

**From the holdings of the Department of Special Collections , Princeton University Library,
Princeton, NJ**

These documents are either in the public domain, or can only be used for educational and research purposes (Fair use) as per U.S. Copyright law (text below). By accessing this file, all users agree that their use falls in the public domain or within fair use as defined by the copyright law. We do not charge any permission or usage fees for the publication of images of material in our collections, including those provided by Princeton University Library, via our website, catalog, or directly from staff. For more information regarding copyright, credit, and citation please review the information on the Copyright, Credit and Citation Guidelines page of our website.

<https://library.princeton.edu/special-collections/policies/copyright-credit-and-citation-guidelines>. If you have any questions, including how to order higher-quality images from the collection, contact Special Collections Public Services via our Ask Us! Form. <https://library.princeton.edu/special-collections/ask-us>

Doctoral Dissertations, 2011-2026

Philosophy

2016

Decision Theory and Deep Ignorance

Isaacs, Yoaav Isaacs

Princeton, NJ : Princeton University

<http://arks.princeton.edu/ark:/88435/dsp01pk02cd22x>

U.S. Copyright law: The copyright law of the United States (Title 17, §108, United States Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research. If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of fair use that user may be liable for copyright infringement. This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve a violation of copyright law. <https://www.copyright.gov/title17/title17.pdf>

Decision Theory and Deep Ignorance

Yoaav Isaacs

A DISSERTATION
PRESENTED TO THE FACULTY
OF PRINCETON UNIVERSITY
IN CANDIDACY FOR THE DEGREE
OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE
BY THE DEPARTMENT OF PHILOSOPHY

[Adviser: Adam Elga]

September 2016

©Copyright by Yoaav Isaacs, 2016. All rights reserved.

Abstract

Decision theory is concerned with how agents should choose under uncertainty. I argue that standard decision theories overlook uncertainty about *oneself*—they offer inadequate guidance to agents who lack self-knowledge, and develop alternative theories that do better. I show that a decision theory for agents who are uncertain about their own credences, utilities, available acts, and decision functions can respond to such uncertainties with unexpected nuance. I develop a decision theory for agents who are uncertain about their evidence which is ideally suited for followers of a deontological ethics. I evaluate the uncertainties occasioned by transformative experiences. Lastly, I defend the decision-theoretic significance of improbable possibilities, without which my fine-grained treatment of an agent's lack of self-knowledge would be unworkable.

Table of Contents

Abstract, i

Table of Contents, ii

Self-Confident Decision Theory, 1

Duty and Knowledge, 28

The Problems of Transformative Experience, 44

Probabilities Cannot Be Rationally Neglected, 62

Introduction

Decision theory is concerned with how agents should choose under uncertainty. I argue that standard decision theories overlook uncertainty about *oneself*—they offer inadequate guidance to agents who lack self-knowledge. I will develop alternative theories that do better. In my first chapter, I develop a decision theory for agents who are uncertain about their own credences, utilities, available acts, and decision functions. In my second chapter, I develop a decision theory for agents who are uncertain about their evidence. In my third chapter, I evaluate the uncertainties occasioned by transformative experiences. In my fourth chapter, I defend the decision-theoretic significance of improbable possibilities, without which my fine-grained treatment of an agent's lack of self-knowledge would be unworkable.

1 Self-Confident Decision Theory

Major arguments in decision theory turn on exploiting an agent's lack of self-knowledge. I show how traditional decision theories can be amended so that even agents lacking self-knowledge can avoid being exploited.

One would expect that any agent with coherent credences who takes a bet on a contingent proposition can win that bet. But this is not so. When agents are uncertain about what they think, what they want, what they can do, or how they choose what to do, they may take bets that they cannot win.

Suppose an agent has high credence that it's raining, but also has high credence that he doesn't have high credence that it's raining. Standardly, that agent would take a bet at even odds that he doesn't have high credence that it's raining. After all, he's being offered a bet at even odds on a proposition he thinks is true. But that agent cannot win that bet. The agent's decision-theoretic state licensed a bet which it ensured he would lose.

Losing a bet about what one thinks about rain may not seem to be a big deal. In fact, it may not be a big deal. It would not be onerous to accept that when agents lack self-knowledge, they reasonably take bets which they have no chance of winning. But not all unfortunate bets are so bland.

Suppose that there are two agents who each have high credence that they are causal decision theorists. One of them is, in fact, a causal decision theorist, while the other is an evidential decision theorist. They are each offered a bet at even odds that they are evidential decision theorists, and they each take the bet. The causal decision theorist is doomed to lose the bet while the evidential decision theorist is not. Of course, it's also easy to alter the setup so that that the evidential decision theorist is doomed to lose the bet while the causal decision theorist is not. Unfortunate bets can be tailored to causal decision theorists or to evidential

decision theorists; these unfortunate bets exploit one sort of agent's lack of self-knowledge, but not the other sort's. Indeed, major arguments in decision theory rely on just this sort of tailored unfortunate bet. Andy Egan's *Psychopath Button* is tailored to exploit causal decision theorists' lack of self-knowledge and Allan Gibbard and William Harper's *Medical Newcomb Problem* is tailored to exploit evidential decision theorists' lack of self-knowledge. Since the exploitation of agents' lack of self-knowledge is at stake in major arguments in decision theory, it's worth seeing what can be done about it.

By modifying standard decision theories, it is possible to provide agents with many of the benefits of self-knowledge despite their lack of it. In particular, it is possible to prevent agents from taking any unfortunate bets. I call such modified decision theories self-confident. Standard decision theories maximize some sort of expected utility (causal, evidential, etc.) relative to an agent's unconditional credences. But self-confident decision theories maximize expected utility relative to an agents' conditional credences, specifically their credences conditional on what they actually think, what they actually want, what they actually can do, and how they actually choose what to do.

Self-confident causal decision theorists will not fall prey to the *Psychopath Button* and self-confident evidential decision theorists will not fall prey to the *Medical Newcomb Problem*. These self-confident decision theories are not subject to counterexample as their standard analogues are. The contest between causal decision theory and evidential decision theory should thus be superseded by a contest between self-confident causal decision theory and self-confident evidential decision theory.

2 Duty and Knowledge

Deontological ethics needs a decision theory. But traditional decision theory is unsuitable for specifying when an agent may perform an action which may or may not violate a deontological scruple. I show that a novel decision theory—one that is based on knowledge-first epistemology—suits deontological ethics much better.

An omniscient agent can easily follow a deontological ethic. All he has to do is abstain from any action which is categorically forbidden. But how should an agent acting under uncertainty follow a deontological ethic?

Three standards present themselves, but none are promising: (1) The agent might be permitted to perform an act so long as that agent was not certain that doing so would violate a deontological scruple, (2) the agent might be permitted to perform an act only so long as he was certain that doing so would not violate a deontological scruple, or (3) the agent might sometimes be permitted to perform an act and sometimes be forbidden to perform an act when that agent is uncertain

of whether or not the act would violate a deontological scruple. The rarity of certainties means that the first standard would permit virtually any act and the second standard would forbid virtually any act. Neither is tenable. And the third standard, though not as egregiously unacceptable, does seem unpalatably arbitrary.

I argue that the decision theory for deontological ethics should be formulated through a knowledge-first epistemology. An agent should maximize expected utility relative to his evidential probabilities (probabilities conditional on what the agent knows) subject to the constraint that the agent abstain from any action which he does not know to be deontologically licit.

This knowledge-first decision theory can only offer imperfect guidance, as agents are not always in a position to know what they know. However, it is plausible that any decision theory for imperfect agents (like us) can only offer imperfect guidance. And a knowledge-first decision theory offers massively more guidance than could be had without a decision theory. There is thus good reason for those who favor deontological ethics to go knowledge-first in epistemology, and, similarly, for those who disfavor knowledge-first epistemology to reject deontological ethics.

3 The Problems of Transformative Experience

Transformative experiences pose a challenge for decision theory. Agents facing transformative experiences are impoverished relative to their decision problems, and decision theory has a hard time handling impoverished agents. Agents facing epistemically transformative experiences lack utilities for experiences they cannot fathom. Agents facing personally transformative experiences lack utilities for being radically different than they are. And it is unclear how decision theory should evaluate agents with utility gaps.

Several changes to decision theory have been proposed in light of the problems of transformative experience. Richard Pettigrew has argued that the problem of epistemically transformative experiences should be addressed by redescribing the decision problems in which epistemically transformative experiences arise. Agents should take the expected value of what their utilities for a transformative experience would be after having that experience, and then use that expected value of a utility as a surrogate for the utility they lack. Pettigrew has also argued that the problem of personally transformative experiences should be addressed by changing the nature of agents' utility functions. This new sort of utility function, which Pettigrew dubs a local utility function, represents how an agent values possible present time slices. Pettigrew's proposal is to aggregate local utility functions together to make what he calls a global utility function, and then to use that global utility function as traditional decision theory uses a traditional utility function.

