Counterpossibles, Functional Decision Theory, and Artificial Agents

Alexander W. Kocurek
University of California, San Diego
San Diego, CA, United States
akocurek@ucsd.edu

1 Introduction

Recently, Yudkowsky and Soares (2018) and Levinstein and Soares (2020) have developed a novel decision theory, *Functional Decision Theory* (FDT). They claim FDT outperforms both *Evidential Decision Theory* (EDT) and *Causal Decision Theory* (CDT). Yet FDT faces several challenges. First, it yields some very counterintuitive results (Schwarz 2018; MacAskill 2019). Second, it requires a theory of counterpossibles, for which even Yudkowsky and Soares (2018) and Levinstein and Soares (2020) admit we lack a "full" or "satisfactory" account.

Here, I focus on the latter problem of counterpossibles. My aim is to establish two claims. First, the problem of counterpossibles does not even arise without a fairly strong assumption—one that rarely applies to human agents, but may apply to artificial agents. And second, even given this assumption, the problem is solv*able*, though how best to solve it remains an open question.

2 Background

To begin, recall a familiar problem from decision theory (Nozick 1969):

> *Newcomb.* Before you are two boxes: one transparent and one opaque. You see \$1K in the transparent box, but you can't see inside the opaque box. You're can take either both boxes or just the opaque box. A reliable predictor put \$1M in the opaque box if they predicted you'd take just the opaque box; otherwise, they put \$0 in it.

What should you do in *Newcomb*? There are two standard answers. According to *Evidential Decision Theory* (EDT), you should do whatever provides the best evidence that good outcomes will obtain if you do it (Jeffrey 1965; Price 1991; Ahmed 2014b). For EDTers, this means the expected utility of an act a is calculated using the agent's credence function *conditionalized* on their choosing a. More precisely, let's say a **decision situation** is a tuple $\Delta = \langle \Omega, P, O, V \rangle$, where $\langle \Omega, P \rangle$ is a probability measure representing the agent's credal state, O is a set of options available to the agent, and V is their utility function for various act-state pairs. According to EDT, the expected utility of an option a in Δ is given as follows:

$$\mathrm{EDT}(\Delta, a) = \sum\nolimits_{s \in \Omega} P(s \mid a) V(a, s)$$

EDT recommends one-boxing: if you take one box, it's very likely the predictor put \$1M in it and you're a millionaire, whereas if you take both boxes, it's very likely the predictor put \$0 in it and you can't even pay your rent. So you should take just the opaque box.

According to *Causal Decision Theory* (CDT), you should do whatever would most likely bring about good outcomes were you to do it (Gibbard and Harper 1978; Lewis 1981; Joyce 1999).[1] For CDTers, this means the expected utility of an act a is calculated using the agent's

[1] Some use "counterfactual" in place of "causal"; see Hedden 2023; Gallow 2024 (cf. Collins 1996).

credence function *imaged* on their choosing a. Imaging can be interpreted in a number of ways (Lewis 1981; Gärdenfors 1982; Joyce 1999; Hitchcock 2016; Joyce and Gibbard 2016). The exact interpretation we adopt will not matter below, so for ease of exposition and simplicity, we'll interpret imaging in terms of counterfactuals (Gibbard and Harper 1978):[2]

$$\mathrm{CDT}(\Delta, a) = \sum_{s\in\Omega} P(s \,||\, a)V(a, s), \text{ where (for simplicity) } P(s \,||\, a) \coloneqq P(a \mathrel{\square\!\!\rightarrow} s)$$

CDT recommends two-boxing: nothing you do now can change what's in the opaque box, and so no matter what, if you were to take both boxes, you'd have \$1K more than if you were to take just the opaque box. So you should take both boxes.

Various counterexamples have been proposed to each decision theory. Against EDT, consider the following case (Egan 2007):

> *Smoking Lesion.* Smoking doesn't cause cancer. Instead, there's a brain lesion some people are born with that both causes cancer and causes people to smoke. Smokers are not more likely to develop cancer than non-smokers with the same lesion-status.

