Success-First Decision Theories

Preston Greene

In Arif Ahmed, editor, *Newcomb's Problem*, pages 115–137. Cambridge University Press, 2018.

0 Abstract

The standard formulation of Newcomb's problem compares evidential and causal conceptions of expected utility, with those maximizing evidential expected utility tending to end up far richer. Thus, in a world in which agents face Newcomb problems, the evidential decision theorist might ask the causal decision theorist: "if you're so smart, why ain'cha rich?" Ultimately, however, the expected riches of evidential decision theorists in Newcomb problems do not vindicate their theory, because their success does not generalize. Consider a theory that allows the agents who employ it to end up rich in worlds containing Newcomb problems and continues to outperform in other cases. This type of theory, which I call a "success-first" decision theory, is motivated by the desire to draw a tighter connection between rationality and success, rather than to support any particular account of expected utility. The primary aim of this paper is to provide a comprehensive justification of success-first decision theories as accounts of rational decision. I locate this justification in an *experimental* approach to decision theory supported by the aims of methodological naturalism.

1 Introduction

The classic formulation of Newcomb's problem is often thought to compare evidential and causal conceptions of expected utility, with those maximizing evidential expected utility tending to end up far richer than those maximizing causal expected utility. Thus, in a world in which agents face classic Newcomb problems, the evidential decision theorist might ask the causal decision theorist: "if you're so smart, why ain'cha rich?" Ultimately, however, the expected riches of evidential decision theorists do not vindicate their theory, because their success does not generalize. For example, in a world in which agents face "transparent" variants of Newcomb's problem, where the contents of both boxes are revealed before a choice must be made, both causal and evidential decision theorists tend to end up poor.[1] Thus, in such a world, those following some other theory could ask of both causal and evidential decision theorists "why ain'cha rich?"

Consider a theory that allows the agents who employ it to end up rich in worlds containing both classic and transparent Newcomb problems. This type of theory is motivated by the desire to draw a tighter connection between rationality and success, rather than to support any particular account of expected utility. We might refer to this type of theory as a "success-first" decision theory. The main aim of this paper is to provide a comprehensive metatheoretical justification of success-first decision theories as accounts of rational decision .

The desire to create a closer connection between rationality and success than that offered by standard decision theory has inspired several success-first decision theories over the past three decades, including those

[1] The transparent Newcomb problem is discussed in Gibbard and Harper, 1981, 181–2. For further examples of a why-ain'cha-rich objection lodged against EDT, see Arntzenius, 2008, 289–90 and Soares and Fallenstein, 2015, 2–3. See Lewis, 1981b and Joyce, 1999, 151–4 for other types of responses to the why-ain'cha-rich objection on behalf of CDT.

of Gauthier [1986], McClennen [1990], and Meacham [2010], as well as an influential account of the rationality of intention formation and retention in the work of Bratman [1999]. McClennen [1990, 118] writes: "This is a brief for rationality as a positive capacity, not a liability—as it must be on the standard account." Meacham [2010, 56] offers the plausible principle "If we expect the agents who employ one decision making theory to generally be richer than the agents who employ some other decision making theory, this seems to be a prima facie reason to favor the first theory over the second." And Gauthier [1986, 182–3] proposes that "a [decision-making] disposition is rational if and only if an actor holding it can expect his choices to yield no less utility than the choices he would make were he to hold any alternative disposition." In slogan form, Gauthier [1986, 187] calls the idea "utility-maximization at the level of dispositions," Meacham [2010, 68–9] a "cohesive" decision theory, McClennen [1990, 6–13] a form of "pragmatism," and Bratman [1999, 66] a "broadly consequentialist justification" of rational norms.

Even though these metatheoretical views have some initial plausibility, they have not been rigorously developed past the slogans (the object-level theories, which I evaluate in Section 3.3, are much more developed), and it is partly for this reason that many theorists working on practical rationality view success-first decision theories with deep suspicion. In Section 2, I provide a comprehensive metatheory for success-first decision theories, which I locate inside an *experimental approach* to decision theory. The experimental approach helps reveal what the why-ain'cha-rich objection is really getting at, and it provides a guiding light for success-first decision theory. I show that the approach is compatible with the general ambitions of both CDT and EDT, although here my primary focus is on the causalist. The experimental approach diverges from the standard methodology of CDT by suggesting that experiments, rather than mathematical axiom systems, are the proper tool for studying causal efficacy.

In Section 3, I outline the sort of object-level theories that follow from the experimental approach. For the causal experimentalist, the difference between one- and two-boxing in Newcomb's problem is not that between maximizing causal and evidential expected utility, but rather that between *acts* and *decision theories* serving as the *independent variable* of the experiment.

2 The Experimental Approach to Decision Theory

Decision theory is concerned with instrumental rationality: the rationality of instrumental aims. As such, it is not concerned with the rationality of ultimate aims. Many decision theorists point out that decision theory is relevant to all forms of decision making: not just in prudential contexts, but in moral contexts as well, *precisely because* decision theory is not concerned with ultimate aims.

Accordingly, the decision theorist's job is like that of an engineer in inventing decision theories, and like that of a scientist in testing their efficacy. A decision theorist attempts to discover decision theories (or decision "rules," "algorithms," or "processes") and determine their efficacy, under certain idealizing conditions, in bringing about what is of ultimate value.

Someone who holds this view might be called a *methodological hypernaturalist*, who recommends an *experimental approach to decision theory*.[2] On this view, the decision theorist is a scientist of a special sort, but their goal should be broadly continuous with that of scientific research. The goal of determining efficacy in bringing about value, for example, is like that of a pharmaceutical scientist attempting to discover the efficacy of medications in treating disease.

[2] "Hyper" because many decision theorists think of themselves as methodological naturalists but do not subscribe to the experimental approach, often due to concerns that I allay in Section 2.2.

For game theory, Thomas Schelling [1960] was a proponent of this view. The experimental approach is similar to what Schelling meant when he called for "a reorientation of game theory" in Part 2 of *A Strategy of Conflict*. Schelling argues that a tendency to focus on first principles, rather than upshots, makes game-theoretic theorizing shockingly blind to rational strategies in coordination problems. A simple example asks respondents to name "heads" or "tails" with the understanding that if a partner in another room does the same they both receive a prize [Schelling, 1960, 56]. Since the heads-heads and tails-tails outcomes are both in equilibrium and Pareto-optimal, the game-theoretic answer is that agents should pick a response at random. This strategy secures the prize half of the time. But respondents tend to do much better than that. A large majority of respondents choose heads [Mehta et al., 1994]. Schelling explains that "heads" represents a point of convergence of expectations for most people, which he calls a "focal point."[3] He introduces many examples to show that attention to focal points is a valuable tool for coordinating, bargaining, and deterring. Focal points, however, have received little attention from game theorists.[4]

Focal points would not disappear under idealization, and so cannot be dismissed as concerning non-ideal theory. Rather, the lack of attention to focal points is due to the difficulty in deriving a systematic theory of their use from axioms of rational choice. In contrast, Schelling imposes no such restriction on how a theory is derived. Instead, he commits to finding the best theory by *induction from its success*. Referring to coordination games like "heads/tails," Schelling [1960, 283–4, emphasis added] writes that his basic premise is that rational players realize "that some rule must be used if success is to exceed coincidence, and that the best rule to be found, *whatever its rationalization*, is consequently a rational rule."

