright Y$ in some possible world if Y can be derived after adding the assumption X.
- Timeless decision theory takes a decision setup in the form of a Pearl-style standard causal model, some of whose nodes are intended to denote logical propositions (and particularly the proposition $\mathbf{Q}$. $\mathbb{P}(X \triangleright Y)$ is then just $\mathbb{P}(Y | \operatorname{do}(X))$.

These special cases respectively suffice for:

- Formalizing Modal agents, and letting us formally simulate agents with common knowledge of each other's code negotiating on the Prisoner's Dilemma and other game-theoretic setups.

We know these two formalization styles aren't complete because:

- Proof-based decision theory has weird edge cases indicating that it only correctly formalizes the intuitive notion of logical conditioning some of the time.
- We don't have a general algorithm for *building* causal models that include logical facts, if we're not giving them to the agent as manna from heaven. It's also not obvious that the probability rules for causal models can usefully model logical correlations more complicated than "run this exact same algorithm in two different places".

In any setup we do understand how to represent, a logical decision agent cheerfully promises to pay the $1,000 to Parfit's driver (and then actually pays). They also cheerfully leave behind Box A in the original Newcomb's Problem. An LDT agent is choosing the best output for their algorithm, and reasoning, "If my algorithm had output 'don't pay' / 'take both boxes', then this would have implied my dying in the desert / Box B being empty."

LDT as the principle of rational choice

The general family of logical decision theories claims to embody *the principle of rational choice*--or at least, embody it better than causal decision theory or any currently known alternative. There remain open problems like a truly general formulation of "logical counterfactuals" / "logical conditioning". But the claim is that *some sort of conditioning on the logical output of your algorithm*, compared to classical causal decision theory, now seems a much more likely candidate for the principle of rational choice--that is, when we have a fully specified principle of rational choice, it's much more likely to look like it belongs somewhere in the family of logical decision theories, rather than belonging to the causal decision family that takes both boxes in Newcomb's problem.

Some of the arguments brought to bear on the philosophy of rationality include:

Argument from fairness of Newcomblike problems

The first and foremost motivation behind LDT is that LDT agents systematically end up rich. The current literature contains rich veins of discourse about "one-boxers" on Newcomb's Problem asking two-boxers "Why aincha rich, if you're so rational?" and various retorts along the lines of "It's not my fault Omega decided to punish people who'd act rationally before I even got here; that doesn't change what my rational choice is now."

That retort may sound less persuasive if we're thinking about Parfit's Hitchhiker instead of Newcomb's Problem. The driver is not making a weird arbitrary choice to punish 'rational' agents. It makes no (selfish) sense to rescue someone you don't predict will pay up afterwards.

If in Newcomb's Problem Omega read the agent's source code and decided to reward only agents with an *algorithm* that output 'one box' *by picking the first choice in alphabetical order,* punishing all agents that behaved in exactly the same way due a different internal computation, then this would indeed be a rigged contest. But in Newcomb's Problem, Omega only cares about the behavior, and not the kind of algorithm that produced it; and an agent can indeed take on whatever kind of behavior it likes; so, according to LDT, there's no point in saying that Omega is being unfair. You can make the logical output of your currently running algorithm be whatever you want, so there's no point in picking a logical output that leaves you to die in the desert.

Argument from freedom of counterfactuals

Within analytic philosophy, the case for causal decision theory rests primarily on the intuition that one-boxing on Newcomb's problem cannot *cause* Box B to be full. Or that on the transparent Newcomb's Problem, with Box B transparently full (or empty), it cannot be reasonable to imagine that by leaving behind Box A and its $1,000 you can *cause* things to be different.

Is it not in some sense *true,* after Parfit's driver has conveyed the LDT agent to the city in the desert, that in the counterfactual world where the LDT agent does not at that time choose to pay, they remain in the city? In this sense, must not an LDT agent be deluded about some question of fact, or act as if it is so deluded?