What should you do in *Smoking Lesion*? Intuitively, you should smoke: after all, smoking will have no impact on whether you have the lesion, so you might as well get the pleasure of smoking. This is what CDT recommends. By contrast, EDT says you should not smoke: smoking is evidence you have the lesion, and thus evidence that you will develop cancer.

Against CDT, consider the following case (Gibbard and Harper 1978):

> *Death in Damascus.* You are in Damascus when Death knocks on your door. Death says, "I'm coming for you tomorrow". You know Death works from a book, written in advance, that lists the time and place of each person's death. You have two options: stay in Damascus or flee to Aleppo. If you stay in Damascus, you will have a peaceful evening. If you flee to Aleppo, you'll be up all night and exhausted.

What should you do in *Death in Damascus*? Intuitively, you should stay put: no matter where you go, Death will likely follow, so you might as well enjoy your final night in peace. This is what EDT recommends. By contrast, CDT does not give a stable recommendation: the more likely you are to stay in Damascus, the higher the expected utility of fleeing and vice versa.

Whether EDT or CDT can overcome these problems, or whether these predictions are even problematic to begin with, is debated (Eells 1984; Egan 2007; Arntzenius 2008; Joyce 2012; Ahmed 2014a). But Yudkowsky and Soares (2018) and Levinstein and Soares (2020) take cases like the above to show there is something wrong with both EDT and CDT and that we should seek an alternative decision theory that will yield the "right" results. While I myself disagree (as a CDTer), for the sake of discussion, I will grant these intuitive judgments and assume they present at least a prima facie problems for EDT and CDT.

3 Functional Decision Theory

Functional Decision Theory (FDT) starts with the idea that agents employ certain *decision algorithms* when making decisions. Little is said about what this means, but roughly, we can think of these algorithms as decision rules that encode deliberative dispositions: agents are disposed to make decisions in line with what their decision algorithm recommends. Formally, we can represent decision algorithms as mathematical functions taking a decision situation and an

[2] Yudkowsky and Soares (2018) and Levinstein and Soares (2020) both instead use causal models to define imaging. This would amount to adding a causal graph to the definition of a decision situation. As these details won't matter below, I'll stick with the simple formulation in terms of counterfactuals.

option as inputs and outputting an expected value for that option in that situation.[3] Both EDT and CDT can be seen as examples of decision algorithms in this sense.

According to FDT, agents should consider not just what would happen if *they were to choose differently*, but also what would happen if their *decision algorithm were to output different choices*. Informally, you should do whatever would most likely bring about good outcomes *were your decision algorithm to output it*. Formally, FDT is like CDT in that it calculates expected utility using imaging. But FDT images on a different proposition. Whereas CDT images the agent's credence function on the proposition that the agent *chooses* a certain action, FDT images on the proposition that *the agent's decision algorithm δ maximizes its output* on that action.

$$\text{FDT}(\Delta, a) = \sum\nolimits_{s\in\Omega} P(s \,||\, \text{argmax}_x \delta(\Delta, x) = a) V(a, s)$$

In a slogan: *be the kind of agent it would be best to be*. Or, less catchily: employ the decision algorithm that would most likely yield good outcomes.

To see how FDT differs from rivals, consider what it recommends in each of the three decision situations above. In *Newcomb*, FDT agrees with EDT that you should one-box. For if your decision algorithm were to output one-boxing, the predictor would very likely predict this, in which case, there would be \$1M in the opaque box. Yet if your decision algorithm were to output two-boxing, the predictor would very likely predict this, in which case, there would be \$0 in the opaque box. So it would be better if your decision algorithm outputted one-boxing.

Likewise, in *Death in Damascus*, FDT recommends staying in Damascus. For if your decision algorithm were to output staying, Death's book would have you down as staying. And if your decision algorithm were to output fleeing, Death's book would have you down as fleeing. So it would be better if your decision algorithm saved you the trouble and outputted staying.