These proposed changes are not viable, as they each face two insurmountable

problems. (1) Both proposals illegitimately give the agent's future preferences immediate decision-theoretic significance. Only an agent's current preferences are of immediate decision-theoretic significance, and an agent's current preferences are seriously constrained by utility gaps. An agent cannot prefer that his future preferences for some option be satisfied when his current utilities for that option are undefined. Any such preference would be a preference regarding his current option, thus contradicting the initial supposition that the agent's preferences regarding that option were undefined. Utility gaps run deep. (2) Both proposals illegitimately take averages across different utility functions. But averages across different utility functions are neither mathematically nor philosophically defined. Mathematically, utilities are unique only up to positive affine transformation—they lack the structure needed to allow the sort of averaging that Pettigrew requires. Philosophically, utilities show only relative desirabilities, and standings in two different hierarchies of relative desirability can't be averaged together. (Local utility functions also face a further, distinctive problem. Local utility functions would not be able to capture the range of preferences that ordinary utility functions can capture. Most notably, local utility functions would make it impossible for an agent to care about what happens after his death. This is obviously unacceptable.)

Pettigrew's proposals are untenable. The problems of transformative experience stand.

4 Probabilities Cannot Be Rationally Neglected

In response to Nicholas Smith, I argue that probabilities cannot be rationally neglected. I show that Smith's proposal for ignoring low probability outcomes must, on pain of violating dominance reasoning, license taking arbitrarily much risk for arbitrarily little reward.

Self-Confident Decision Theory¹

There are two major positions in decision theory: causal decision theory and evidential decision theory. Causal decision theorists have worked to make evidential decision theory look foolish, evidential decision theorists have worked to make causal decision theory look foolish, and they've each been depressingly successful. Causal decision theory seems to prescribe a foreseeably foolish course of action in Andy Egan's *Psychopath Button*, and evidential decision theory seems to prescribe a foreseeably foolish course of action in Allan Gibbard and William Harper's *Medical Newcomb Problem*. Neither maximizing causal expected utility (as causal decision theory would have us do) nor maximizing evidential expected utility (as evidential decision theory would have us do) gets us everything we want from our decision theory.

I will argue that the *Psychopath Button* and the *Medical Newcomb Problem* share an important feature—they each exploit an agent's uncertainty about his own decision-theoretic state. If an agent is uncertain about his decision theoretic state, then if he maximizes expected utility (be it causal expected utility or evidential expected utility) he can be made to act foolishly. I will show how to modify causal decision theory and evidential decision theory to safeguard agents from the perils of this particular sort of ignorance. These modified decision theories require agents to maximize *conditional* expected utility—expected utility conditional upon an agent's actual decision-theoretic state. I call such modified decision theories “self-confident”. Self-confident causal decision theorists will not fall prey to the *Psychopath Button* and self-confident evidential decision theorists will not fall prey to the *Medical Newcomb Problem*. These self-confident decision theories are not subject to counterexample as their standard analogues are. The conflict between causal decision theory and evidential decision theory should thus be superseded by a conflict between self-confident causal decision theory and self-confident evidential decision theory.

5 A Little Predicament

The great decision theorist, Mick Jagger, once said, “You can't always get what you want. But if you try sometimes, you just might find you get what you need.” Jagger's point, I take it, was this: A rational agent may well lose some of the bets he takes. (You can't always get what you want.) However, a rational agent will not take any bets which he could not possibly win. (You just might find you get what you need.)

¹Delivered as a talk at the Society for Exact Philosophy in June 2014.

Jagger’s Principle: A rational agent will not take any bet which he cannot win.

Jagger’s principle seems plausible. Unfortunately, Jagger’s principle conflicts with another plausible decision-theoretic principle: the principle of expected utility maximization.

Expected Utility Maximization: A rational agent will perform an act only if his expected utility conditional upon that act is greater than or equal to his expected utility conditional upon any alternative act.

If agents always act so as to maximize their expected utilities, they will sometimes take bets which they cannot win.

It may seem outlandish that Jagger’s principle and expected utility maximization should conflict. After all, it is well known that if an agent is representable as an expected utility maximizer then that agent cannot be Dutch booked, and a bet that an agent is guaranteed to lose looks an awful lot like a Dutch book.²

Here’s the trick: A Dutch book is a collection of bets which guarantee a loss for any possible agent. But some bets can guarantee a loss for some specific agent without guaranteeing a loss for any possible agent. Thus an agent can violate Jagger’s principle without being Dutch booked. But this is possible in only one sort of circumstance—when an agent takes a losing bet about his own decision-theoretic state.³

A decision-theoretic state is composed of the factors which determine how agents act. The acts of an agent are determined by what the agent thinks, what the agent wants, what the agent can do, and how the agent chooses from among the things he can do on the basis of what he thinks and what he wants.⁴ So why does an agent’s ignorance about these factors cause problems which his ignorance about other matters doesn’t? Some cases will clarify the matter:

Case 1:

Suppose that Bob has credence $\frac{1}{2}$ that it will rain. Bob is then offered a bet about the rain at favorable odds—he’ll win \$2 if it rains and only

²See Lehman (1955) and Kemeny (1955).

³More formally, a bet is a Dutch book if it guarantees a loss given any consistent assignment of truth-values to the sentences in the agents language. A bet violates Jagger’s principle if it guarantees a loss given the actual truth-values of sentences about the agents decision-theoretic state and any consistent assignment of truth-values to the other sentences in the agents language. See Mahtani (2014) for more about the formal underpinnings of Dutch books.

⁴More precisely, the acts of an agent are determined by the agent’s credence function, the agent’s utility function, the set of acts available to that agent, and the agent’s decision function—a function from the agent’s credence function, utility function, and set of available acts to an element of the set of available acts. See the technical appendix for details.

lose \$1 if it doesn't rain. Bob does the sensible thing, and takes the bet. As it happens, it doesn't rain, and Bob loses \$1.

Has Bob violated Jagger's principle? Not at all. Bob lost the bet in the actual world, but he does not lose the bet in every possible world. There are worlds in which someone who thinks what Bob thinks, wants what Bob wants, can do what Bob can do, and chooses his acts the way Bob chooses his acts wins the bet that it will rain.

Case 2:

Suppose that Bob has credence $\frac{1}{2}$ that he is a causal decision theorist. Bob is then offered a bet about his decision theory at favorable odds—he'll win \$2 if he is a causal decision theorist and only lose \$1 if he is not a causal decision theorist. Bob does the sensible thing, and takes the bet. As it happens, Bob is not a causal decision theorist, and Bob loses \$1.

Has Bob violated Jagger's principle? Yes. Not only did Bob lose the bet in the actual world, he loses the bet in every possible world. There are no worlds in which someone who thinks what Bob thinks, wants what Bob wants, can do what Bob can do, and chooses his acts the way Bob chooses his acts wins the bet that he is a causal decision theorist.

The problem is this: Bob's decision-theoretic state can't vary freely the way the outside world can. Bob, just as he is, might face rain. But Bob, just as he is, could never be a causal decision theorist. Bob's ignorance about rain is therefore formally different than Bob's ignorance about Bob. Bob's attitudes towards rain don't guarantee that he'll lose a bet about the rain, but Bob's attitudes towards his decision-theoretic state do guarantee that he'll lose a bet about his decision-theoretic state.

Jagger's Principle conflicts with Expected Utility Maximization. So which principle should we hold to? It certainly makes a great deal of sense to take Expected Utility Maximization over Jagger's Principle. Expected Utility Maximization—unlike Jagger's Principle—is solidly established and well-motivated. It's not so strange to think that in contrived circumstances in which an agent is uncertain about his decision-theoretic state and yet is taking bets about his decision-theoretic state, one can rig up bets that the agent can't win. The violation of Jagger's Principle I outlined is trivial. If all violations of Jagger's Principle were comparably trivial, well, you might suspect that Mick Jagger wasn't such a great decision theorist after all.

As it turns out, though, violations of Jagger's Principle are at the heart of major arguments in decision theory. The two most prominent positions in deci-

sion theory—causal decision theory and evidential decision theory—each face situations which they seem to mishandle. Causal decision theory seems to go wrong in Andy Egan’s *Psychopath Button*, and evidential decision theory seems to go wrong in Allan Gibbard and William Harper’s *Medical Newcomb Problem*. These cases each involve violations of Jagger’s Principle, and these cases are highly non-trivial. In the *Psychopath Button*, causal decision theorists—but not evidential decision theorists—violate Jagger’s Principle, and in the *Medical Newcomb Problem*, evidential decision theorists—but not causal decision theorists—violate Jagger’s principle. In order to figure out what to make of these problem cases, we will have to see how decision theories can be amended so that they satisfy Jagger’s Principle.

6 Self-Confidence

We’ve seen that an agent who maximizes expected utility and is uncertain about what he thinks, what he wants, what he can do, or how he chooses what to do will take bets which he is sure to lose. But suppose we give up the maximization of expected utility—what do we do instead? It’s easiest to think about a specific case first. Let’s imagine a decision theory which (somehow) obeys Jagger’s principle. I’ll call it “X-decision theory”. Let’s imagine an agent who is an X-decision theorist, but who is not certain that he is an X-decision theorist. How can we make sense of an X-decision theory which would forbid our agent from betting against being an X-decision theorist, even at apparently favorable odds?