The experimental approach to decision theory departs from standard methodology in the way imagined by Schelling. Rather than attempting to justify theories through deduction from a priori axioms, the main justification lies in the results. Only after expected results are determined should foundations be inferred. To borrow from William James [1896, 726]: the strength of the standard decision theorist's system "lies in the principles, the origin, the *terminus a quo* of his thought; for us the strength is in the outcome, the upshot, the *terminus ad quem*. Not where it comes from but what it leads to is to decide."[5]

2.1 Experiments

The best metatheory for success-first decision theories develops Schelling's insights into a rigorous experimental approach. First, note how experiments in decision theory differ from those of other scientists. The decision theorist uses thought experiments, not concrete experiments, and they determine the efficacy of decisions or decision theories from the specification of problems, rather than through controlled trials.

These methodological differences owe to decision theory's goal of instrumental value-maximization across possible worlds. In comparison, the goal of treating a disease as it actually exists, for example, is severely circumscribed. Nevertheless, in each instance the function of thought experiments and concrete experiments is the same. In empirical science, experiments and thought experiments share a functional description [El Skaf and Imbert, 2013].

An experiment aims to investigate how a *dependent variable* depends on an *independent variable*. For example, a research pharmacist attempts to discover medications (independent variable) and determine their efficacy in

[3] In another famous example, Schelling asked respondents to pick a time and place to meet a partner in New York City without prior communication. Most respondents selected Grand Central Station at noon [Schelling, 1960, 55–6].

[4] Exceptions include Sugden, 1995 and Sugden and Zamarrón, 2006.

[5] Quoted in Sugden and Zamarrón 2006, 620.

treating disease (dependent variable). This requires careful use of controls that help isolate effects of the independent variable from others. It is not the role of the pharmacist to define "disease" and "cure," or even "therapeutic effect" and "adverse effect." Rather, the pharmacist is in the first place simply interested in determining effects of a treatment.

The same should hold of decision theory. As decision theory concerns instrumental rationality, it is not the place of the decision theorist to determine which states of the world are better than others. Rather, the decision theorist studies the efficacy of an action or decision theory (independent variable) in bringing about states of affairs (dependent variable). The value of these states is often taken as determined by the agent's preferences; however, in line with the idea that decision theory concerns all forms of decision making, values can be determined in many ways. The idealizations (in idealized decision theory) act to isolate effects of the independent variable from others, like an agent's belief accuracy or cognitive abilities.

Consider James Joyce's [1999, 146–7] fascinating description of Newcomb's problem:

> Suppose there is a brilliant (and very rich) psychologist who knows you so well that he can predict your choices with a high degree of accuracy. One Monday as you are on the way to the bank he stops you, holds out a thousand-dollar bill, and says: "You may take this if you like, but I must warn you that there is a catch. This past Friday I made a prediction about what your decision would be. I deposited $1,000,000 into your bank account on that day if I thought you would refuse my offer, but I deposited nothing if I thought you would accept.[6]

In Nozick's [1969, 114–5] classic formulation, the bank account is an opaque box and the money is in a transparent box. Joyce's otherwise equivalent formulation does an excellent job of highlighting the causal inefficacy of refusing the thousand dollars.

When determining causal efficacy under the experimental approach, there is a choice regarding the independent variable, and this is best highlighted by Newcomb's problem. As Joyce's formulation reveals, if we set the independent variable to *acts*, then the causally efficacious option is to accept the thousand. We deduce, using causal dominance, that the effect of accepting (i.e., "two-boxing") is to leave you a thousand richer than you would be had you chosen the other option. The good thing about causal dominance arguments is that they leave no room for doubt. If we set up this experiment and controlled for variables that are causally independent of acts, we would observe that everyone who accepts the thousand gets a thousand more than they would have had they chosen the other option . This all follows from the specification of the case.[7]

The story about causal efficacy remains the same in the transparent Newcomb problem.[8] Suppose the psychologist informs his potential beneficiaries whether the money has been deposited before they decide, and

[6] Adapted from a version presented by Sobel [1985, 198–9 n. 6]. Sobel points out that the classic formulation of Newcomb's problem can lead to confusions that make one-boxing seem more appealing.

[7] Dominance arguments only apply if the states are independent of the available acts. Causal decision theorists hold that this is satisfied when the states are *causally* independent of the acts, as in Newcomb's Problem. Evidential decision theorists, however, require *evidential* independence of the acts, and therefore reject dominance reasoning in Newcomb's problem (see Joyce, 1999, 150–1). See footnote 10 for a type of dominance argument that is compatible with success-first decision theories.

[8] Gibbard and Harper [1981, 181] write: "Consider a variant on Newcomb's story: the subject of the experiment is to take the contents of the opaque box first and learn what it is; he then may choose either to take the thousand dollars in the second box or not to take it." Note that in their discussions of the transparent Newcomb problem, Broome [2001, 101], Gauthier [1986, 157–89], and Parfit [1984, 7] all use "transparency" to denote the fact that the agent can be predicted, and not the fact that the agent can see into the boxes..

as before, that he is an accurate predictor. Again, causal dominance applies. There are two possible conditions: you learn that the money has been deposited or it has not. In both conditions, you receive a thousand more by accepting than you would by refusing. This is what we would observe were we to set up the experiment and control for irrelevant factors. Thus, when studying causal act efficacy with the experimental approach, two-boxing is most efficacious in classic and transparent Newcomb problems.

An experimental rationale can also be given for the split recommendations of evidential decision theorists who favor one-boxing in the classic Newcomb problem but two-boxing in the transparent version. Rather than determining an act's causal efficacy, experiments can determine an act's "foreseeable actual expected return" [Ahmed, 2014b, 182, fn. 34]. To determine that, the relevant experiment would observe the average return of one- and two-boxing while controlling only for variables that are *evidentially* independent of the act (most importantly, the prediction is left uncontrolled). Since the predictor is accurate, these experiments would show a greater average return for one-boxing. Such an experiment would be "nonstandard" in that it would aim to establish correlation, rather than a causal relationship. However, this is unlikely to trouble evidential decision theorists, who are already committed to the idea that rational decisions are not determined by causal relationships.[9]

Two possibilities remain: a recommendation to two-box in the transparent version but one-box in the classic version, and a recommendation to one-box in both. I know of no experimental rationale for the former, but the key, I believe, to understanding the motivation behind the latter is to recognize the impulse to study the causal efficacy of *decision theories* rather than acts.[10] Understood this way, the experiment should not test the effects of the act of one-boxing, as when studying act-efficacy, but rather of employing a decision-making theory that recommends one-boxing. Thus, we set the independent variable to decision theories, not acts, and imagine experiments that isolate the causal consequences of decision theories. That changes the story about who tends to end up with more money. As is assumed in the literature, those who employ a decision-making theory that recommends one-boxing tend to get a million, while those who employ a theory that recommends two-boxing tend to get a thousand. This follows from the specification of the case—what it means for a predictor to be "accurate" etc. Furthermore, these effects can be observed, in principle, in the same way that the effects of the different acts can be observed: by controlling for causally independent factors, applying accurate predictions to each type of agent, and noting the differences in wealth that result.[11]

[9] What can be said against the evidential experimental approach? From my perspective, the largest hurdles are i) the intuition that rational decision making is a subjunctive endeavor in which the relevant considerations concern causal properties [Joyce, 1999, 252–3], and ii) the concern that this will result in a "one-box" recommendation for medical Newcomb problems (see Section 2.3.1). Whether (ii) is a problem depends on whether there exists a sound version of the "tickle defense" argument (See Ahmed, 2014b, 91–7 for a recent example and further references).