The LDT agent responds:

- There aren't any actual "worlds where I didn't pay" floating out there. The only *real* world is the one where my decision algorithm had the output it actually had. To imagine other worlds is an act of imagination, computing a description of a certain world that doesn't exist; there's no corresponding world floating out there, so my description of that impossibility can't be true or false under a correspondence theory of truth. "Counterfactuals were made for humanity, not humanity for counterfactuals". That being the case, I can decide to condition on the actions my algorithm doesn't take in whatever way produces the greatest wealth. I am free to say "In the nonexistent world where I don't pay now, I already died in the desert".
- I don't one-box in Newcomb's Problem *because* I think it physically causes Box B to be full. I one-box in Newcomb's Problem because I have computed this as the optimal output in an entirely different way. It begs the question to assume

To this reply, the LDT agent adds that it *is* desirable for the action-conditionals we compute to match the one real world. That is, if it's a fact that your decision algorithm outputs action a_x, then your imagination of "The world where I do a_x" inside that term of the expected utility formula should match reality. Arguendo, CDT agents must either violate this rule, or sometimes go into infinite loops.

Argument from coherence

LDT agents never need to resort to precommitments (since LDT agents never wish their future selves would act differently from the LDT algorithm). LDT agents always calculate a positive value of information. Where, e.g., on the transparent Newcomb's Problem, an evidential decision agent might beg you to *not* render Box B transparent, since then the evidential agent will choose both boxes and therefore box B will be revealed to be empty.

Or similarly: suppose that with 10% probability, Omega is actually still present, and will fill or empty box B *after* you make your choice. In this case, all agents take only box B (of course!) Now suppose that Omega hasn't yet set up the experiment, and you don't have access to any means of precommitment or of forcing your future self to do anything. I threaten to reveal whether or not Omega is going to fill box B before or after the experiment. As a causal decision agent, you will pay me thousands of dollars *not* to reveal this decision to you.

A CDT agent will also, given the chance to set up before Omega's arrival, pay a $10 fee to have a precommitment assistant stand around with a gun threatening to shoot them if they take both boxes in Newcomb's Problem. Naturally, given the chance by surprise, the same agent would *later* (after Omega's departure) pay $10 to make the gun-toter go away. [3]

From the standpoint of LDT, these anomalies reveal a kind of dynamic inconsistency or reflective inconsistency in EDT and CDT. If Omega is setting up a Newcomb's Problem at 8am, the CDT agent at 7am and the CDT agent making the choice at 9am have different preferences for how many boxes they wish the CDT agent would take at 9am. The EDT agent in the desert, and the EDT agent once it has actually arrived in the city, have different preferences for what the EDT agent ought to do in the city.

The upshot is that both EDT agents and CDT agents will pay to force their future selves to do things, or alter their own source code if they can. This is one reason why LDT was originally studied by researchers looking into the theory of reflectively stable self-modifying agents.

One may, of course, try to retort that all Newcomblike problems are just weird special cases (including voting in elections?)--but it's not clear why we should, considering that on *non*-Newcomblike problems, EDT and CDT and LDT all give the same answer. In this sense, LDT seems to dominate by performing strictly better on the larger problem class, and equally well on the smaller problem class.

We might also just say, "Well, there's no reason to *expect* that the principle of rational choice ought to be reflectively coherent--who says that rational agents always want their future selves to be rational?" The modern economics literature so takes for granted 'precommitments' and 'usefully irrational' behavior at the bargaining table, that it may sound odd to claim that dynamic consistency and reflective consistency are desirable properties for the principle of rational choice to have. There are just two entirely different subjects of study, the study of 'What is rational behavior', and the study of 'What rationalists wish they behaved like' or rationality as modified by precommitments, 'useful irrationality', 'social rationality', 'superrationality', etcetera.

But it would nonetheless be a remarkable fact if the study of 'the dispositions rational agents wish they had' pointed strongly in the direction of a simple general algorithm, with philosophically appealing properties, that happened to be much more self-consistent in various ways, whose corresponding agents stand around saying things like "Why on Earth would a rational agent wistfully wish that their future selves would be irrational?" or "What do you mean *useful irrationality?* If a choice pattern is useful, it's not irrational!" [4]

A truly pure causal decision agent, with no other thoughts but CDT, will wave off all that argument with a sigh; you can't alter what Fairbot2 has already played in the Prisoner's Dilemma and that's that. But if we actual humans let ourselves blank our minds of our previous thoughts and try to return to an intuitive, pretheoretic standpoint, we might suspect from looking over this situation that we have made a mistake about what to adopt as our explicit theory of rationality.