Here’s a plausible line of reasoning: It would be wrong for X-decision theory to prescribe taking such a bet under any circumstances whatsoever. Taking the bet is, after all, guaranteed to be bad for X-decision theorists. The goodness of taking the bet for non-X-decision theorists should not count in its favor as far as X-decision theory is concerned. X-decision theory only recommends acts to X-decision theorists. In effect, X-decision theory should be aware that only X-decision theorists will follow its recommendations, even if those following its recommendations are not always aware that they are X-decision theorists. X-decision theory should not be in the business of maximizing expected utility in general, but rather in the business of maximizing expected utility for X-decision theorists.

Any decision theory which obeys Jagger’s principle must take itself for granted. It must ignore any possibilities in which it is not an agent’s decision procedure. For similar reasons, such a decision theory must take what an agent thinks, wants, and can do for granted, and ignore any possibilities in which the agent thinks, wants, or can do other than he actually thinks, wants, and can do. In order to obey Jagger’s principle, a decision procedure must act as though the only possibilities of consequence are those in which the agent’s thoughts, desires, abilities, and decision procedure are as they actually are. And this can be done—a decision procedure has

all the information required to ignore possibilities in which an agent is other than he actually is. After all, a decision procedure is some way of taking what an agent thinks, what an agent wants, and what an agent can do, and using those to pick an action. A decision procedure can respond to more information about an agent than that agent is aware of.

We can easily formulate decision theories which ignore all possibilities in which the agent is other than he actually is. I will call such decision theories “self-confident”. What self-confident decision theories do with possibilities they consider will vary from theory to theory, just as what standard decision theories do with a full space of possibilities varies from theory to theory. There is a self-confident decision theory that treats its possibilities the way causal decision theory treats its possibilities, a self-confident decision theory that treats its possibilities the way evidential decision theory treats its possibilities, and so on. Instead of prescribing acts which maximize some sort of expected utility across all worlds, self-confident decision theories will prescribe acts which maximize some sort of expected utility across worlds in which the agent’s decision-theoretic state is what it actually is. Instead of evaluating acts relative to the agent’s unconditional credences, self-confident decision theories evaluate acts relative to the agent’s credences conditional upon the agent’s actual decision-theoretic state.⁵

Regarding matters that are probabilistically independent of the agent’s decision-theoretic state, the self-confident analogue of a standard decision theory will give the same prescriptions as the standard decision theory. For propositions independent of the agent’s decision-theoretic state, self-confident decision theories prescribe that an agent with credence of x in p be willing to pay up to x dollars for a bet that pays 1 dollar if p is true and nothing otherwise.⁶ But when it comes to matters that are not probabilistically independent of the agent’s decision-theoretic state, self-confident decision theories produce different behavior. Consider the limiting case of dependence: a self-confident decision theory’s prescriptions for bets regarding the agent’s decision-theoretic state. If the proposition being bet on is that the agent’s beliefs are whatever they actually are, that the agent’s desires are whatever they actually are, that the agent’s available acts are whatever they actually are, or that the agent’s decision theory is whatever it actually is, then a self-confident decision theory will prescribe that the agent be willing to pay up to 1 dollar for a bet that that pays 1 dollar if p is true and nothing otherwise. Put a bit loosely, a self-confident decision theory will have the agent act as though he were supremely confident about his decision-theoretic state, however much doubt the agent may entertain. But the affected confidence that a self-confident decision theory prescribes cannot lead an agent astray; self-confident decision theories always get an agent’s decision-theoretic state right. Self-confident decision theories act as though some unfalsified worlds

⁵See the technical appendix for details.

⁶Taking for granted that the agent’s utilities are linear in money.

were falsified, but those unfalsified worlds are always false.

7 Evaluating Self-Confidence

Self-confidence has obvious advantages. In general, one expects self-confident agents to fare better than non-self-confident agents. Self-confident agents will win bets that non-self-confident agents will lose.

Suppose that you are building a faithful robot servitor, Hal. You can program Hal to have credences, utilities, a set of available acts, and a function from his credences, utilities, and available acts to an act. Unfortunately, you just don't have the materials to make sure that Hal is always certain about his credences, utilities, set of available acts, and decision function. Hal will be insuperably uncertain of his precise decision-theoretic state. Yet despite this limitation, you can program Hal with any decision function you want. If you want Hal to make the best decisions possible, you should make sure that Hal's decision function is self-confident. By giving Hal a self-confident decision function, you can provide Hal with the benefits of the self-knowledge which you cannot actually provide him. But if self-confidence is a good thing for Hal's decision function to have, then it stands to reason that self-confidence is a good thing for our decision functions to have as well.

So what's not to like about self-confidence? Most obviously, self-confident decision functions seem completely divorced from the agent's point of view, and (traditionally) decision theory is all about the agent's point of view. Frank Jackson puts this sentiment well:

[T]he fact that a course of action would have the best results is not in itself a guide to action, for a guide to action must in some appropriate sense be present to the agent's mind. We need, if you like, a story from the inside of an agent ... and having the best consequences is a story from the outside.

A great virtue of decision theory is that it offers guidance, but agents performing strange feats of self-confidence will not understand the force that moves them. Self-confident agents may be completely unable to explain their actions. If given the opportunity to get a nickel, a self-confident agent will bet his life that he has the exact decision-theoretic state he has. He may—and likely will—expect to die, and curse the inexplicable madness which led him to take the apparently insane bet.

A great virtue of decision theory is that it offers guidance, but that virtue is already abandoned if we allow agents to be uncertain of their decision-theoretic states. Any agent who is uncertain about his credences, utilities, available acts, and decision procedure is—by definition—less than fully conscious of the factors underpinning his decision. Standard decision theories are no better-suited for guidance than self-confident decision theories are. Consider an agent who has high

credence that its going to rain tomorrow, but who is wildly uncertain about what his credence about the rain is. Suppose this agent is offered a bet at even odds that its going to rain tomorrow. This is a proposition that the agent thinks is true, so, standardly, the agent should take the bet. But the agent is in no position to be guided by that verdict. A standard decision theory makes its prescription for the bet about the rain on the basis of the agent's credence about the rain—but the agent has no idea what that credence is.

A self-confident decision procedure does not eliminate guidance value, it merely takes advantage of a circumstance in which guidance value was already eliminated. Any agent who is less than certain of his decision-theoretic state must be something of a somnambulist, moving without understanding. Why not be a somnambulist who takes winning bets?

Of course, it may not be worth thinking about any sort of somnambulist. If the value of decision theory is that it offers guidance, then circumstances in which agents cannot be guided by decision theory may simply be unimportant. There may be no point to a decision theorist evaluating an agent's choice in light of more information about the agent's decision-state than that agent had available.⁷ Not for nothing do economists make strong, idealizing assumptions which ensure that agents are completely certain of their decision-theoretic states.⁸ But if we are going to consider cases in which agents will not fully understand why they do whatever they do, the fact that self-confidence makes a difference in such cases should not be held against self-confidence.

Why should agents be self-confident, and act as though they were sure of their actual decision-state? Why not have agents be height-confident, and act as though they were sure of their actual height? Or actuality-confident, and act as though they were sure of absolutely everything? I answer that self-confidence is meaningfully different than height-confidence or actuality-confidence. Self-confidence is formulable given only the standard apparatus of decision theory, whereas height-confidence and actuality-confidence require more apparatus. But there might well be circumstances in which more than the standard apparatus should be allowed. One might be able to build an agent whose actions are sensitive to his height, and in such a case it might be a good idea to make that agent height-confident. It would, of course, be rather harder to build an actuality-confident agent. But I don't balk at height-confidence or actuality-confidence. One needs to specify the domain of decision theory—to specify the tools that decision theorists can use when coming up with their theories—and only then can one determine which decision theory is

⁷The cost of dismissing cases in which agents are uncertain about their decision-theoretic states is that it precludes decision theory from having any direct applicability for agents like us. See Lewis (1981) and Williamson (2012) for more.

⁸The weird, problematic cases are left to philosophers. Thus it is more fun to be a philosopher than an economist.

appropriate. The standard formalism of decision theory does not allow for height-confidence or actuality-confidence, but it does allow for self-confidence.

Now that we have a grip on self-confidence, we can see what difference it makes to the dialectic between causal decision theory and evidential decision theory.

8 Causal Decision Theory and Evidential Decision Theory

Causal decision theorists and evidential decision theorists agree that rational agents should maximize expected utility. They disagree only about what expected utility is, that is, about how one should calculate expected utility.

Causal decision theorists calculate the expected utility of an act according to the expected value of its effects. Evidential decision theorists calculate the expected utility of an act according to the expected value of the worlds in which that act is performed. Causal decision theorists use conditional credences which are constrained to follow their unconditional credences about the causal structure of the world. Evidential decision theorists use conditional credences which are not subject to any such constraint. Causal decision theorists evaluate acts with their credences about the causal structure of the world held fixed, whereas evidential decision theorists evaluate acts with their credences about the causal structure of the world allowed to vary.

Causal decision theory seems to prescribe a foreseeably foolish course of action in Andy Egan's *Psychopath Button*, and evidential decision theory seems to prescribe a foreseeably foolish course of action in Allan Gibbard and William Harper's *Medical Newcomb Problem*. Unfortunately, both the *Psychopath Button* and the *Medical Newcomb Problem* are loosely described, thus making analyzing them quite difficult. By analyzing more perspicuous cases first, the *Psychopath Button* and the *Medical Newcomb Problem* will become clearer.