[10] A type of dominance argument that is compatible with one-boxing in both versions is the "principle of full information" endorsed by Fleurbaey and Voorhoeve [2013, 114]: "When one lacks information, but can infer that there is a particular alternative one would invariably regard as best if one had full information, then one should choose this alternative." This principle, when combined with the requirement of causal act-state independence, would merely require that a theory give the same recommendation in both classic and transparent Newcomb problems.

[11] If one does not understand this motivation, and views act-efficacy as the only possible option for evaluating rationality of decision making, then one is likely to be befuddled by the many seemingly intelligent people working in fields related to artificial intelligence who view one-boxing in the transparent Newcomb problem as rational. Arntzenius [2008, 290], for example, calls these people "insane." (The same applies, mutatis mutandis, to those befuddled by the many seemingly intelligent philosophers who view two-boxing

Thus, from the causal experimentalist perspective, disagreement over Newcomb's problem is best attributed to disagreement over the appropriate independent variable, and not over the appropriate conception of expected utility. In fact, from that perspective, casting Newcomb's problem as a disagreement over expected utility is a needless distraction. Endorsement of one-boxing is compatible with the general motivations of CDT and the use of causal notions to define expected utilities.

2.2 The Epistemic Problem with the Experimental Approach

The above might strike many as a nice idea, but unworkable as we move to more complex cases involving epistemic uncertainty. The problem is this. Most decision theorists believe that decision situations are defined by an agent's epistemic perspective, not the agent's actual situation. (This is the analog, in decision theory, of the idea that practical rationality concerns a "subjective ought," on which what an agent ought to do depends on what the agent believes, not what the world is actually like). However, the experimental approach may seem compatible only with a more "objective" conception of decision situations,[12] since it is unclear how to experiment with an epistemic perspective. So, the experimental approach might seem hopeless as an account of subjective rationality.

This is the largest obstacle to the experimental approach, and one main reason knowledgeable theorists remain skeptical of success-first decision theories. Thus, the justification for the experimental approach or success-first decision theories *will never be complete* without a proper specification of the connection between an agent's epistemic perspective and the conditions by which the agent succeeds or fails.

The solution lies in an "experimental resolution" of the epistemic perspective. To create an experimental resolution, take a description of the agent's epistemic perspective, including probabilities, causal dependency hypotheses, etc.; in short, everything that is relevant to the decision situation. The *idealized experimental resolution* of the case is a specification of a situation that eliminates the effects of inaccuracies in the agent's epistemic perspective. Using this, we deduce the effects of the independent variable on average success and failure.

Some theorists will, of course, disagree about which epistemic factors are relevant to the decision situation. The experimental approach is neutral between these different conceptions. It only suggests that, for any proposed decision situation, the description be made detailed enough (if possible) to allow for experimental resolution.

In step form:

1. The case should describe in as much detail as necessary the agent's epistemic perspective.

2. Form an idealized experimental resolution of the case, controlling for the effects of inaccuracies or limitations in the agent's epistemic perspective.

3. Deduce the averages of good and bad states for different decisions or decision makers under the experimental resolution.

This process is not as revisionary as it may seem. Theorists have long been forming idealized experimental resolutions of decision problems and making claims about them. Consider the orthodox claim that one-boxers tend to be richer than two-boxers. In descriptions of Newcomb's problem, the predictor is often called "accurate," "reliable," or similar. If these terms are meant to describe the agent's epistemic perspective and not

in the classic Newcomb problem as rational.)

[12] See Mellor, 1991 for a defense of "objective decision theory."

the actual situation, then the inference from the description to one-boxers' success is obviously invalid. Here Nozick's [1969, 114, emphasis added] original presentation of the case is particularly apt. He writes: "Suppose a being in whose power to predict your choices you have enormous confidence" and "One might tell a longer story, but all this *leads you to believe* that almost certainly this being's prediction about your choice in the situation to be discussed will be correct." Notice that we cannot directly infer any actual average monetary payouts from this description.

When theorists describe the average success of one- and two-boxers, therefore, these claims *must* concern an idealized experimental resolution of Nozick's description, which controls for effects of belief inaccuracy. Here is one such. To call the predictor reliable is exactly to say that there is a high chance of it predicting one/two-boxing in situations where one/two-boxing in fact occurs.[13] Under this assumption, the idealized experimental resolution demands a high chance of a one-box prediction given that one-boxing occurs, and a high chance of a two-box prediction given that two-boxing occurs. So, in the resolution, one-boxers tend to get a million dollars while two-boxers tend to get only a thousand (two-boxers do not *always* end up poorer, of course, and that is why the focus is on average payout across repeated hypothetical experiments, as in Step 3 above). The nature of the chances need not be metaphysically loaded: it could simply be a high hypothetical relative frequency of correct predictions. Something like this is necessary to make sense of the orthodox claim that one-boxers tend to be richer than two-boxers.

Can we always form experimental resolutions of well-formed decision problems? My main reply is that, ultimately, it does not matter. But let me first introduce the reasons for optimism about forming experimental resolutions

Causal decision theorists share a common idea, and differ mostly on emphasis [Lewis, 1981a, 5]. Some, for example, emphasize chances in their formulations, and some counterfactual conditionals. But these are interrelated and, on some views, interdefinable. The idea that causal decision theorists share is that the agent's epistemic perspective yields a causal story that must be respected when determining rational decisions. This story is often called a "dependency hypothesis." Dependency hypotheses are formulated in various ways, but all involve the possibility of constructing an experimental resolution from the story.

For example, the experimental approach is particularly friendly to "chance" readings of dependency hypotheses as maximally detailed specifications of conditional chances.[14] For *objectivists*, these comprise propositions that do not vary from person to person, while for *subjectivists*, they are based on an agent's estimation of chances. Either way, we can create an idealized experimental resolution of the agent's decision situation by assuming that the conditions determining the agent's actual payout match the conditional chances under the agent's dependency hypothesis.