Pragmatic consequences (overview)

A very quick overview of some of the pragmatic consequences of adopting logical decision theory:

⚫ The rare textbook that goes into enough detail to present causal decision theory instead of just using the simple wrong formulation of expected utility, ought to present the equation for functional decision theory instead.

⚫ Various economic and social dilemmas that involve *people similar to you* or *people who might guess how you're thinking or people trying to gain a public reputation for acting a certain way* will come out to different answers. See e.g. the analysis of

⚫ We ought to stop going around telling people that two rational agents defect at each other in the Prisoner's Dilemma. LDT isn't a panacea and two LDT agents might end up defecting against each other because, e.g., I'm sure you're an LDT agent but I'm not sure you *know* I'm an LDT agent. But on, e.g., the Iterated Prisoner's Dilemma with a known time horizon, even a small chance that the other agent is also LDT-rational is enough to try playing Cooperate on the first round.

⚫ Similarly, we can stop going around saying that rational agents don't vote, or that rational agents give in to blackmail, or that we need 'social rationality' instead of plain old rationality in order for society to cohere, or that there's such a thing as being 'usefully irrational' at the negotiating table, and so on. LDT generally calculates equilibria on economic dilemmas of this sort such that there never seems to be a visible advantage to being less rational.

⚫ We can expect agents that know each other's code, or can calculate out the probable cases for each other's code, to be extremely good at coordinating with *each other*. This might apply to e.g. machine intelligences or to future Distributed Autonomous Organizations.

⚫ Future work on self-modifying agents should start with LDT agents rather than CDT (or EDT) agents.

Further reading

You've now seen an overview of many aspects of logical decision theories. From here, you can go on to read about:

- How proof-based agents work, and why we can simulate them in polynomial time. (a)
  - How to derive game theory from expected utility maximization and agents reasoning about each other.
  - Open questions in the modal agents formalism.
- How LDT actually handles Newcomblike problems.
  - Updateless decision theory; how to not go into infinite loops. (b)
  - Timeless decision theory; how and why to factor logical uncertainty into causal models. (c)
  - A bestiary of Newcomblike problems and how to represent and reason about them using LDT. (requires b, c)
    - Negotiations and 'fairness', in the Ultimatum Game and elsewhere. (d)
    - The problem of the no-blackmail equilibrium. (e)
- The philosophical debate about Newcomblike problems and principles of rationality.
  - Why LDT is more reflectively and dynamically coherent than CDT. (f)
  - Controlling something does not require changing it - why it makes sense to talk about controlling the output of a logical algorithm.
- A brief history of LDT and who invented what. (requires a, b, c, d, f)
- Why LDT matters to issues of sufficiently advanced machine intelligence. (requires b, d, e, f)
- Ldt citations.

Parents: Logical decision theoriesChildren: none

20changes by4 authors
4119 views

6 PROPOSE EDIT

Learn more

1 Death in Damascus — Teaches: Logical dec / Relies on: Causal dec
Death tells you that It is coming for you tomorrow. You can stay in Damascus or flee to Aleppo. Whichever decision you actually make is the wrong one. This gives some

3 Parfit's Hitchhiker — Teaches: Logical decisior
You are dying in the desert. A truck-driver who is very good at reading faces finds you, and offers to drive you into the city if you promise to pay $1,000 on arrival. Yo

0 Toxoplasmosis dilemma — Relies on: Logical decision theories, Causal decision t
A parasitic infection, carried by cats, may make humans enjoy petting cats more. A kitten, now in front of you, isn't infected. But if you *

0 Ultimatum Game — Relies on: Logical decision theories
A Proposer decides how to split $10 between themselves and the Responder. The Responder can take what is offered, or refuse, in which case both p

0 Absent-Minded Driver dilemma — Relies on: Causal decisi
A road contains two identical intersections. An absent-minded driver wants to turn right at the second intersection. "With what probability should the driver turn right?

0 Transparent Newcomb's Problem — Relies on: Evidential decision th
Omega has left behind a transparent Box A containing $1000, and a transparent Box B containing $1,000,000 or nothing. Box B is full iff Omega thinks you on

WATCH DISCUSSION 1

PROPOSE COMMENT