9 Problem Cases for Causal Decision Theory

Causal decision theory mandates that an agent act so as to maximize the expected goodness of the effects of his actions. Yet this decision procedure sometimes prescribes acts which reliably cause bad effects.

The Tricky Button:

Jim, a causal decision theorist, is debating whether to press the tricky button. The tricky button causes a moderately bad result if pushed by someone who would two-box in *Newcomb's Problem* and a moderately

good result if pushed by someone who would one-box in *Newcomb's Problem*.⁹ Suppose that it is common knowledge that all one-boxers will refrain from pushing the tricky button (perhaps it is known that the Angel of Death killed all the one-boxers who would push the button). Now suppose that Jim's credence that he is a two-boxer is extremely low. Should Jim push the button?

Standard causal decision theory prescribes that Jim push the tricky button. Holding Jim's credences about the causal structure of the world fixed, the act of pushing the button has positive expected utility. Unfortunately, if Jim is a causal decision theorist then Jim is a two-boxer, and will inevitably suffer bad effects due to pushing the button. What's more, Jim is in a position to foresee these bad effects—he already knows that there is no epistemically possible world in which he pushes the button and thereby causes good effects. But if Jim is a causal decision theorist then his foresight will not avail him.

Suppose that Jim were then able to pay a dollar to nullify the effect of pushing the button. He would gladly do so. Doesn't it seem that Jim should have been able to avoid wasting that dollar? Consider a new sort of decision theorist, the Y-decision theorist, who is exactly like a causal decision theorist in all cases other than the *Tricky Button*, but who will not choose to push the tricky button. It seems that it is better to be a Y-decision theorist than a causal decision theorist. Y-decision theorists and causal decision theorists generally do the same, but in one case Y-decision theorists save a buck.

Y-decision theory is a lot like causal decision theory, but in one case Y-decision theory prescribes the act that causal decision theory would prescribe if it were taking an agent's decision-theoretic state for granted. The decision theory which always prescribes the act that causal decision theory would prescribe if it were taking an agent's decision-theoretic state for granted is self-confident causal decision theory. Self-confident causal decision theory is the logical extension of Y-decision theory. A self-confident causal decision theorist will not push the tricky button.

Now that we've got a grip on the *Tricky Button*, let's take a look at the *Psychopath Button*.

The Psychopath Button:

⁹From the Stanford Encyclopedia of Philosophy's entry on Causal Decision Theory: "In Newcomb's Problem an agent may choose either to take an opaque box or to take both the opaque box and a transparent box. The transparent box contains one thousand dollars that the agent plainly sees. The opaque box contains either nothing or one million dollars, depending on a prediction already made. The prediction was about the agent's choice. If the prediction was that the agent will take both boxes, then the opaque box is empty. On the other hand, if the prediction was that the agent will take just the opaque box, then the opaque box contains a million dollars. The prediction is reliable. The agent knows all these features of his decision problem."

Paul is debating whether to press the kill all psychopaths button. It would, he thinks, be much better to live in a world with no psychopaths. Unfortunately, Paul is quite confident that only a psychopath would press such a button. Paul very strongly prefers living in a world with psychopaths to dying. Should Paul press the button?

Holding fixed Paul's beliefs about the which underlying state he is in (in this case, whether or not he is a psychopath), the expected utility of pushing the button exceeds the expected utility of not pushing the button. Causal decision theory would thus have Paul push the button.

The *Psychopath Button* presents a stranger situation than its prose suggests. Paul's credence that he is a psychopath is low enough that pushing the button would maximize his expected utility, but Paul's credence that he is a psychopath would be raised enough if he pushed the button he would regret doing so. What is the basis of Paul's expectation that there is a correlation between pushing the button and being a psychopath? Paul's act is the product of what he thinks, what he wants, what he can do, and how he chooses from among the things he can do on the basis of what he thinks and what he wants. The correlation between pushing the button and being a psychopath must therefore be a correlation between combinations of thoughts, desires, abilities, and ways of choosing actions which result in the act of pushing the button. But the correlation is not (or at least need not be) with what Paul thinks. What Paul thinks is stipulated, and Paul can be certain of what he thinks yet still think that he is not a psychopath and still think that if he were to push the button that would be strong evidence that he is a psychopath. The correlation is not (or at least need not be) with what Paul wants. What Paul wants is stipulated, and Paul can be certain of what he wants yet still think that he is not a psychopath and still think that if he were to push the button that would be strong evidence that he is a psychopath. The correlation is not (or at least need not be) with what Paul can do. Paul's options are stipulated, and Paul can be certain of his options yet still think that he is not a psychopath and still think that if he were to push the button that would be strong evidence that he is a psychopath. The correlation between pushing the button and being a psychopath must therefore be a correlation between ways of choosing actions and being a psychopath.

The situation in *Psychopath Button* is quite similar to the situation in *Tricky Button*. In each case there is an act which tends to be bad for certain sorts of decision theorists, standard causal decision theorists are among those for whom the act tends to be bad, and standard causal decision theory nonetheless prescribes performing that act. Yet while a self-confident causal decision theorist is sure not to push the tricky button, it remains to be seen whether or not a self-confident causal decision theorist would push the psychopath button. The *Psychopath Button* is underdescribed, so a determinate analysis will require that Paul's credences be

specified more fully.

(1.) Would Not Push The Button:

If Paul's credences are such that, conditional upon his having the decision procedure which he does in fact have, his credence that he is a psychopath is sufficiently high that pushing the button would not maximize expected utility, Paul would not push the button. If, say, causal decision theorists of all sorts were likely to be psychopaths, Paul would not push the button.

(2.) Would Push The Button:

If Paul's credences are such that, conditional upon his having the decision procedure which he does in fact have, his credence that he is a psychopath is sufficiently low that pushing the button would maximize expected utility, Paul would push the button. If, say, the majority of causal decision theorists were likely to be psychopaths, but self-confident causal decision theorists were not likely to be psychopaths, Paul would push the button.

Self-confident causal decision theory seems to do better than standard causal decision theory either way.¹⁰

10 Problem Cases for Evidential Decision Theory

Evidential decision theory mandates that an agent act so as to maximize the expected goodness of the world given his act. Yet this decision procedure sometimes prescribes acts which reliably make the world worse.

A Cigarette Before Dying:

The Predictor (whose unnatural insights have driven him quite mad) has imprisoned a group of decision theorists. The Predictor tells the imprisoned decision theorists that he has predicted what they would do if presented with *Newcomb's Problem*, and that in the morning he will execute all the predicted two-boxers and release all the predicted one-boxers. The decision theorists are all highly convinced of the Predictor's

¹⁰A third, more problematic case is also possible. If Paul's credences are such that, conditional upon his having a self-confident decision function which prescribes pushing the button he is likely a psychopath, but conditional upon his having a self-confident decision function which prescribes not pushing the button he is likely not a psychopath, then both acts fail to stably maximize causal expected utility. For a fuller treatment of such problematic cases, see the technical appendix.

reliability, and highly uncertain about whether they are one-boxers or two-boxers. (They each find it overwhelmingly plausible that they are either standard causal decision theorists or standard evidential decision theorists, and find each of those two possibilities comparably plausible.) In a peculiar gesture, the Predictor provides each decision theorist with a cigarette. The Predictor further notes that if any of the decision theorists throw their cigarette away, he is willing to sell them another for one dollar.¹¹ All of the decision theorists would rather smoke a cigarette than not, and would much rather live than die.

How will standard causal decision theorists reason in this case? The standard causal decision theorists cannot affect the Predictor's prediction, and thus smoking the cigarette dominates not smoking the cigarette. Standard causal decision theorists will therefore smoke the cigarette. How will standard evidential decision theorists reason in this case? They know that standard causal decision theorists will smoke the cigarette, therefore they know that throwing their cigarette away will prove that they are not standard causal decision theorists, and thus be good evidence that they are going to survive the upcoming slaughter of the two-boxers. Throwing the cigarette away does not cause any good effect, but it is *very* good news. The expected utility conditional upon throwing the cigarette away thus exceeds the expected utility conditional upon smoking the cigarette, and thus the standard evidential decision theorists will throw their cigarettes to the ground. They will then gladly pay a dollar for a new cigarette, and ultimately survive the Predictor's madness.

At first glance *A Cigarette Before Dying* looks exactly like *Newcomb's Problem*. And there are indeed substantial similarities. In both cases standard causal decision theorists do and standard evidential decision theorists do not do the act which dominance reasoning demands, and in both cases standard evidential decision theorists wind up doing better than standard causal decision theorists. But in *Newcomb's Problem* the evidential connection between performing the dominated act and being in the desirable underlying state is much closer. In *A Cigarette Before Dying*, the Predictor isn't responding to what the decision theorists do with the cigarette, but rather to what they would do in *Newcomb's Problem*. As a result, were a standard evidential decision theorist to behave uncharacteristically in *A Cigarette Before Dying* things would work out even better for him.