We can also formulate experimental resolutions using the counterfactual dependence version. At first glance, creating a usable experimental resolution from counterfactual conditionals may seem a challenge, since the relevant chances, which are required to calculate the averages of the states obtaining, are not supplied directly by the agent's epistemic perspective but must be inferred from credences over counterfactuals. These inferences can seem especially difficult if the agent's credences are divided over conflicting counterfactuals. Nevertheless, epistemologists are doing promising research on the connection between credence and objective chance. Hájek [manuscript] suggests that credence "aligns with the truth" to the extent that it matches objective chances, and Mellor [1991, 274] proposes something similar. Other theories focus on when a credence is justified. Following Hájek and Mellor, perhaps a justified credence is one that matches or nearly matches objective chance, or is

[13] Cf. Joyce, 170, fn. 36.

[14] See Joyce, 1999, 166–7 for discussion.

produced by a process that produces a high proportion of credences nearly matching objective chances. Following van Fraassen [1983, 1984] and Lange [1999], perhaps a justified credence is one that is the output of a process that is calibrated; i.e., a process that produces credences that match actual or potential frequencies. Or perhaps, as Tang [2016] suggests, expanding on Alston [1988, 2005], a justified credence is one that is based on some ground, where the objective probability of the credence having true content given that it is based on that ground matches the credence. [15] In each instance, we have a general formula for creating experimental resolutions: stipulate that the agent's credences are justified or at least align with the truth, and so match objective chances.

Even if one does not accept these theories of the justification of credences—for example, one prefers the concept of *accuracy*[16]—they can generate idealized experimental resolutions. The important point is that an agent's credences over propositions can be used in an idealized experimental resolution by stipulating that they match objective chances.

Now the main point. Inability to construct idealized experimental resolutions of certain decision problems should not affect our evaluation of problems where it is possible to construct such a resolution. This is analogous to a common observation regarding decision making under risk (where probabilities are known) versus decision making under ignorance (where they are not). Even if one accepts a principle for decision making under ignorance, such as minimax regret, one should not insist on it when probabilities are known. (Rather, in those cases one should maximize expected utility.) Similarly for experimental resolutions: when an experimental resolution *can* be formed, there is no need to apply principles meant to deal with cases in which it cannot.

Furthermore, just as some believe that supposed decisions under ignorance can always be transformed into decisions under risk—perhaps via a principle of indifference—some may believe that one can always create experimental resolutions from well-formed decision problems. Others may believe that it is not always possible. An experimentalist about decision theory need not take a stand on this.[17]

[15] Beebee and Papineau [1997] make a similar proposal.

[16]See Joyce, 1998. Also see Tang, 2016, 74–6 for objections to this view.

[17]Examples of decision problems where it is difficult to construct idealized experimental resolutions suitable for studying act-efficacy involve *causal unratifiability*. For example, in Gibbard and Harper's [1981, 185–6] *Death in Damascus,* rational deliberation results in shifting calculations of the causal consequences of each act, and thus shifting idealized experimental resolutions of the decision problem. The same is true in "asymmetric" variants of *Death in Damascus*, in which one location is more pleasant than the other. (See also Egan's [2007] *Psychopath Button.* Interestingly, the suggestion that such cases lack an idealized experimental resolution suitable for studying act-efficacy supports Joyce's [2012] contention that it is permissible in these cases for the agent to choose any available act.)

In contrast, it is easy to construct an idealized experimental resolution of *Death in Damascus* that is suitable for studying *decision-theory efficacy*. If Death's reliability is nearly perfect, then nearly all decision makers end up dead regardless of their decision-making theory. The idealized experimental resolution of *Death in Damascus* suitable for studying decision-theory efficacy, therefore, does not support any theory. However, in the asymmetric variant, all decision makers end up dead, but some end up dead *and* spend their last day in a less pleasant location. Thus, from the perspective of decision-theory efficacy, we want a theory that recommends the more pleasant location.

Similar points concerning decision-theory efficacy apply to Ahmed's [2014a] variant in which the agent is offered the chance to pay $1 to base their decision on the flip of an indeterministic coin, which Death cannot predict. As Ahmed [2014a, 589] points out, half of those who purchase the coin survive, while nearly all of those who do not die. Therefore, from the perspective of decision-theory efficacy, we want a decision theory that recommends paying.

2.3 Further Applications

We have seen how the experimental approach applies to the classic and transparent versions of Newcomb's problem. Let us now apply it to the medical Newcomb problem and in elucidating the difference between idealized and non-idealized decision theory.

2.3.1 Medical Newcomb Problems

Here is Andy Egan's [2007, 94] version of the medical Newcomb problem:

> *The Smoking Lesion*
>
> Susan is debating whether or not to smoke. She believes that smoking is strongly correlated with lung cancer, but only because there is a common cause—a condition that tends to cause both smoking and cancer. Once we fix the presence or absence of this condition, there is no additional correlation between smoking and cancer. Susan prefers smoking without cancer to not smoking without cancer, and she prefers smoking with cancer to not smoking with cancer. Should Susan smoke?

It seems clear that Susan should smoke, but *why*? According to nonexperimental approaches to decision theory, the answer is that smoking is more intuitive, or that smoking is supported by a set of axioms that are more intuitive than those that support not smoking. The experimentalist has a better answer.

The description of the case makes clear that Susan takes her current decisions to have no causal effect on the formation of the lesion. Susan also believes that her decision theory has no causal effect (at any point) on the formation of the lesion (otherwise, this would be a classic Newcomb problem and not a medical one). We incorporate this information into the idealized experimental resolution of the decision problem. In the idealized experimental resolution, the formation of the lesion is determined by an independent process that results in a percentage of the population forming the lesion. (To assume otherwise would be to allow Susan's belief inaccuracy to affect the dependent variable, exactly the sort of the effect that we hope to remove in idealized decision theory). When we analyze the case given this assumption, we infer that adjustments to the decision Susan makes or the decision theory she employs have no causal effect on the development of cancer. Rather, changes to these variables only result in changes to Susan's pleasure in smoking.

Refraining from smoking, and any decision theory which recommends doing so, are thus irrational. Contrary

However, it is less clear that paying is causally *act efficacious*. This is because, at least prima facie, the causal consequences of traveling to Damascus (Aleppo) after viewing the coin are the same as the causal consequences of traveling to Damascus (Aleppo) without viewing the coin—except for the lost $1. Those committed to the idea that paying for the coin is rational may therefore need to deny that rationality is determined by causal act efficacy.

A possible response on behalf of CDT would dispute Ahmed's characterization of the alternatives. Ahmed imagines the alternatives to be i) go to Damascus, ii) go to Aleppo, or iii) pay to flip the coin. But according to Joyce [2012], deliberation about whether to choose Damascus or Aleppo results in a "tie," and either is permissible. The agent must therefore engage some sort of tiebreaking procedure. Thus, the agent's available alternatives ultimately are i) engaging a deterministic tiebreaking procedure which outputs Damascus or Aleppo, or ii) paying $1 to engage an indeterministic tiebreaking procedure which outputs Damascus or Aleppo. CDT recommends (ii) because that has a lower chance of outputting Death's location. The same treatment would then need to be used to amend Lewis' [1981a, 29–30] proposed solution to the Hunter-Richter problem.

to standard methodology, refraining from smoking is not irrational because it is directly or indirectly unintuitive, but because it is efficacious in preventing smoking pleasure and not cancer. These effects can all be observed, in principle, in the idealized experimental resolution of *The Smoking Lesion*.