The reason FDT differs from CDT is that the output of one's decision algorithm is causally (or explanatorily) upstream from the decision itself. In effect, the output of the agent's algorithm is a "common cause" of both the agent's decision and the predictor's prediction, who, we might suppose, has some reliable indicator of how their decision algorithm behaves.

By contrast, in *Smoking Lesion*, FDT agrees with CDT that you should smoke. One's decision algorithm does not cause/explain the presence of the relevant brain lesion from birth. Thus, regardless of what your decision algorithm were to recommend, the chance that you would have the lesion remains the same, and so, you might as well smoke.

Advocates of FDT defend these verdicts: you should one-box in *Newcomb*, smoke in *Smoking Lesion*, and stay in Damascus in *Death in Damascus*. Again, while there is substantial debate over this point, for the sake of discussion, I'm granting advocates these verdicts are indeed the correct ones. Still, FDT faces two main problems.

The first major problem is there are cases where FDT yields highly counterintuitive verdicts. Consider *Transparent Newcomb*, which is exactly like Newcomb except both boxes are transparent: you can see clearly that one box contains \$1M while the other contains \$1K. Intuitively, in this case, you should take both boxes—otherwise, you're leaving money on the table! Indeed, in this case, both EDT and CDT recommend two-boxing, as your action is both evidentially and causally independent of the contents of the box.

By contrast, FDT *still* recommends one-boxing in *Transparent Newcomb*. Even though you can clearly see \$1M inside, FDT says that doesn't matter: what matters is what *would* likely be inside *were* your decision algorithm to output such-and-such. If your decision algorithm were to output two-boxing, the predictor would very likely have predicted this and put \$0 in that box. If your decision algorithm were to output one-box, the predictor would very likely have predicted this and put \$1M in that box. So it would be better if your decision algorithm outputted one-boxing. Thus, according to FDT, you should one-box.

[3] For certain purposes, we may want a more fine-grained notion of algorithm on which multiple algorithms may produce the same input-output pairs. I'll set this complication aside.

Whether examples like this ultimately undermine FDT is a matter of debate. Yudkowsky and Soares (2018) argue that one-boxing in *Transparent Newcomb* is the correct result. By contrast, Schwarz (2018) and MacAskill (2019) both present variants of *Transparent Newcomb* that make the counterintuitiveness more pressing. For lack of space, I set aside concerns about whether FDT makes the intuitively correct verdicts in these cases for another time.

The second major problem for FDT, which is the one I take up here, is its reliance on counterpossibles. FDT's calculation of expected utility requires imaging on a mathematically impossible proposition, viz., that a certain mathematical function δ, given certain inputs, yields different outputs than it actually yields. Yet, by Yudkowsky and Soares's (2018) and Levinstein and Soares's (2020) own admissions, we do not have a "full" or "satisfactory" theory of counterpossibles. Yudkowsky and Soares (2018) describe it as "the main drawback of FDT relative to CDT". Similarly, Levinstein and Soares (2020) admit that "for FDT to be successful, a more worked out theory [of counterpossibles] is necessary". Critics likewise see this as a major issue for the view (Schwarz 2018; MacAskill 2019).

The problem rests on two premises: (1) FDT requires agents to evaluate the probability of counterpossibles; and (2) we have no "full" or "satisfactory" account of counterpossibles. I will challenge each premise and, in so doing, argue the problem is at least solvable, though it remains to be seen which of the solutions on the table are most appealing.

4 Are counterpossibles necessary for FDT?

Why think FDT requires counterpossibles? The idea is this. Let δ be the function representing the agent's decision algorithm. Whether δ maximizes its value on an input a i.e., whether $\text{argmax}_x\delta(\Delta, x) = a$, is a mathematical fact. Mathematical facts are not contingent. So the claim '$\text{argmax}_x\delta(\Delta, x) = a$', if false, is mathematically impossible.