Consider a new sort of decision theorist, the Z-decision theorist, who is exactly like an evidential decision theorist in all cases other than *A Cigarette Before Dying*,

¹¹One may think of throwing the cigarette away as betting a dollar on the false. One can easily modify *A Cigarette Before Dying* so that it is the act of betting a dollar on the false rather than the act of throwing away a cigarette which allows evidential decision theorists to learn that they are not causal decision theorists. This dubious prescription provides a new argument against evidential decision theory—it seems wrong for a decision theory to license betting on the false.

but who will choose to smoke the cigarette in *A Cigarette Before Dying*. The Z-decision theorist will also survive the Predictor's madness (Z-decision theorists are one-boxers) and will also enjoy a cigarette. But the Z-decision theorists will get their cigarette for free. It seems that it is better to be an Z-decision theorist than an evidential decision theorist. Z-decision theorists and evidential decision theorists generally do the same, but in one case Z-decision theorists save a buck.

Z-decision theory is a lot like evidential decision theory, but in one case Z-decision theory prescribes the act that evidential decision theory would prescribe if it were taking an agent's decision-theoretic state for granted. The decision theory which always prescribes the act that evidential decision theory would prescribe if it were taking an agent's decision-theoretic state for granted is self-confident evidential decision theory. Self-confident evidential decision theory is the logical extension of Z-decision theory. A self-confident causal evidential theorist will smoke the cigarette.

Now that we've got a grip on *A Cigarette Before Dying*, let's take a look at the *Medical Newcomb Problem*.

The Medical Newcomb Problem:

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?

The *Medical Newcomb Problem* presents a stranger situation than its prose suggests. Susan thinks that an underlying condition causes both smoking and cancer. Her credences about this condition are crucial. Suppose, for example, that Susan thinks that the underlying condition is one which aliens reliably detect, and that these aliens then abduct those with the condition, force them to smoke cigarettes, and then expose them to radiation. There would then be a correlation between smoking and cancer, but one without direct causation between smoking and cancer. But the threat of alien abductions has no bearing on whether or not Susan should smoke. The choice to smoke neither causes an alien abduction nor is it evidence that an alien abduction will occur. If the correlation between smoking and cancer were due to alien abductions, then both causal decision theory and evidential decision theory would prescribe that Susan smoke. In order to drive causal decision theory and evidential theory apart, the correlation between smoking and cancer must be one which bears on Susan's decision whether or not to smoke.

What sort of correlation can there be between Susan's decision whether or not to smoke and whether or not she has this underlying, cancer-causing condition?

Susan's act is the product of what she thinks, what she wants, what she can do, and how she chooses from among the things she can do on the basis of what she thinks and what she wants. The correlation between smoking and the underlying condition must therefore be a correlation between combinations of thoughts, desires, abilities, and ways of choosing actions which result in smoking. But the correlation is not (or at least need not be) with what Susan thinks. What Susan thinks is stipulated, and Susan can be certain of what she thinks yet still think that she does not have the underlying condition and still think that if she were to smoke that would be strong evidence that she has the underlying condition. The correlation is not (or at least need not be) with what Susan wants. What Susan wants is stipulated, and Susan can be certain of what she wants yet still think that she does not have the underlying condition and still think that if she were to smoke that would be strong evidence that she has the underlying condition. The correlation is not (or at least need not be) with what Susan can do. Susan's options are stipulated, and Susan can be certain of her options yet still think that she does not have the underlying condition and still think that if she were to smoke that would be strong evidence that she has the underlying condition. The correlation between smoking and having the underlying condition must therefore be a correlation between ways of choosing actions and having the underlying condition.

If the correlation between smoking and having the underlying condition is a correlation between ways of choosing actions and having the underlying condition, then evidential decision theory may well recommend that Susan not smoke, for then the expected utility of not smoking may well exceed the expected utility of smoking. Not smoking will have no causal influence over whether or not Susan has the underlying, cancer-causing condition, and smoking is preferable to not smoking whether or not Susan has the underlying, cancer-causing condition, yet evidential decision theory will prescribe that Susan not smoke. This is hardly surprising, however. In *Newcomb's Problem*, just as in the *Medical Newcomb Problem*, evidential decision theory advises performing a strictly dominated act. One may well think that it is irrational to perform a strictly dominated act, but the *Medical Newcomb Problem* contributes nothing to that case.¹²

The only novel problem the *Medical Newcomb Problem* presents is that the evidential connection between the underlying condition and smoking may be looser, thus potentially making not smoking a needless sacrifice. Suppose, for example,

¹²One might wonder why so many people are convinced that the *Medical Newcomb Problem* presents a special problem for evidential decision theory. I strongly suspect that such people are simply confused. In *Newcomb's Problem* it is clear that one-boxers will generally do better than two-boxers. In the *Medical Newcomb Problem* it is less clear (although equally true) that non-smokers will generally do better than smokers. I believe that people mistakenly believe that smokers and non-smokers will foreseeably have comparable cancer rates. But this is not so. Were there no expected advantage to not smoking, evidential decision theory would not prescribe not smoking.

that the underlying, cancer-causing condition is caused by the disposition to two-box in *Newcomb's Problem*. Then the situation in the *Medical Newcomb Problem* is quite similar to the situation in *A Cigarette Before Dying*. If one performs a small, wasteful act then one gets evidence that one's decision-theoretic state is an advantageous one. But since self-confident evidential decision theorists take their decision-theoretic state for granted, they would not perform the small, wasteful act.

It thus remains to be seen whether or not a self-confident evidential decision theorist would smoke. The *Medical Newcomb Problem* is underdescribed, so a determinate analysis will require that Susan's credences be specified more fully. Hold everything about Susan's decision-theoretic state fixed except whether or not she smokes in the *Medical Newcomb Problem*. If one of these two versions of Susan expects herself to do better than the other version of Susan expects herself to do, then self-confident evidential decision theory will prescribe the act performed by the version of Susan with the more favorable expectations.

(1.) Would Not Smoke:

If Susan's credences are such that the expected utility of not smoking conditional upon her being a non-smoker exceeds the expected utility of smoking conditional upon her being a smoker, then Susan would not smoke.

(2.) Would Smoke:

If Susan's credences are such that the expected utility of smoking conditional upon her being a smoker exceeds the expected utility of not smoking conditional upon her being a non-smoker, then Susan would smoke.

Self-confident evidential decision theory seems to do at least as well as standard evidential decision theory either way.¹³

11 Conclusion

Causal decision theory and evidential decision theory are each subject to counterexample. Causal decision theory prescribes a foreseeably foolish course of action in Andy Egans *Psychopath Button* and evidential decision theory prescribes a foreseeably foolish course of action in Allan Gibbard and William Harpers *Medical Newcomb Problem*. These counterexamples rely on exploiting an agent's lack of self-knowledge. But self-confident decision theorists cannot be so exploited, even when they lack self-knowledge. Self-confident causal decision theory and self-confident

¹³No problematic third possibility arises for self-confident evidential decision theory.

evidential decision theory are not subject to counterexample as their standard analogues are. The conflict between causal decision theory and evidential decision theory should thus be superseded by a conflict between self-confident causal decision theory and self-confident evidential decision theory.

Many of the arguments against causal decision theory also apply to self-confident causal decision theory and many of the arguments against evidential decision theory also apply to self-confident evidential decision theory. It remains to be seen if any decision theory will be fully satisfying. You can't always get what you want. But if you try sometimes you just might find you get what you need.

Technical Appendix

1 Formal Setup

A deliberate agent chooses his actions on the basis of what he thinks, what he wants, and what he can do. Put formally, the acts of a deliberate agent are determined by four factors: the agent's credence function, the agent's utility function, the set of acts available to that agent, and the agent's decision function—a function from the agent's credence function, utility function, and set of available acts to an element of the set of available acts.

The agent's credence function ranges over possible worlds, assigning to each world a number which represents the agent's degree of belief that it is actual. The numbers assigned must be non-negative and sum to 1. The agent's credence function is some element of the set of all possible credence functions, some $c_i \in \mathcal{C}$.

The agent's utility function ranges over possible worlds, assigning to each world a number which represents how much the agent would want it to be actual. These values are defined on an interval scale, unique only up to positive affine transformation. The agent's utility function is some element of the set of all possible utility functions, some $u_i \in \mathcal{U}$.

The agent's set of available acts is a set of equivalence classes of possible worlds, with those equivalence classes carved up according to the maximally specific description of how the agent acts in the worlds. The agent can act so as to ensure that the actual world falls into whichever equivalence class he chooses, and has no finer degree of control. The agent is guaranteed to perform exactly one act from among his available acts. The agent's set of available acts is some subset of the set of all possible acts, some $A_i \subseteq \mathcal{A}$.

The agent's decision function represents how the agent chooses his actions on the basis of what he thinks, what he wants, and what he can do. The decision function takes as inputs the agent's credence function, utility function, and set of available acts, and outputs an act from the set of available acts. The agent's decision function is some element of the set of all possible decision functions, some $f_i \in \mathcal{F}$, where $\mathcal{F} := \{f : \mathcal{C} \times \mathcal{U} \times \mathcal{A} \rightarrow \mathcal{A}\}$.

Any act, a_i , of a deliberate agent is determined by a 4-tuple $\langle c_i, u_i, A_i, f_i \rangle$, that determination being made by evaluating the 3-tuple $\langle c_i, u_i, A_i \rangle$ at f_i , so that $f_i(\langle c_i, u_i, A_i \rangle) = a_i \in A_i$.

Any decision theory picks out some set of decision functions.