2.3.2 Difference between Idealized and Non-Idealized Decision Theory

Not all of decision theory employs idealizing conditions. There is research into non-idealized decision theory, typified by the work of Pollock [2006] and Weirich [2004]. On the experimental approach, we can neatly distinguish idealized and non-idealized decision theory, and in a way that respects the idea that both types of theories are studying the same concept of rationality.[18]

Idealized decision theory attempts to control for effects of limitations in the agent's epistemic perspective in determining the efficacy of acts or decision theories. Thus, for idealized decision theory, cognitive limitations and inadequate time for calculation represent confounding variables that should be controlled for in an experimental resolution. In contrast, non-idealized decision theory attempts to determine efficacy without controlling for these factors. Thus, for non-idealized decision theory, cognitive limitations and inadequate time for calculation are the conditions under which efficacy is to be determined. In short, on the experimental approach, *both idealized and non-idealized decision theory study efficacy*, but under different experimental resolutions.

For example, when studying the efficacy of decision theories under time limitations, a non-idealized experimental resolution would stipulate a set of decision problems, the cognitive abilities of the agent, and a time limit for calculating decisions. Then, various decision theories would be evaluated as the independent variable. The theories that produce the best weighted average of success—in relation to the decision problems, cognitive ability, and time limit—are the better theories.

When studying the efficacy of acts, rather than decision theories, the distinction between idealized and non-idealized decision theory is less clear. While the weighted average of success produced by a *decision theory* depends on the agent's cognitive abilities and calculation time, that produced by an *act* does not. An act produces the same results whatever process is used to select it. Therefore, the "best act" for a non-idealized agent with cognitive limitations and time constraints matches the best act for an idealized agent.

The experimental approach to non-idealized decision theory, therefore, focuses on decision theories as the independent variable, not acts. In the next section, we explore reasons to adopt decision theories as the independent variable for idealized decision theory. A positive upshot of doing so would be the creation of an even more unified research program for studying rational decision making.

3 Decision Theories as the Independent Variable in Idealized Decision Theory

In some contexts it is best to study the efficacy of acts and in others that of decision theories. In this section, I explain the motivation for thinking of decision theories as the appropriate object of study for determining decision-making rationality in idealized decision theory (e.g., in determining whether it is rational to one- or two-box), and I discuss the sorts of theories that follow.

[18] Non-idealized decision theory is often thought to take inspiration from the "bounded rationality" idea, but this can be misleading. In fact, Pollock and Weirich see themselves as studying decision-making rationality in "realistic" situations, not as studying a different type of rationality. Weirich [2004, 6] explains: "A standard of bounded rationality does not introduce a new concept of rationality but rather applies the usual concept in cases where idealizing assumptions are relaxed and agents face obstacles such as a shortage of time for reaching a decision."

3.1 Why Study the Efficacy of Decision Theories when Determining Rationality?

Clearly an agent's physical and cognitive abilities, false beliefs, and bad luck can prevent her from attaining her goals, but it is odd to regard her *rationality* as a further hindrance. Instead, we might think that if we specify the goals and eliminate all differences in abilities, beliefs, and luck, then the decision makers that tend to do best are those who are most rational. This, anyway, is my construal of the best general motivation behind success-first decision theories, especially those of Gauthier [1986] and McClennen [1990].

Nevertheless, even if it is odd to regard practical rationality as a hindrance, this is exactly what it must be if we identify rational decision making by determining the efficacy of acts. There are many cases, like Newcomb's problem, where those employing a decision-making theory that recommends the most efficacious acts leads to bad results (in comparison to the results of employing certain other decision-making theories). These cases often involve—as in Newcomb's problem—predictions of an agent's choices.

One might respond that here the predictor is "rewarding for irrationality." Sometimes the remarks of Gibbard and Harper [1981, 181] and Lewis [1981b, 377] are interpreted in this way. However, I interpret them as claiming merely that the predictor rewards for a predicted one-box decision, which is irrational, rather than claiming that the predictor is "rewarding for" irrationality. The latter claim, after all, is false. The reward condition in Newcomb's problem is not a prediction of irrationality (a case with this reward condition would be very different from Newcomb's problem). Rather, the reward condition is a prediction of one-boxing. Since two-boxing is the causally efficacious act, those who think rationality is determined by the causal efficacy of acts should just accept that rational agents end up worse off, without attempting to soften the blow by appealing to the false claim that the predictor's reward condition is irrationality.[19] If, on the contrary, one is inclined to reject a decision theory because it leads agents to disaster (under idealizing conditions), then the way to do that is to study the efficacy of decision theories.[20] The ultimate justification for this perspective is that decision theories are the proper independent variable.

3.2 Appropriate Experimental Resolutions

As we saw in Section 2, a strength of the experimental approach is that it makes clear the value of identifying suitable idealizing conditions when studying idealized decision theory. One obvious idealizing condition is belief accuracy. It would be absurd to demand that idealized decision theories perform well despite belief inaccuracy. This would be like demanding that a flu treatment cure patients when they mistakenly believe that they have the flu. As discussed in Section 2, the effects of belief inaccuracy should be controlled for by studying experimental

[19] Cases involving an infinite number of decisions, such as Arntzenius et al.'s [2004, 262] *Satan's Apple*, provide non-prediction examples in which employing a decision-making theory that recommends always choosing the most causally efficacious acts leads to bad results. Arntzenius et al.'s [2004, 267] suggestion that such agents are being "punished for" their inability to self-bind is just as problematic as the claim that agents in Newcomb's problem are being punished for their rationality. As Meacham [2010] points out, decision problems are individuated partly by the acts available, and thus cases in which binding acts are available are simply different decision problems. Cases like *Satan's Apple* represent a problem for agents who employ causal act-efficacy decision theories *precisely because* each act on the road to disaster is optimally efficacious. They have this feature *by design*. If, on a metatheoretic level, one cares only about the causal efficacy of acts, then there is no puzzle to be solved: one should grant that such agents are led to disaster.

[20] Presumably, even supporters of focusing on act efficacy like Arntzenius et al. [2004] feel some of this impulse, or they would not attempt to blame the causal decision theorists' disasters on their inability to self-bind.

resolutions where agents' credences match objective chances.

The other obvious idealizing condition is unbounded cognitive processing. Idealized decision theory determines what agents should do when they are not facing processing constraints. Accordingly, such constraints should be controlled for in experimental resolutions of decision problems. To include effects of such constraints in experimental resolutions would be to engage in non-idealized decision theory.

