

New paper: “Cheating Death in Damascus”

March 18, 2017 | Rob Bensinger (<https://intelligence.org/author/robby/>) | Papers (<https://intelligence.org/category/papers/>)

(<https://intelligence.org/files/DeathInDamascus.pdf>) MIRI Executive Director Nate Soares and Rutgers/UIUC decision theorist Ben Levinstein have a new paper out introducing *functional decision theory* (FDT), MIRI’s proposal for a general-purpose decision theory.

The paper, titled “**Cheating Death in Damascus** (<https://intelligence.org/files/DeathInDamascus.pdf>),” considers a wide range of decision problems. In every case, Soares and Levinstein show that FDT outperforms all earlier theories in utility gained. The abstract reads:



Evidential and Causal Decision Theory are the leading contenders as theories of rational action, but both face fatal counterexamples. We present some new counterexamples, including one in which the optimal action is causally dominated. We also present a novel decision theory, Functional Decision Theory (FDT), which simultaneously solves both sets of counterexamples.

Instead of considering which physical action of theirs would give rise to the best outcomes, FDT agents consider which output of their decision function would give rise to the best outcome. This theory relies on a notion of subjunctive dependence, where multiple implementations of the same mathematical function are considered (even counterfactually) to have identical results for logical rather than causal reasons. Taking these subjunctive dependencies into account allows FDT agents to outperform CDT and EDT agents in, e.g., the presence of accurate predictors. While not necessary for considering classic decision theory problems, we note that a full specification of FDT will require a non-trivial theory of logical counterfactuals and algorithmic similarity.

“Death in Damascus” is a standard decision-theoretic dilemma. In it, a trustworthy predictor (Death) promises to find you and bring your demise tomorrow, whether you stay in Damascus or flee to Aleppo. Fleeing to Aleppo is costly and provides no benefit, since Death, having predicted your future location, will then simply come for you in Aleppo instead of Damascus.

In spite of this, causal decision theory often recommends fleeing to Aleppo — for much the same reason it recommends defecting in the one-shot twin prisoner’s dilemma (<https://intelligence.org/2016/03/31/new-paper-on-bounded-lob/>) and two-boxing in Newcomb’s problem (http://lesswrong.com/lw/gu1/decision_theory_faq/#newcomblike-problems-and-two-decision-algorithms). CDT agents reason that Death has already made its prediction, and that switching cities therefore can’t *cause* Death to learn your new location. Even though the CDT agent recognizes that Death is inescapable, the CDT agent’s decision rule forbids taking this fact into account in reaching decisions. As a consequence, the CDT agent will happily give up arbitrary amounts of utility in a pointless flight from Death.

Causal decision theory fails in Death in Damascus, Newcomb’s problem, and the twin prisoner’s dilemma — and also in the “random coin,” “Death on Olympus,” “asteroids,” and “murder lesion” dilemmas described in the paper — because its counterfactuals only track its actions’ causal impact on the world, and not the rest of the world’s causal (and logical, etc.) structure.

While evidential decision theory succeeds in these dilemmas, it fails in a new decision problem, “XOR blackmail.”¹ FDT consistently outperforms both of these theories, providing an elegant account of normative action for the full gamut of known decision problems.

The underlying idea of FDT is that an agent’s decision procedure can be thought of as a mathematical function. The function takes the state of the world described in the decision problem as an input, and outputs an action.

In the Death in Damascus problem, the FDT agent recognizes that their action cannot *cause* Death’s prediction to change. However, Death and the FDT agent are in a sense computing the same function: their actions are correlated, in much the same way that if the FDT agent were answering a math problem, Death could predict the FDT agent’s answer by computing the same mathematical function.

This simple notion of “what variables depend on my action?” avoids the spurious dependencies that EDT falls prey to. Treating decision procedures as multiply realizable functions does not require us to conflate correlation with causation. At the same time, FDT tracks real-world dependencies that CDT ignores, allowing it to respond effectively in a much more diverse set of decision problems than CDT.

The main wrinkle in this decision theory is that FDT’s notion of dependence requires some account of “counterlogical” or “counterpossible” reasoning.