2 Self-Confident Decision Functions

There is a general procedure for modifying standard decision theories so as to insulate them against prescribing unfortunate courses of action when agents are unaware of their decision-theoretic state. Even if an agent is not sure what his decision-theoretic state is, his decision function can take his decision-theoretic state for granted. I call decision functions that an agent's decision-theoretic state for granted *self-confident*. The act that a self-confident decision function, f , prescribes given any triple of credences, utilities, and available acts will be the same as the act it prescribes given those credences conditional upon the agent having those credences, those utilities, those available acts, and that decision function. That is, $f_i(\langle c_i, u_i, A_i \rangle) = f_i(\langle c_i^*, u_i, A_i \rangle)$, where $Pr_{c_i^*}(p) = Pr_c(p \mid c_i \text{ are my credences, } u_i \text{ are my utilities, } A_i \text{ is my set of available acts, and } f_i \text{ is my decision function})$, for all credences, utilities, sets of available acts, and propositions.

Most self-confident decision functions are crazy and uninteresting. There are innumerable self-confident decision functions that prescribe guzzling hot sauce any time that act is available. Interesting self-confident decision functions are those whose prescriptions mirror the prescriptions of some other decision theory in a limited domain—prescriptions conditional upon the agent's decision function being the agent's decision function. The self-confident analogue of causal decision theory always has $f_i(\langle c_i, u_i, A_i \rangle)$ such that $f_i(\langle c_i^*, u_i, A_i \rangle)$ maximizes causal expected utility, the self-confident analogue of evidential decision theory always has $f_i(\langle c_i, u_i, A_i \rangle)$ such that $f_i(\langle c_i^*, u_i, A_i \rangle)$ maximizes evidential expected utility, and so on. The function's prescriptions conditional upon the agent's decision-theoretic state maximize some sort of expected utility, and those expected utility maximizing conditional prescriptions directly determine what the function's unconditional prescriptions are.

Unfortunately, things can't quite work as I've laid them out. Agents cannot have any possible credences regarding what their credences are. Whatever the domain of possibilities is, the cardinality of the number of possible credences about those possibilities is greater than the cardinality of those possibilities. The easiest way to deal with this cardinality problem is through indefinite extensibility. Let us say that agents can have Lv_0 credences about the world, Lv_1 credences about their Lv_0 credences, Lv_2 credences about their Lv_1 credences, and so on. If at any finite stage the decision-theoretic relevance of an agent's credences drops out, then restricting the credences through which c_i^* is formulated to the finite number of decision-theoretically relevant credences will solve the cardinality problem. Situations of infinite complexity remain problematic, but I set them aside.

3 Self-Confident Causal Decision Theory

Self-confident causal decision theory is a set of decision functions. There are, however, some difficulties in determining which set of decision functions it is. I will say that self-confident causal decision theory *accepts* the decision functions which make it up and that it *rejects* the decision functions which do not make it up. (If your decision function is one that self-confident causal decision theory accepts, then you're a self-confident causal decision theorist. If your decision function is one that self-confident causal decision theory rejects, then you're not a self-confident causal decision theorist.)

Consider an arbitrary situation, S , in which an agent will make a decision. That is, consider an arbitrary ordered triple of credences, utilities, and available acts. Partition all decision functions into equivalence classes according to what actions they prescribe in all situations other than S . That is, divide up decision functions into groups that differ only in what act is performed in situation S . If the set of available acts in situation S is $\{a_1, a_2, \dots, a_n\}$, then each cell of the partition will contain decision functions $\{f_1, f_2, \dots, f_n\}$, one function prescribing each act available in S .

Consider an arbitrary equivalence class. Suppose that, conditional upon any decision function within the equivalence class, the expected utility of some available act a_i is greater than the expected utility of any other available act.¹⁴ In this case, self-confident causal decision theory will reject any function which does not prescribe a_i , as act a_i unproblematically maximizes self-confident causal expected utility.

If all circumstances worked out so neatly, then determining which decision functions constituted self-confident causal decision theory would be unproblematic. Self-confident causal decision theory would reject all the functions which fail to maximize causal expected utility in any circumstance and accept all the functions which always maximize causal expected utility. Unfortunately, not all circumstances work out so neatly. Different acts can maximize expected utility given different decision functions within a cell of the partition, and in such cases it is quite unclear which decision functions (if any) self-confident causal decision theory would reject.

Most starkly, it can be the case that all decision functions prescribe acts which, conditional upon those functions, do not maximize expected utility. In the simplest of these cases there are only two acts available, and if you have a decision function that makes you do a_1 , you'd have been better off if you'd done a_2 instead, and if you have a decision function that makes you do a_2 , you'd have been better off if you'd done a_1 instead. For example,

Can't Go Right:

¹⁴Note that these conditional expected utility calculations should be made having also conditioned on situation S .

James may choose to receive the contents of one of two opaque boxes. If the Predictor predicted that James would take the box on the left, there will be \$100 in the box on the left and \$500 in the box on the right. If the Predictor predicted that James would take the box on the right, there will be \$100 in the box on the right and \$500 in the box on the left.¹⁵

No matter what James does, he will fail to maximize self-confident causal expected utility. Both taking the box on the left and taking the box on the right would fail to maximize self-confident causal expected utility. Thus if self-confident causal decision theory rejects all decision functions which sometimes fail to maximize self-confident causal expected utility, self-confident causal decision theory must reject all decision functions. If self-confident causal decision theory is understood strictly, then there can be no such thing as self-confident causal decision theory.

There is, however, no good reason to be so strict. It is unreasonable for self-causal decision theory to reject decision functions for choosing the box on the left or for choosing the box on the right; at the very least the symmetrical character of the case above should render either act acceptable. Consider the following situation,

Can't Go Wrong:

James may choose to receive the contents of one of two opaque boxes, one of which is sure to contain money. If the Predictor predicted that James would take the box on the left, there will be \$100 in the box on the left and nothing in the box on the right. If the Predictor predicted that James would take the box on the right, there will be \$100 in the box on the right and nothing in the box on the left.

In this case, both taking the box on the left and taking the box on the right succeed at maximizing self-confident causal expected utility. This is not, however, because of a tie. If James takes the box on the left, then taking the box on the left will be the only act which maximizes self-confident causal expected utility, and if James takes the box on the right, then taking the box on the right will be the only act which maximizes self-confident causal expected utility. Under any plausible understanding of self-confident causal decision theory, either act is permitted. But there is no meaningful difference between the first case and the second. However he acts in either case James is going to get \$100.

To my mind, the most plausible requirement for self-confident causal decision theorists is a version of ratificationism developed (for another purpose) by Johan

¹⁵This case is a variant of the famous Death in Damascus case.

Gustafsson¹⁶:

Let an act a_i be *generally ratifiable*₀ if and only if there is no alternative act a_j such that for every act a_k , the expected utility of a_j exceeds the expected utility of a_i given the supposition that a_k is performed.

Let an act a_i be *generally ratifiable* _{$n+1$} if and only if there is no generally ratifiable _{n} alternative act a_j such that for every generally ratifiable _{n} act a_k , the expected utility of a_j exceeds the expected utility of a_i given the supposition that a_k is performed.

Let an act a_i be *generally ratifiable*_{*} if and only if for every natural number n , a_i is generally ratifiable _{n} .

Self-confident causal decision theory will reject a decision-function f_i if and only if any of the following three conditions are met:

- (1.) That decision-function is not self-confident.
- (2.) That decision function prescribes some act a_i which is not generally ratifiable_{*}.
- (3.) That decision function prescribes some generally ratifiable_{*} act a_i and there is some other generally ratifiable_{*} act a_j such that the expected utility of a_j exceeds the expected utility of a_i .

Situations in which different acts maximize expected utility given different decision functions are complicated, and alternative characterizations of self-confident causal decision theory may be worthwhile.

4 Self-Confident Evidential Decision Theory

The formal characterization of self-confident evidential decision theory works out more neatly than that of self-confident causal decision theory. Again, consider an arbitrary situation, S , in which an agent will make a decision. That is, consider an arbitrary ordered triple of credences, utilities, and available acts. Partition all decision functions into equivalence classes according to what actions they prescribe in all situations other than S . That is, divide up decision functions into groups that differ only in what act is performed in situation S . If the set of available acts

¹⁶Standard ratificationism is no help. When all decision functions prescribe acts which, conditional upon those functions, do not maximize expected utility, there are no ratifiable acts. Thus if self-confident causal decision theory required decision functions to only prescribe ratifiable acts, it would reject all decision functions.

in situation S is $\{a_1, a_2, \dots, a_n\}$, then each cell of the partition will contain decision functions $\{f_1, f_2, \dots, f_n\}$, one function prescribing each act in S .

Consider an arbitrary equivalence class. Consider the expected utilities of each act conditional upon the decision functions which prescribe that act. Self-confident causal decision theory will reject a decision-function f_i if and only if any of the following two conditions are met:

- (1.) That decision-function is not self-confident.
- (2.) That decision function prescribes some act a_i and there is some alternative act a_j such that the expected utility of a_j exceeds the expected utility of a_i .

Duty and Knowledge¹⁷

1 Motif

Deontological ethics needs a decision theory. But it has a hard time getting one.