The last restriction on appropriate experimental resolutions that I will discuss is *theory-extensionality*. (I will not refer to this as an "idealizing condition" because it seems to be important for almost all forms of non-idealized decision theory as well.) An experimental resolution is theory-extensional when the conditions under which an agent is rewarded or punished do not essentially reference the decision-making theory that the agent employs. Otherwise, it is theory-intensional. Almost all decision problems studied by decision theorists yield theory-extensional experimental resolutions. Even decision problems that seem theory-intensional are actually theory-extensional. Newcomb's problem is an example. The reward condition in the experimental resolution of Newcomb's problem is a prediction of one-boxing by a reliable predictor. To say that the predictor is reliable is exactly to specify a high chance of a one/two-box prediction when one/two-boxing will in fact occur. What matters for the reward condition is the predicted one-boxing; the reward condition makes no essential reference to the decision-making theory that the agent employs.

A theory-intensional resolution would be one where the reward condition is, for example, *being an evidential decision theorist*. This sort of case is far less interesting, and is not relevant when determining the efficacy of decision theories. Theory-intensional resolutions in which agents are punished for employing certain decision theories are similarly irrelevant when determining efficacy. For example, the fact that a demon might kill those who employ CDT is not a mark against the efficacy of the theory. To think otherwise would be like denying the efficacy of a flu treatment because it is outperformed by a placebo when all and only the patients receiving the treatment are poisoned.

It remains an open question whether, within idealized decision theory, any decision theory has optimal efficacy across all appropriate experimental resolutions. I suspect that there are no such theories, and thus that there is symmetry between the study of act- and decision-theory efficacy (since it is also true that no acts are necessarily efficacious). On the experimental approach, the necessary truths need not concern efficacy but rather *evaluation*. The act experimentalist claims: "necessarily, the rational act is contingently optimally efficacious within the resolution" The decision-theory experimentalist claims: "necessarily, the rational act is recommended by a contingently optimally efficacious decision theory within the resolution."

To sum up, there are at least three conditions for appropriate idealized experimental resolutions of decision problems: belief accuracy, unbounded cognitive processing, and theory-extensionality. If decision-theory efficacy determines rationality, then a decision theory is a better guide to rationality to the extent that it outperforms others relative to the experimental resolutions being studied.

3.3 Success-First Decision Theories

CDT attempts to isolate causal consequences of acts, while causal success-first decision theory attempts to isolate causal consequences of decision theories. Hypothetical experiments should, in each instance, aim to capture all the effects of the independent variable on the dependent variable, and so the experiment "starts" at a point in which all the effects of the independent variable can be considered.

In the case of act efficacy, this is the point an action can be performed, and potential consequences of available acts are measured from then forward. In imagining the experiment, we hold everything constant that is causally independent of the acts (or evidentially independent, if using the evidential experimental approach discussed in Section 2.1) in order to avoid confounding variables. Similarly, for decision-theory efficacy, the experiment starts at the point when a decision theory can be adopted, and potential consequences of the available acts are measured from then forward.

On the causal experimentalist approach, there exist Newcomb-like prediction problems in which it is impossible for the decision theory that an agent employs to causally influence the dependent variable, because, for example, the prediction occurs before the agent was born. When we study experimental resolutions like that, we are not studying decision-theory efficacy, but rather the efficacy of something else. We might, for example, be studying the causal efficacy of an *agent generator*, not of the agent itself. In some contexts, this is a fascinating question, but it is not the study of the resulting agents' decision-making rationality.

In the rest of this section, I evaluate the potential object-level success-first decision theories mentioned in the introduction. Armed with the experimental approach, we will be able to better appreciate the degree to which existing success-first decision theories succeed, and to formulate a plausible way to perfect them.

3.3.1 Gauthier's "Constrained Maximizer"

Gauthier's [1986] "constrained maximizer" is an agent that is disposed to consider not simply which acts maximize expected utility in the present, but also which acts it would have maximized expected utility to plan on choosing in the past. In much of *Morals by Agreement*, Gauthier's discussion exclusively concerns strategic interaction between bargaining partners, but he provides a generalization of his basic idea that covers decision making more generally. He proposes the following:

> *Gauthier's Proposal*: if at some time t_0 it maximizes expected utility to follow a plan that involves A-ing at some subsequent time t_1, then the agent should A at t_1. [1986, 1988/89]

I have taken a few liberties with Gauthier's proposal in light of the discussion above. In *Morals by Agreement,* Gauthier's formulation is in terms of an agent's dispositions to act. He claims that if it is rational to adopt a decision-making disposition then it is rational to subsequently act in accordance with that disposition. Formulating the theory in terms of dispositions raises problems beyond the ones discussed here.[21]

In Newcomb's problem, there is divergence between the best act in the present and the best plan in the past. At any time before the prediction is made, it would maximize expected utility to follow a one-boxing plan. Therefore, if we set t_0 to a time prior to prediction, Gauthier's proposal would recommend one-boxing. However, two-boxing is recommended if t_0 occurs after the prediction. Thus, if Gauthier's proposal is to give a different recommendation than straightforward maximization in Newcomb's problem, the point at which plans are evaluated must not be in this post-prediction region. But as it stands the proposal leaves the point of evaluation for the utility of plans unspecified. We therefore turn to suggestions for specifying this point.

3.3.2 McClennen's "Resolute Chooser"

One suggestion is to tie evaluation of the utility of plans to actual commitments or "resolutions" made by the agent. This is McClennen's [1990] proposal. When one of McClennen's "resolute choosers" determines that plan p is utility maximizing, she resolves to follow p. When subsequently faced with a decision, the agent ends

[21] For a discussion of some of the problems of the disposition formulation, see Smith, 1991. It should be noted that in his later work Gauthier [1993] retreated from this proposal. His later view is that it is only rational to follow a plan at t_1 if by so acting *one is better off than one would be had one never committed to the plan at all.* This theory retains causal efficacy in Newcomb-like problems that involve "assurances." However, it is causally inefficacious in Newcomb-like problems that involve "threats."

up “intentionally choosing to act on that resolve” by choosing the act specified by *p* [McClennen, 1990, 15]. I suggest that for our purposes it is best to discard references to psychological aspects of “resolve,” and focus on how McClennen’s proposal amounts to an alternative to standard decision theory. In this spirit, we can imagine that a “resolution” manifests in an agent dropping a temporal anchor when she recognizes a utility-maximizing plan. In the future, when the agent faces a decision, the utility-maximizing status of the action at the time of the anchor dictates what to do. So, in Newcomb’s problem, a resolute chooser who is temporally located before the prediction may see that a one-boxing plan has greater expected utility than a two-boxing plan. She resolves to one-box, thus dropping the anchor. When the decision arrives, resolute choice recommends one-boxing due to the utility-maximizing status of one-boxing at the time of the anchor.