Note, however, when we say "Let δ be the function…", we're treating 'δ' as a rigid designator, denoting a fixed mathematical function that represents the algorithm that the agent *actually* employs. This is how Yudkowsky and Soares (2018) and Levinstein and Soares (2020) formulate FDT. But if we instead treat 'δ' as a non-rigid designator, so that, for each world w where the agent exists, 'δ' denotes the function representing the agent's algorithm at w, '$\text{argmax}_x\delta(x) = a$' becomes a contingent statement since agents may employ different algorithms at different worlds. In other words, if 'the agent's decision algorithm' is interpreted *de dicto* (rather than *de re*) in the relevant counterfactual, it isn't a counterpossible.

Arguably, the de dicto interpretation is enough to make the predictions its advocates want it to make (cf. Schwarz 2018): it still predicts you should one-box in *Newcomb*, smoke in *Smoking Lesion*, and stay home in *Death in Damascus*. Moreover, the de re interpretation does not seem relevant for most agents in most contexts. Most agents do not have introspective access to their own decision-making process (if only!): they are *uncertain* regarding which decision algorithm they do employ or would employ. Reasoning about what would happen if (per impossibile) a certain function f had yielded different outputs is only relevant if the agent is certain that f represents their decision-making process. If an agent doesn't know which function represents their decision algorithm, this uncertainty will need to be reflected in how they calculate expected utility, which is what the de dicto (but not the de re) interpretation does.

More precisely, let $\hat{\delta}$ be a non-rigid term denoting, at each world w, the decision algorithm the agent uses at w. Let δ be a rigid term denoting the decision algorithm the agent actually uses. The de re interpretation of FDT for a decision situation Δ seems to assume the following:[4]

Counterfactual Robustness:

$$P(\hat{\delta} = \delta \,||\, \text{argmax}_x\hat{\delta}(\Delta, x) = a) = P(\hat{\delta} = \delta \,||\, \text{argmax}_x\delta(\Delta, x) = a) = 1$$

[4]This violates the Strangeness of Impossibility Condition from Nolan 1997; see Kocurek 2021a for discussion.

The de dicto and de re interpretations collapse given Counterfactual Robustness together with modest assumptions.[5] But in most contexts, human agents do not obey this principle.

This severely limits the force of the problem of counterpossibles, but it does not eliminate it. For one, there may be limited circumstances where Counterfactual Robustness is correct even for rational human agents. Perhaps a human agent learns what their decision algorithm is after years of therapy or after a series of invasive brain scans. Or perhaps a decision theorist becomes so convinced of the arguments for a certain decision theory that they become certain they would employ that theory regardless of its outputs. Moreover, the initial motivation for developing FDT was to develop a decision theory for *artificial* agents, who may have access to their own source code (Soares and Fallenstein 2015). For such agents, Counterfactual Robustness may be more a reasonable assumption. Indeed, advocates of FDT often talk not about what an agent's decision algorithm recommends, but what *FDT itself* recommends, suggesting they have in mind agents who know their decision algorithm is FDT. So, Counterfactual Robustness may be appropriate in special applications. In that case, advocates of FDT have to tackle counterpossibles head on.

5 Are counterpossibles problematic?

Even if we formulate FDT so as to require counterpossibles, this isn't necessarily problematic. There are many accounts of counterpossibles in the literature (see Kocurek 2021a for overview), including: similarity accounts that extend the classic ordering semantics with impossible worlds (Nolan 1997; Krakauer 2012; Brogaard and Salerno 2013; Kment 2014; Berto et al. 2018); grounding accounts that extend the causal modeling framework with noncontingent nodes (Schaffer 2016; Wilson 2018; Baron, Colyvan, and Ripley 2020; Khoo 2022); and counterconventional accounts that simulate counterpossible reasoning without direct appeal to "worldly" impossibilities (Kocurek and Jerzak 2021; Kocurek 2019, 2021b). Still, there are two related problems that might lead one to think none of these accounts is satisfactory for current purposes.