2 Motivating Deontological Decision Theory

It's fairly easy to see what a deontological ethic demands of an omniscient agent. If killing an innocent person were absolutely forbidden, then the ethic would require that the agent never choose to kill an innocent person no matter what. If promise-breaking were absolutely forbidden, then the ethic would require that the agent never choose to break a promise no matter what. And so on. An omniscient agent can follow a deontological ethic just by abstaining from any action which is categorically forbidden.¹⁸

It's far less easy to see how an agent acting under uncertainty would follow a deontological ethic. If killing an innocent person were absolutely forbidden, under what conditions can an agent who is unsure whether or not some person is innocent kill that person? If promise-breaking were absolutely forbidden, under what conditions can an agent who is unsure whether or not some act would break a promise perform that act? What should a deontologist do when he is unsure whether or not some possible act violates his deontological scruples? If deontological ethics has no decision theory—if there's no way for deontology to tell an agent how he should act in light of the limited information he has available rather than in light of the fullness of all truth—then deontology leaves its adherents rather in the lurch. Without a decision theory, deontology is a theory of rightness and wrongness which is of no use to agents trying to decide what to do. And conscientious deontologists need some way to figure out what to do.

If deontological ethics doesn't have a decision theory, then deontological ethics is on rather worse footing than consequentialist ethics, which does have a decision theory. Expected utility theory provides a well-established way to take an agent's state of uncertainty (represented as a credence function) and facts about value

¹⁷Delivered as a talk at ANU's philosophical society in August 2014 and published in *Philosophical Perspectives* in December 2014.

¹⁸There are some niceties to be worked out regarding moral dilemmas, but the basic cases are unproblematic.

(represented as a utility function) and thereby specify which course of action an agent should undertake.¹⁹

Consider a trivial case. Suppose that you're given the opportunity to bet your life for a dollar on the outcome of a fair coin toss. If it landed Heads you die, if it landed Tails you get a buck. Should you take the bet? One answer (sensible in its own, odd way) is that it depends: if the coin landed Heads then you shouldn't take the bet, and if the coin landed Tails then you should take the bet. Everyone can agree that taking the bet results either in death or in a dollar, and everyone can agree that death is a bad outcome and that a dollar is a good outcome. But if no other sort of answer to that question were available, then there'd be no way for an uncertain agent to determine whether or not he should take that bet. You'd be in bad shape indeed if you had no way to evaluate that bet without unavailable information.

Deontologists have an easy time characterizing the objective ought, but a much more difficult time characterizing the subjective ought. And the subjective ought is where the real action is—the subjective ought is strictly more informative than the objective ought. One can derive the objective ought from the subjective ought. Suppose that an agent must decide whether or not to perform some act ϕ . That agent objectively ought to ϕ if and only if an agent who was certain of the actual state of the world subjectively ought to ϕ .²⁰ The limiting case of the subjective ought is the objective ought. But while the subjective ought can accommodate any degree of uncertainty (down to none at all), the objective ought cannot accommodate any uncertainty whatsoever.

One note in passing: The import and irreducibility of the subjective ought shows how misguided the project of “consequentializing” ostensibly non-consequentialist ethical theories is.²¹ Since the “consequentializing” occurs in the limited domain of the objective ought, it amounts to nothing more than constructing a total order over worlds relative to which a maximizing choice function can be defined. But the subjective ought must be a choice function over probability functions over worlds, and it's not even clear what properties the “consequentialized” theory's subjective ought should have.²² By considering the objective ought but not the subjective

¹⁹There are, of course, debates in decision theory about how expected utilities should be calculated. But the vast majority of cases are universally agreed upon.

²⁰Note that an agent who was certain of the actual state of the world might have inaccurate credences. For example, if in the actual world no agent was certain of anything, then this agent (who would obviously have to inhabit some other possible world) would be certain that no agent was certain of anything.

²¹“Consequentializers” attempt to show that ostensibly non-consequentialist ethical theories are (or at least can be thought of as being) consequentialist ethical theories. For more see Campbell Brown's “Consequentialize This”.

²²A *prima facie* plausible standard for a “consequentialized” subjective ought is that it ought to make the theory representable as maximizing expected utility. But there are two reasons why

ought, “consequentializers” create the impression that consequentialist and deontological ethical theories are more similar than they actually are.

3 The Problem

How should a deontological ethic influence decision-making in conditions of uncertainty? There are, broadly speaking, three norms through which deontological scruples could be applied—and none of them are any good.²³

Permissive Norm: The ethic forbids ϕ -ing if and only if the agent is certain that ϕ -ing would violate a deontological scruple.

Restrictive Norm: The ethic forbids ϕ -ing if and only if the agent is not certain that ϕ -ing would not violate a deontological scruple.

Intermediate Norm: The ethic sometimes forbids and sometimes permits ϕ -ing when the agent is not certain whether or not ϕ -ing would violate a deontological scruple.

The permissive norm is no good because it’s *hugely* too permissive. Certainty is rare. If the prohibition against killing innocents applies only when an agent is certain that someone is innocent, then that agent may kill just about anyone without violating the prohibition. If the prohibition against lying applies only when an agent is certain that an assertion is false, then that agent may assert just about anything without violating the prohibition. The permissive norm pretty much never prohibits any action. That’s bad.

The restrictive norm is no good because it’s *hugely* too restrictive. Certainty is rare. If the prohibition against killing innocents applies whenever an agent is not certain that someone is not innocent, then that agent is basically prohibited from killing. If the prohibition against lying applies whenever an agent is not certain an assertion is true, then that agent is basically prohibited from making assertions. The restrictive norm pretty much never permits any action. That’s bad.

We are therefore left with the intermediate norm. And while the intermediate norm isn’t as terrible as either the permissive norm or the restrictive norm, it still

this standard is unavailable to contemporary ethicists. First, John von Neumann and Oskar Morgenstern showed the conditions under which an agent is representable as an expected utility maximizer in 1947, so there’s not much work for contemporary ethicists to do on that front. Secondly, deontological ethical theories are wildly incompatible with the prerequisites of expected utility maximization. More on this later.

²³Here I assume that deontological agents should choose from among their permissible acts so as to maximize expected utility, and that the role of the deontological norm is to determine which acts are permissible. I will consider an alternative approach to deontological norms in section 6.

looks pretty bad. It faces two major difficulties, which I will dub the *Arbitrariness Problem* and the *Conjunction Problem*.

The *Arbitrariness Problem* is that there seems to be no good reason for some particular intermediate level of confidence to separate permissibility from impermissibility. Why should it be permissible to kill someone when one has a credence of .6372 that they are innocent but impermissible to kill someone when one has a credence of .6373 that they are innocent? Whether there is a constant threshold or whether there are different thresholds for different acts (or even different thresholds for different circumstances) there doesn't seem to be a good reason for *any* threshold. The thresholds seem arbitrary, and arbitrary thresholds seem implausible.²⁴

The *Conjunction Problem* is that thresholds interact poorly with multiplicities of acts. Jackson and Smith press this problem forcefully:

Suppose that there are two skiers, X and Y, whose probability of innocence is individually just below the threshold...and whose paths threaten separate groups of people. On the threshold view, you ought to shoot X if the number saved is large enough because the probability of innocence is below the level at which the absolute prohibition kicks in; ditto for Y.

Since X's and Y's probabilities of innocence are each individually just below the threshold, the probability that at least one of them is innocent is over the threshold.²⁵ Is it, then, permitted to kill each of them with two distinct actions (say, with two bullets) but forbidden to kill both of them with a single action (say, with a grenade)? This seems ridiculous. While there are various possible norms for multiplicities of acts, no extant candidate seems particularly palatable.²⁶

4 Knowledge Norms

Let's take it for granted that there's no good way for a deontological ethic to influence decision-making in conditions of uncertainty.²⁷ Let's also take it for granted that deontological ethics are in serious trouble if they cannot accommodate a decision theory. Even granting all of this it does not follow that deontological ethics are in serious trouble, but merely that to avoid serious trouble deontological ethics will

²⁴In contrast, consequentialists can offer a perfectly sensible explanation for the cutoffs they endorse—the cutoffs are where the risks and the rewards equal out.

²⁵Assuming that their probabilities of innocence are not maximally correlated, which is easily stipulated.

²⁶See Ron Aboudi, Adi Borer, and David Enoch's "Deontology, Individualism, and Uncertainty: A Reply to Jackson and Smith."

²⁷I will expand on this case in the next section.

need a decision theory which does not deal in uncertainties. Perhaps deontology will need its decision theory to be formulated in terms of some other epistemological ideology. I contend that a knowledge-based decision theory will fit the bill.

How should a deontological ethic influence decision-making in conditions of limited knowledge? Again there are, broadly speaking, three norms through which deontological scruples could be applied—but these look *much* better.

Permissive Norm: The ethic forbids ϕ -ing if and only if the agent knows that ϕ -ing would violate a deontological scruple.

Restrictive Norm: The ethic forbids ϕ -ing if and only if the agent doesn't know that ϕ -ing would not violate a deontological scruple.

Intermediate Norm: The ethic sometimes forbids and sometimes permits ϕ -ing when the agent does not know whether or not ϕ -ing would violate a deontological scruple.