Search

Browse

- All (/all-posts/)
- Analysis (<https://intelligence.org/category/analysis/>)
- Conversations (<https://intelligence.org/category/conversations/>)
- Guest Posts (<https://intelligence.org/category/guest-posts/>)
- MIRI Strategy (<https://intelligence.org/category/miri/>)
- News (<https://intelligence.org/category/news/>)
- Newsletters (<https://intelligence.org/category/newsletter>)
- Papers (<https://intelligence.org/category/papers/>)
- Video (<https://intelligence.org/category/video/>)

Subscribe

Join newsletter subscribers

- Follow @MIRIBerkeley
- RSS (<http://feeds.feedburner.com/miriblog>)

The prescription of FDT is that agents treat their decision procedure as a deterministic function, consider various outputs this function could have, and select the output associated with the highest-expected-utility outcome. What does it mean, however, to say that there are different outputs a deterministic function “could have”? Though one may be uncertain about the output of a certain function, there is in reality only one possible output of a function on a given input. Trying to reason about “how the world would look” on different assumptions about a function’s output on some input is like trying to reason about “how the world would look” on different assumptions about which is the largest integer in the set {1, 2, 3}.

In garden-variety counterfactual reasoning, one simply imagines a different (internally consistent) world, exhibiting different physical facts but the same logical laws. For counterpossible reasoning of the sort needed to say “if I stay in Damascus, Death will find me here” as well as “if I go to Aleppo, Death will find me there” — even though only one of these events is logically possible, under a full specification of one’s decision procedure and circumstances — one would need to imagine worlds where *different logical truths* hold. Mathematicians presumably do this in some heuristic fashion, since they must weigh the evidence for or against different conjectures; but it isn’t clear how to formalize this kind of reasoning in a practical way.²

Functional decision theory is a successor to timeless decision theory (<https://intelligence.org/files/TDT.pdf>) (first discussed in 2009 (http://lesswrong.com/lw/135/timeless_decision_theory_problems_i_cant_solve/)), a theory by MIRI senior researcher Eliezer Yudkowsky that made the mistake of conditioning on observations. FDT is a generalization of Wei Dai’s *updateless decision theory*.³

We’ll be presenting “Cheating Death in Damascus” at the Formal Epistemology Workshop (<http://www.mayowilson.org/FEW.htm>), an interdisciplinary conference showcasing results in epistemology, philosophy of science, decision theory, foundations of statistics, and other fields.⁴

Update April 7: Nate goes into more detail on the interpretive questions raised by functional decision theory in a follow-up conversation: Decisions are for making bad outcomes inconsistent (<https://intelligence.org/2017/04/07/decisions-are-for-making-bad-outcomes-inconsistent/>).

Update November 25, 2019: A revised version of this paper has been accepted to *The Journal of Philosophy* (https://en.wikipedia.org/wiki/The_Journal_of_Philosophy). The *JPhil* version is **here** (<https://intelligence.org/files/DeathInDamascus.pdf>), while the 2017 FEW version is available here (<https://intelligence.org/files/obsolete/DeathInDamascus2019-11-27.pdf>).

Sign up to get updates on new MIRI technical results

Get notified every time a new technical paper is published.

1. Just as the variants on Death in Damascus in Soares and Levinstein’s paper help clarify CDT’s particular point of failure, XOR blackmail drills down more exactly on EDT’s failure point than past decision problems have. In particular, EDT cannot be modified to avoid XOR blackmail in the ways it can be modified to smoke in the smoking lesion problem. ↔

2. Logical induction (<https://intelligence.org/2016/09/12/new-paper-logical-induction/>) is an example of a method for assigning reasonable probabilities to mathematical conjectures; but it isn’t clear from this how to define a *decision theory* that can calculate expected utilities for inconsistent *scenarios*. Thus the problem of reasoning under logical uncertainty is distinct from the problem of defining counterlogical reasoning. ↔

3. The name “UDT” has come to be used to pick out a multitude of different ideas, including “UDT 1.0 (http://lesswrong.com/lw/15m/towards_a_new_decision_theory/)” (Dai’s original proposal), “UDT 1.1 (http://lesswrong.com/lw/1s5/explicit_optimization_of_global_strategy_fixing_a/)”, and various proof-based (<https://agentfoundations.org/item?id=50>) approaches to decision theory (which make useful toy models, but not decision theories that anyone advocates adhering to).