First, for the expected utility calculation to work in FDT, the probability function imaged on $\operatorname{argmax}_x \delta(\Delta, x) = a$ must still be a *probability* function. But Schwarz (2018) raises the worry that imaging on the impossible could lead to violations of the Kolmogorov axioms. For example, these axioms entail that $P(A \wedge B) \leqslant P(A)$. So if $P(A \wedge \neg A \,||\, A \wedge \neg A) = 1$, then $P(A \,||\, A \wedge \neg A) = P(\neg A \,||\, A \wedge \neg A) = 1$. But this violates another implication of the Kolmogorov axioms: $P(\neg A) = 1 - P(A)$. More generally, the Kolmogorov axioms assume the logical connectives $\neg$, $\wedge$, and so on are interpreted classically: $\neg A$ is true at w iff A is not true at w, $A \wedge B$ is true at w iff A and B are both true at w, and so on. Yet at logically impossible worlds, these assumptions can fail: there can be worlds where both $\neg A$ and A are true (or both false), worlds where $A \wedge B$ is true and yet A and/or B are false, and so on.

Second, as Yudkowsky and Soares (2018) note, while there are methods for discovering causal structure (Pearl 2000), we have no such thing for counterpossible dependencies. How is an agent to assign rational credences to counterpossibles? Arguably, there is no universal similarity ordering, grounding graph, etc. that works in every case and we should instead focus on developing ones tailored to specific applications. We may even have artificial agents *learn*, via standard machine learning techniques, which ordering, graphs, etc. are most effective at maximizing utility in a range of typical scenarios. Still, it would be more satisfactory to have a story regarding how agents should assess the probability of a counterpossible.

[5] Here are the modest assumptions: First, $P(\cdot \,||\, A)$ is a probability function (see section 5). Second, $P(A \,||\, A) = 1$ (success). Third, $P(A \,||\, B) = P(B \,||\, A) = 1$ implies $P(C \,||\, A) = P(C \,||\, B)$ (replacement of counterfactually equivalent antecedents). Finally, perhaps redundantly, $P(a = b \,||\, A) = 1$ implies $P(\phi(a) \,||\, A) = P(\phi(b) \,||\, A)$ (substitution of counterfactual identicals; this may or may not follow from the first assumption, depending on how its formulated). I don't claim these assumptions hold universally (e.g., if the antecedents express the failures of these very rules), but they seem plausible in this context.

To solve the first problem, we need to clarify how we're interpreting logical operators used in probability statements. Suppose we want to characterize the credences of an intuitionistic logician. When we write '$P(\neg A)$', what does '$\neg$' refer to? According to an *object-level* interpretation, '$\neg$' refers to an operation the *agent* uses in their own thought, e.g., intuitionistic negation. According to a *meta-level* interpretation, '$\neg$' refers to an operation the *theorist* uses in the metalanguage, viz., classical negation. The problem arises from conflating these two interpretations. When we, as *theorists*, state the Kolmogorov axioms, we adopt a meta-level interpretation, where the connectives are interpreted as the theorist interprets them in their (meta)language. By contrast, when discussing logically deviant or logically uncertain agents, we often want to represent their credal state from *their* perspective, in *their* language, and thus adopt an object-level interpretation, where the connectives stand for the representations agents themselves use.

One way to avoid this confusion is to introduce a different set of symbols for operators with different interpretations. For example, we may use $-$, $\cap$, $\cup$, and so on, for *classically rigid* connectives, which are always interpreted according to the (classical) metalanguage even under the scope of a counterpossible supposition. We may then reserve $\neg$, $\wedge$, $\vee$, and so on for the operations that the agent themselves use, which they may interpret nonclassically. This allows us to simultaneously state the Kolmogorov axioms in the metalanguage while representing the language the agent uses "from within". For example, for intuitionistic agents, $P(A \cup -A) = 1$ even if $P(A \vee \neg A) \neq 1$. And $P(-A) = 1 - P(A)$ even if $P(\neg A) \neq 1 - P(A)$.[6] Thus, even if an agent images on a classically impossible proposition, the result can still be a probability function. For example, $P^*(\cdot) \coloneqq P(\cdot \,||\, A \wedge \neg A)$ can still satisfy the Kolmogorov axioms stated in the classical metalanguage (e.g., while $P^*(A \wedge \neg A) = 1$, still $P^*(A \cap -A) = 0$).