McClennen’s proposal makes clear the past point from which plans are to be evaluated, and it does sometimes create better payouts in Newcomb problems than those received by causal decision theorists. However, McClennen’s resolute choice gives the same recommendations as CDT in situations where agents are not forewarned of predictions. This is the case in the classic Newcomb problem, where it is commonly assumed that the agent does not know of the prediction until offered the choice. The agent thus has no opportunity to make a resolution, and without a prior resolution, McClennen’s theory recommends two-boxing. This is part of a general pattern: decision rules that require an actual pre-commitment will in some cases be less success-conducive than one that allows for merely hypothetical pre-commitments [Andreou, 2008, 415–22]. In other words, any theory that requires that an agent *actually* “resolve,” “intend,” or otherwise “commit” to a plan may fail if the agent is not forewarned of impending predictions.

3.3.3 Meacham’s “Cohesive Decision Theorist”

To solve this problem with McClennen’s formulation, a theory needs to focus on hypothetical rather than actual commitments. Such a theory is formulated by Meacham [2010, 68–9], in an attempt to advise agents to choose actions that they would have bound themselves to choose. According to Meacham’s “cohesive decision theory,” agents should make decisions according to the “comprehensive strategy” (a function that maps every decision problem to one of its available acts) they would choose for themselves from the perspective of their “initial credence function.” Meacham imagines that such agents would maximize “cohesive expected utility”:

$$\mathrm{CoEU}(CS) = \sum\nolimits_{\mathrm{i}}\mathrm{ic}(w_i{:}CS)\mathrm{u}(w_i)$$

Where “ic” is the agent’s initial credences and “*CS*” is a comprehensive strategy for selecting acts given decision problems. Meacham uses “:” as a neutral connective between strategies and worlds, leaving open whether this should be understood causally or evidentially. Meacham’s formulation of expected utility builds on the standard formulation [i.e., $\mathrm{EU}(A) = \sum_i\mathrm{cr}(w_i{:}A)\mathrm{u}(w_i)$] by giving control to the agent’s initial credences rather than her current credences, and by evaluating plans (“comprehensive strategies”) instead of acts. As Meacham [2010, 69, fn. 33] notes, we might understand these initial credences as the agent’s “ur-priors.”

Meacham [2010, 69] explains that a specification of cohesive decision theory with the “standard form” would rely on an agent’s current credences about her initial credences:

$$\mathrm{CoEU}(CS) = \sum\nolimits_i\mathrm{cr}(\mathrm{ic}_i)\sum\nolimits_j\mathrm{ic}_i(w_j{:}CS)\mathrm{u}(w_j)$$

Where “cr(ic)” is the agent’s current credences about her initial credences.

Under the experimental approach, we should immediately apply two Newcomb tests when considering the efficacy of cohesive decision theory. The first, as revealed in Section 2.1, is that a theory should recommend one-boxing in the classic Newcomb problem. The second, as revealed in Section 2.3.1, is that the theory should recommend the two-boxing equivalent in medical Newcomb problems. These represent two of the most basic measures when testing a success-first decision theory.

For cohesive decision theory to pass the second test, we must interpret ":" as recommended by CDT. Otherwise, the theory will fail to recommend the two-boxing equivalent in medical Newcomb problems. Thus, the connection between comprehensive strategies and worlds must be spelled out by the agent's causal dependency hypothesis, and more precisely, the agent's *ur-causal dependency hypothesis* (more precisely still, by the agent's current credences about the agent's ur-causal dependency hypothesis). Doing so will satisfy the medical Newcomb problem requirement. For example, as we saw in *The Smoking Lesion*, in the idealized experimental resolution the decision to refrain from smoking is efficacious only in preventing smoking pleasure, not in preventing cancer. The same is true of a nonsmoking plan from the ur-priors. This allows cohesive decision theory to pass the second test, but it prevents it from passing the first.

Appealing to ur-causal dependency hypotheses is a cumbersome way to attempt to produce a one-box recommendation in the classic Newcomb problem, and it is unclear that it succeeds. For example, consider a Newcomb's problem in which the prediction takes place on July 1st, 2017 at noon. Now consider which comprehensive strategy would maximize causal expected utility from the ur-priors. If the prediction has already occurred, then a two-boxing strategy is most efficacious; whereas, if the prediction has not occurred, then a one-boxing strategy is most efficacious. Whether it is better to "bind" oneself to a one- or two-boxing strategy thus depends on where the agent is in time—from the ur-priors, the agent asks, "Am I located before or after noon on July 1st, 2017?"

Giving control to the ur-priors is like stripping away the agent's evidence, and it is difficult to think about the probabilities an agent would assign to locations in time if she had no evidence whatsoever. However, given that the best strategy for the causalist in Newcomb problems crucially depends on these probabilities, we must find some way to determine them if cohesive decision theory is to produce a recommendation. These credences seem vague or impossible to specify.[22] This problem might be solved by a new form of success-first decision theory—the "functional decision theory" of Soares and Levinstein [manuscript] and Soares and Yudkowsky [manuscript]—which does not focus on causation and time but rather logical counterfactuals regarding the output of the agent's decision function.[23]

[22] Even if we grant the possibility of determinate locational probabilities from the ur-priors, the recommendations of cohesive decision theory would still fall far short of an optimally efficacious decision theory. From the ur-priors, presumably, the probability that an agent is located prior to the prediction is the sum of the probabilities of the possible locations prior to noon on July 1st, 2017. If the probability of being in this region falls below a certain threshold (e.g., .001 in the classic Newcomb problem if the predictor has near-perfect reliability), then cohesive decision theory recommends two-boxing. We can lower the probability threshold by imagining that the prediction occurs nearer to the beginning of the total possible days that an agent could be located, or by lessening the difference between the money in the transparent box and the potential reward in the opaque box. These changes affect what cohesive decision theory recommends, even though the causal efficacy of employing a decision theory that recommends one-boxing remains optimal.

[23] The *motivations* for functional decision theory seem to align perfectly with cohesive decision theory. Soares and Yudkowsky [manuscript, 9] write: "Pre-commitment requires foresight and planning, and can require the expenditure of resources—relying on ad-hoc pre-commitments to increase one's expected utility is inelegant, expensive, and impractical...FDT agents simply act as they would have ideally pre-committed to act." Whether functional decision theory makes good on this claim is best

To sum up, the experimental approach to studying the efficacy of decision theories provides a backdrop from which to evaluate the success-first decision theories of Gauthier, McClennen, and Meacham. Their analyses go astray but are moving in the right direction. The justification of a success-first decision theory, I suggested, is not to be located, precisely, in its "cohesiveness," in its being an "asset" rather than a "liability," or related notions. Rather, success-first decision theories succeed exactly when they produce the best average payout in appropriate experimental resolutions of decision problems.

4 Conclusion

This paper has presented an experimental approach to decision theory, which is compatible with the general aims of both CDT and EDT. When evaluating causal efficacy, we should use decision-theoretic experiments to isolate the causal consequences of acts or decision theories. For the causal experimentalist, the disagreement between one- and two-boxing in Newcomb's problem lies in a disagreement over the appropriate independent variable for decision-theoretic experiments. In line with CDT, two-boxing is supported when acts serve as the independent variable. In line with success-first decision theory, one-boxing is supported when decision theories serve as the independent variable. Several success-first decision theories have been evaluated using the causal experimental approach, and none has been found entirely satisfactory, though the potential for research seems bright. Ultimately, the goal for success-first decision theories is maximization of success in appropriate experimental resolutions of decision problems.[24]

References

Arif Ahmed. Dicing with Death. *Analysis*, 74(4):587–92, October 2014a.