FDT captures a lot (but not all) of the common ground between these ideas, and is intended to serve as a more general umbrella category that makes fewer philosophical commitments than UDT and which is easier to explain and communicate. Researchers at MIRI do tend to hold additional philosophical commitments that are inferentially further from the decision theory mainstream (which concern updatelessness and logical prior probability), for which certain variants of UDT are perhaps our best concrete theories, but no particular model of decision theory is yet entirely satisfactory. ↔

4. Thanks to Matthew Graves and Nate Soares for helping draft and edit this post. ↔

The basic reasons I expect AGI ruin

6 months ago

I've been citing AGI Ruin: A List of Lethalities to explain why the situation with AI ...

Six Dimensions of Operational Adequacy ...

a year ago

Editor's note: The following is a lightly edited copy of a document written by ...

Conversation on technology ...

2 years ago

This post is a transcript of a multi-day discussion between Paul Christiano, ...

Biology-Inspired AGI Timelines: The Trick ...

2 years ago

– 1988 – Hans Moravec: Behold my book Mind Children. Within, I project ...

G Join the discussion...

LOG IN WITH

OR SIGN UP WITH DISQUS ?

Name


Share Best Newest Oldest

 **hrurahaalm** 7 years ago — 🚩
 You mean, "can be modified to smoke"?


1 0 Reply • Share >

 **Rob Bensinger** Mod → hrurahaalm 7 years ago — 🚩
 Yes, fixed. Thanks, hrurahaalm!

0 0 Reply • Share >

 **ViRiX_Dreamcore** 7 years ago — 🚩
 This may sound dumb, but it sounds like using FDT, it just takes a much closer look at all the variables within the situation, but it didn't exactly solve the problem or get out of the situation. In the example, was its goal to escape death? I'll have to look at the other scenerios in the paper to see how they play out.


0 0 Reply • Share >

 **Rob Bensinger** Mod → ViRiX_Dreamcore 7 years ago edited — 🚩
 In decision theory, it's just as important that agents not waste their time trying to solve unsolvable problems (or unduly difficult to solve problems) as it is for agents to not squander opportunities to solve solvable problems. In cases where you're threatened by a predictor and there are cost-effective ways to escape, we'd like agents to escape successfully. Death in Damascus, though, is a case where you're threatened by a predictor and it's assumed that there are no ways to escape.


If you're sufficiently confident that you can't escape Death, then the rational choice is to maximize your expected utility given that fact. You don't waste time hand-wringing over Death if you don't expect it to do you any good; you just move on to other questions and look for ways to achieve your achievable desires. In the standard Death in Damascus problem, staying in Damascus can stand in for the entire "achievable desires" category, provided that fleeing to Aleppo has nonzero cost. CDT can be made to give up arbitrary amounts of actually achievable utility in the pursuit of a goal (escaping Death) that the CDT agent itself knows is impossible.

The other dilemmas in the paper might help draw this point out more. The random coin variant, for example, shows that in cases where the agent finds a loophole that does allow them to cost-effectively escape Death at least some of the time, standard formulations of CDT in the literature will not reliably take advantage of this loophole, whereas FDT will.

0 0 Reply • Share >

 **Zachary Jacobi** 7 years ago — 🚩
 Typo in the paper on page 9: (in the counterfactual world that the imagines)

0 0 Reply • Share >

 **Sniffnoy** 7 years ago — 🚩
 The links in footnote 3 are broken due to smart quotes. Would you mind fixing this? Thank you!

0 0 Reply • Share >

 **Rob Bensinger** Mod → Sniffnoy 7 years ago — 🚩
 Fixed, thanks!

0 0 Reply • Share >

[Subscribe](#)

[Privacy](#)

[Do Not Sell My Data](#)