This resolution to the first problem helps resolve the second. When representing agents entertaining counterpossible suppositions, it seems most appropriate to use the object-level operators that they use to represent those impossibilities. In that case, we can understand their counterpossible suppositional reasoning as a kind of shift in the interpretation of the terms they use to state those suppositions. This is effectively the expressivist approach to counterlogicals adopted by Kocurek and Jerzak (2021) and Kocurek (2021b). Indeed, Kocurek and Jerzak (2021) show that this approach can represent the same hyperintensional phenomena as the standard impossible worlds approach: both approaches generate the same hyperintensional logic over the propositional counterfactual language (though they diverge in more powerful languages; see Kocurek 2021b). By allowing re-interpretation of the object-language terms, it is thus possible to simulate all counterpossible reasoning without appealing to impossible worlds.

At the same time, interpretation-shifting illuminates why certain counterpossibles are true and also how we may be able to know which counterpossible dependencies hold: our ability to know counterpossibles stems from our general ability to consider and reason about alternative interpretations. This approach therefore offers a general method for establishing counterpossible dependencies via the manipulation of interpretations. How is an agent to determine whether a counterpossible holds? By exploring alternative interpretation-world pairings that make the antecedent true and evaluating whether the consequent is true at those pairs.

Of course, this is not a complete theory of counterpossibles. The space of possible interpretations is vast. Yet agents only select a small set of reasonable alternative interpretations when engaged in counterpossible reasoning. What counts as a "reasonable" alternative interpretation pair will be a context-sensitive matter: not anything goes. So, in a sense, the problem of counterpossible lingers. But this approach at least offers a method via which we can start to address some of the thorny questions surrounding context-resolution that would help settle how to uncover counterpossible dependencies in a more objective manner.[7]

[6] It is possible to mix meta-level and object-level connectives in a rigorous and formally consistent manner, as in *hyperlogic* (Kocurek 2019, 2021, 2024a, 2024b).

[7] Thanks to Alan Hájek, Ethan Jerzak, Rachel Rudolph, and the audience members at the Dartmouth Summer

References

Ahmed, Arif (2014a). "Dicing with death". In: *Analysis* 74.4, pp. 587–592.
— (2014b). *Evidence, Decision and Causality*. Cambridge: Cambridge University Press.
Arntzenius, Frank (2008). "No Regrets, or: Edith Piaf Revamps Decision Theory". In: *Erkenntnis* 68, pp. 277–297.
Baron, Sam, Mark Colyvan, and David Ripley (2020). "A Counterfactual Approach to Explanation in Mathematics". In: *Philosophia Mathematica* 28.1, pp. 1–34.
Berto, Francesco et al. (2018). "Williamson on Counterpossibles". In: *Journal of Philosophical Logic* 47, pp. 693–713.
Brogaard, Berit and Joe Salerno (2013). "Remarks on Counterpossibles". In: *Synthese* 190.4, pp. 639–660.
Collins, J. (1996). "Supposition and Choice: Why 'Causal Decision Theory' is a Misnomer". Manuscript.
Eells, Ellery (1984). "Newcomb's Many Solutions". In: *Theory and Decision* 16.1, pp. 59–105.
Egan, Andy (2007). "Some counterexamples to causal decision theory". In: *The Philosophical Review* 116.1, pp. 93–114.
Gallow, J. Dmitri (2024). "Counterfactual Decision Theory is Causal Decision Theory". In: *Pacific Philosophical Quarterly* 105, pp. 115–156.
Gärdenfors, Peter (1982). "Imaging and Conditionalization". In: *Journal of Philosophy* 79.12, pp. 747–760.
Gibbard, Allan and William L Harper (1978). "Counterfactuals and Two Kinds of Expected Utility". In: *Foundations and Applications of Decision Theory*. Ed. by Clifford Alan Hooker, James J Leach, and Edward Francis McClennen. Dordrecht: D. Reidel, pp. 125–162.
Hedden, Brian (2023). "Counterfactual Decision Theory". In: *Mind* 132.527, pp. 730–761.
Hitchcock, Christopher (2016). "Conditioning, Intervening, and Decision". In: *Synthese* 193, pp. 1157–1176.
Jeffrey, Richard C. (1965). *The Logic of Decision*. Chicago: University of Chicago Press.
Joyce, James (1999). *The Foundations of Causal Decision Theory*. Cambridge: Cambridge University Press.
— (2012). "Regret and Instability in Causal Decision Theory". In: *Synthese* 187, pp. 123–145.
Joyce, James and Allan Gibbard (2016). "Causal Decision Theory". In: *Readings in Formal Epistemology*. Ed. by Horacio Arlø-Costa, Vincent F. Hendricks, and Johan van Benthem. Berlin: Springer, pp. 457–491.
Khoo, Justin (2022). *The Meaning of If*. Oxford: Oxford University Press.
Kment, Boris (2014). *Modality and Explanatory Reasoning*. Oxford: Oxford University Press.
Kocurek, Alexander W. (2019). "Hyperlogic: A System for Talking about Logics". In: *Proceedings for the 22nd Amsterdam Colloquium*. Ed. by Julian J. Schlöder, Dean McHugh, and Floris Roelofsen. ILLC. Amsterdam, pp. 238–247.
— (2021a). "Counterpossibles". In: *Philosophy Compass* 16.11, e12787.
— (2021b). "Logic Talk". In: *Synthese* 199.5–6, pp. 13661–13688.
— (2024a). "The Logic of Hyperlogic. Part A: Foundations". In: *Review of Symbolic Logic* 17.1, pp. 244–271.
— (2024b). "The Logic of Hyperlogic. Part B: Extensions". In: *Review of Symbolic Logic* 17.3, pp. 654–681.
Kocurek, Alexander W. and Ethan J. Jerzak (2021). "Counterlogicals as Counterconventionals". In: *Journal of Philosophical Logic* 50.4, pp. 673–704.
Krakauer, Barak (2012). "Counterpossibles". PhD thesis. University of Massachusetts, Amherst.