Arif Ahmed. *Evidence, Decision and Causality*. Cambridge University Press, 2014b.

William P. Alston. An Internalist Externalism. *Synthese*, 74:265–83, 1988.

William P. Alston. *Beyond "Justification": Dimensions of Epistemic Evaluation*. Cornell University Press, 2005.

Chrisoula Andreou. The Newxin Puzzle. *Philosophical Studies*, 139(3):415–22, 2008.

Frank Arntzenius. No Regrets, or: Edith Piaf Revamps Decision Theory. *Erkenntnis*, 68(2): 277–297, 2008.

Frank Arntzenius, Adam Elga, and John Hawthorne. Bayesianism, Infinite Decisions, and Binding. *Mind*, 113(450):251–83, 2004.

Helen Beebee and David Papineau. Probability as a Guide to Life. *The Journal of Philosophy*, 94(5):217–43, May 1997.

Michael E. Bratman. *Intentions, Plans, and Practical Reason*. CSLI Publications, 1999.

John Broome. Are Intentions Reasons? And How Should We Cope with Incommensurable Values? In Christopher Morris and Arthur Ripstein, editors, *Practical Rationality and Preference: Essays for David Gauthier*, pages 98–120. Cambridge University Press, 2001.

Andy Egan. Some Counterexamples to Causal Decision Theory. *Philosophical Review*, 116(1): 93–114, 2007.

Rawad El Skaf and Cyrille Imbert. Unfolding in the Empirical Sciences: Experiments, Thought Experiments and Computer Simulations. *Synthese*, 190(16):3451–74, November 2013.

Marc Fleurbaey and Alex Voorhoeve. Decide as You Would with Full Information! An Argument against *Ex Ante* Pareto. In Nir Eyal, Samia A. Hurst, Ole F. Norheim, and Dan Wikler, editors, *Inequalities in Health: Concepts, Measures, and Ethics*, pages 113–28. Oxford University Press, 2013.

judged within the experimentalist framework.

[24] I would like to thank Arif Ahmed, Adam Bales, and Andy Egan for providing detailed comments on earlier drafts of this paper. I also thank Matthew Graves, James Joyce, Ben Levinstein, Mary Salvaggio, Nate Soares, Jeremy Strasser, and David Thorstad for helpful comments and discussion.

David Gauthier. *Morals by Agreement*. Oxford University Press, 1986.

David Gauthier. In the Neighbourhood of the Newcomb-Predictor (Reflections on Rationality). *Proceedings of the Aristotelian Society*, 89:179–94, 1988/89.

David Gauthier. Assure and Threaten. *Ethics*, 104(4):690–721, July 1993.

Allan Gibbard and William L. Harper. Counterfactuals and Two Kinds of Expected Utility. In Robert Stalnaker William L. Harper and Glenn Pearce, editors, *Ifs: Conditionals, Belief, Decision, Chance, and Time*, pages 153–90. D. Reidel, 1981.

Alan Hájek. A Puzzle about Degree of Belief. Manuscript.

William James. The Will to Believe. In J. J. McDermott, editor, *The Writings of William James*. Random House, 1896.

James M. Joyce. A Nonpragmatic Vindication of Probabilism. *Philosophy of Science*, 65(4): 575–603, December 1998.

James M. Joyce. *The Foundations of Causal Decision Theory*. Cambridge University Press, 1999.

James M. Joyce. Regret and Instability in Causal Decision Theory. *Synthese*, 187(1):123–45, 2012.

Gregory S. Kavka. The Toxin Puzzle. *Analysis*, 43(1):33–6, 1983.

Marc Lange. Calibration and the Epistemological Role of Bayesian Conditionalization. *The Journal of Philosophy*, 96(6):294–324, June 1999.

David Lewis. Causal Decision Theory. *Australasian Journal of Philosophy*, 59(1):5–30, March 1981a.

David Lewis. 'Why Ain'cha Rich?'. *Nous*, 15(3):377–380, September 1981b.

Edward F. McClennen. *Rationality and Dynamic Choice*. Cambridge University Press, 1990.

Christopher J. G. Meacham. Binding and its Consequences. *Philosophical Studies*, 149(1): 49–71, 2010.

Judith Mehta, Chris Starmer, and Robert Sugden. The Nature of Salience: An Experimental Investigation of Pure Coordination Games. *The American Economic Review*, 84(3):658–73, June 1994.

D. H. Mellor. Objective Decision Making. In *Matters of Metaphysics*, pages 269–87. Cambridge University Press, 1991.

Robert Nozick. Newcomb's Problem and Two Principles of Choice. In Nicholas Rescher, editor, *Essays in Honor of Carl G. Hempel*, pages 114–146. D. Reidel, 1969.

Derek Parfit. *Reasons and Persons*. Oxford University Press, 1984.

John L. Pollock. *Thinking about Acting: Logical Foundations for Rational Decision Making*. Oxford University Press, 2006.

Thomas C. Schelling. *The Strategy of Conflict*. Oxford University Press, 1960.

Holly Smith. Deriving Morality from Rationality. In Peter Vallentyne, editor, *Contractarianism and Rational Choice: Essays on David Gauthier's Morals by Agreement*, pages 229–53. Cambridge University Press, 1991.

Nate Soares and Benja Fallenstein. Toward Idealized Decision Theory. *arXiv: 1507.01986 [cs.AI]*, 2015.

Nate Soares and Ben Levinstein. Cheating Death in Damascus. *Available at https://intelligence.org/files/DeathInDamascus.pdf*, Manuscript.

Nate Soares and Eliezer Yudkowsky. Functional Decision Theory: A New Theory of Instrumental Rationality. Manuscript.

Jordan Howard Sobel. Circumstances and Dominance in a Causal Decision Theory. *Synthese*, 63(2):167–202, May 1985.

Robert Sugden. A Theory of Focal Points. *The Economic Journal*, 105(430):533–50, May 1995.

Robert Sugden and Ignacio E. Zamarrón. Finding the Key: The Riddle of Focal Points. *Journal of Economic Psychology*, 27(5):609–21, October 2006.

Weng Hong Tang. Reliability Theories of Justified Credence. *Mind*, 125(497):63–94, 2016.

Bas van Fraassen. Calibration: A Frequency Justification for Personal Probability. In R. Cohen and L. Laudan, editors, *Physics, Philosophy, and Psychoanalysis*, pages 295–319. D. Reidel, 1983.

Bas van Fraassen. Belief and the Will. *Journal of Philosophy*, 81(5):235–56, May 1984.

Paul Weirich. *Realistic Decision Theory: Rules for Nonideal Agents in Nonideal Circumstances*. Oxford University Press, 2004.