Workshop on Artificial Intelligence, Language, and Cognition for helpful feedback on this project.

Levinstein, Benjamin A. and Nate Soares (2020). "Cheating Death in Damascus". In: *The Journal of Philosophy* 117.5, pp. 237–266.

Lewis, David K (1981). "Causal Decision Theory". In: *Australasian Journal of Philosophy* 59.1, pp. 5–30.

MacAskill, William D. (2019). "A Critique of Functional Decision Theory". https://www.lesswrong.com/posts/ySLYSsNeFL5CoAQzN/a-critique-of-functional-decision-theory.

Nolan, Daniel (1997). "Impossible Worlds: A Modest Approach". In: *Notre Dame Journal of Formal Logic* 38.4, pp. 535–572.

Nozick, Robert (1969). "Newcomb's Problem and Two Principles of Choice". In: *Essays in Honor of Carl G. Hempel*. Ed. by Nicholas Rescher. Dordrecht: Springer, pp. 114–146.

Pearl, Judea (2000). *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.

Price, Huw (1991). "Agency and Probabilistic Causality". In: *British Journal for Philosophy of Science* 42, pp. 157–176.

Schaffer, Jonathan (2016). "Grounding in the Image of Causation". In: *Philosophical Studies* 173, pp. 49–100.

Schwarz, Wolfgang (Dec. 2018). "On Functional Decision Theory". https://www.umsu.de/wo/2018/688.

Soares, Nate and Benja Fallenstein (2015). "Toward Idealized Decision Theory". In: *arXiv* 1507.01986[cs.AI], pp. 1–15. URL: https://arxiv.org/abs/1507.01986.

Wilson, Alastair (2018). "Grounding Entails Counterpossible Non-Triviality". In: *Philosophy and Phenomenological Research* 46.3, pp. 716–728.

Yudkowsky, Eliezer and Nate Soares (2018). "Functional Decision Theory: A New Theory of Instrumental Rationality". In: *arXiv* 1710.05060[cs.AI], pp. 1–36. URL: https://arxiv.org/abs/1710.05060.