Both the permissive norm and the restrictive norm look *much* better when framed in terms of knowledge than when framed in terms of certainty. Agents both know a great deal and fail to know a great deal. Agents often know that some act would violate a deontological scruple and often do not know that some act would violate a deontological scruple, so the permissive norm is non-trivial. Agents often know that some act wouldn't violate a deontological scruple and often do not know that some act wouldn't violate a deontological scruple, so the restrictive norm is non-trivial. Both the permissive norm and the restrictive norm are non-arbitrary, and since both the permissive norm and the restrictive norm are viable, there is no need to worry about the intermediate norm's problems with arbitrariness. But while both the permissive norm and the restrictive norm evade the arbitrariness problem, each can face the conjunction problem, so let's turn to that.

The conjunction problem arises for the permissive norm if it's possible for an agent to have more knowledge about a conjunction of acts than he does about either of the acts conjoined.²⁸

²⁸For these purposes I assume that licitness respects agglomeration. That is, I assume that if ϕ -ing is licit and ψ -ing is licit then ϕ -ing and ψ -ing is licit. I thus leave aside situations in which an agent is explicitly permitted to do either of two acts but forbidden to do both. Such situations can arise, but they just don't have anything to do with the conjunction problem. Strictly speaking, the formalism of decision theory only allows an agent to perform a single act in any decision situation, thus making acts fully fine-grained. Nonetheless, thinking in a coarser grain is often helpful. This coarse grain is merely a heuristic; in situations in which licitness does not respect agglomeration the heuristic should be dropped.

Permissive Problem Case: The agent doesn't know that ϕ -ing would violate a deontological scruple and doesn't know that ψ -ing would violate a deontological scruple but does know that both ϕ -ing and ψ -ing would violate a deontological scruple.

It's obviously possible to have more knowledge about a conjunction of acts than about either of the acts conjoined. There's nothing questionable about saying that an agent doesn't know that buying the first ticket in a two ticket lottery would lead to his winning and that the agent doesn't know that buying the second ticket in the two ticket lottery would lead to his winning but that the agent does know that buying both the first ticket and the second ticket in the two ticket lottery would lead to his winning.

How would this sort of problem case play out? Let's return to the skier case. Suppose that the agent knows that exactly one of the two skiers is innocent, but that he doesn't know which. The agent would be permitted to kill either skier, but forbidden to kill both. This is a bit of an unpalatable result, but not completely horrible.²⁹

The conjunction problem arises for the restrictive norm if it's possible for an agent to have less knowledge about a conjunction of acts than he does about each of the acts conjoined.

Restrictive Problem Case: The agent knows that ϕ -ing wouldn't violate a deontological scruple and knows that ψ -ing wouldn't violate a deontological scruple but doesn't know that both ϕ -ing and ψ -ing wouldn't violate a deontological scruple.

It's debatable whether or not it's possible to have less knowledge about a conjunction of acts than about each of the acts conjoined. Can an agent know that ϕ -ing is licit and know that ψ -ing is licit without knowing that both ϕ -ing and ψ -ing are licit? There's some measure of controversy over whether knowledge is closed under entailment,³⁰ and such a strong closure principle would obviously rule this sort of case out. A weaker principle can still help with this case. If agents are always in a position to know the entailments of what they know, this would suffice to ensure that whenever an agent knows that ϕ -ing is licit and knows that ψ -ing is licit, he is at least in a position to know that both ϕ -ing and ψ -ing are licit. Timothy

²⁹Since an agent might perform a sequence of acts which he knew to violate a deontological scruple even though no act within it was known to violate a deontological scruple, it is likely necessary to forbid agents from performing acts which will cause them to know that they violated a deontological scruple.

³⁰See David Lewis' *Elusive Knowledge* and Jaakko Hintikka's *Knowledge and Belief* for more.

Williamson’s knowledge-first epistemology employs just such a framework.³¹ One might, however, employ a notion of knowledge according to which it’s possible for an agent to know p , know q , but still be quite agnostic about $p \wedge q$.

How would this sort of problem case play out? Let’s return to the skier case again. Suppose that the agent knows that each of the two skiers is not innocent, but that he doesn’t know that both skiers are not innocent. The agent would be permitted to kill either skier, but forbidden to kill both. This is a bit of an unpalatable result, but not completely horrible.

Although both the permissive norm and the restrictive norm are viable, I prefer the restrictive norm. Williamson’s knowledge-first epistemology is the best knowledge-based epistemology in which knowledge does real work.³² Under the auspices of knowledge-first epistemology the restrictive norm evades the conjunction problem while the permissive norm does not. More importantly, the restrictive norm seems more appropriate for deontological ethics than the permissive norm. Deontologists take the idea of killing an innocent person *very* seriously. A deontologist would not knowingly take an innocent life *pereat mundus*—though the world perish.³³ It therefore seems inappropriately cavalier for a deontologist to think that he is permitted to kill someone because he doesn’t know whether or not that person is innocent.³⁴ If the deontological subjective ought is formulated through the permissive norm, then an agent who obeyed the subjective ought might still violate the objective ought. But if the deontological subjective ought is formulated through the restrictive norm, then an agent who obeyed the subjective ought would be guaranteed to obey the objective ought.

5 Battling Bayesianisms: Certainty-Based vs. Knowledge-Based

Knowledge norms are best thought of as part of a larger, knowledge-first epistemology. That framework is therefore worth exploring in a bit more detail. And it’s

³¹Although Williamson allows that knowledge is extendible under competent deduction, this principle is not necessary for a solution to the conjunction problem. Even if a conjunction were not known despite its conjuncts being known, the knowledge of the conjuncts would guarantee that the conjunction had an evidential probability (that is, a probability conditional on what the agent knows) of 1. And probability 1 makes for a non-arbitrary threshold.

³²In contrast to, say, David Lewis’ knowledge-based epistemology, which leaves the real work to be done by credences.

³³*Fiat justitia, et pereat mundus*—let there be justice, though the world perish—was the personal motto of the Holy Roman Emperor Ferdinand I. This motto is famously (and approvingly) quoted by Immanuel Kant in his *Perpetual Peace*.

³⁴I note that the restrictive norm is favored by those who support a knowledge norm for assertion. Surely the norm for killing someone should not be more lax than the norm for assertion.

best explored through contrast.

The standard framework for decision theory employs Bayesian probabilities. Probabilities provide enough detail about an agent's epistemic state to do the fine-grained work that decision theory requires.³⁵ While there are various probabilistic epistemologies in the offing, they share a common formal structure.

Traditional Bayesians impose two kinds of coherence requirements on rational credences, a synchronic constraint which applies at each time and a diachronic constraint which applies over different times. The synchronic requirement is **probabilistic coherence**: that the agent's credences at any time should form a probability function. That is, an agent's degrees of belief must (at a minimum) conform to the following axioms of the probability calculus:

- (1) $\Pr(p) \geq 0$, for any proposition p .
- (2) $\Pr(t) = 1$, for any tautology t .
- (3) $\Pr(p \vee q) = \Pr(p) + \Pr(q)$, for any inconsistent propositions p and q .

The diachronic requirement is **conditionalization**. This constraint relies on a notion of conditional probability, which we can define given the axioms above. The probability of p conditional on q is equal to the probability of p and q divided by the probability of q (assuming that $\Pr(q) > 0$).

$$\Pr(p \mid q) =_{df} \frac{\Pr(p \wedge q)}{\Pr(q)}.$$

Conditionalization mandates that when an agent has evidence E at time a and evidence $E+$ at time b (where $E+$ implies E), then the probability of any proposition p at b should equal the conditional probability of p given $E+$ at a . That is, $\Pr_b(p) = \Pr_a(p \mid E+)$.

Putting this all together, the basic story of Bayesian epistemology goes something like this. Rational agents assign non-negative credence to various possible worlds in such a way that the sum of those credences is 1. When rational agents gain evidence which is inconsistent with some of those possible worlds, they adjust their credences in those worlds to 0, and proportionally increase their credences in the remaining worlds so that the sum of their credences returns to 1.

Conditionalization presupposes a notion of evidence. But what, exactly, is evidence? Different epistemologies offer different answers. Just as earlier there were certainty-based norms and knowledge-based norms, there is the certainty-based conception of evidence and knowledge-based conception of evidence.

³⁵Outright beliefs and deduction are insufficient. For more, see Gilbert Harman's *Change in View*.

The more traditional notion of evidence (at least among decision theorists) is certainty-based.³⁶ The basic idea is that an agent's evidence is what that agent has complete, utter, bet-your-life-for-a-nickel confidence in. An agent takes his evidence for granted. The traditional thought is that agents are certain of their phenomenal states. (This view is often called the phenomenal conception of evidence.) Suppose an agent is looking at a box and that the box looks red. The agent might be unsure whether the box really is red or whether the box merely looks red due to some tricky lighting, but (the traditional thought goes) the agent can be certain that the box looks red. Agents are supposed to have privileged, infallible access to their phenomenal states.

The certainty-based conception of evidence faces a serious problem. There's a strong case to be made that agents (or at least agents like us) are never justified in being certain of anything non-trivial. We may well have privileged access to our own phenomenal states, but it's quite clear that that access is not infallible. Suppose that an agent takes a quick glance at a painting with 137 black dots.³⁷ That agent's phenomenal state will have all 137 black dots represented (at least supposing that the agent's eyesight is decent). Nonetheless, it seems completely reasonable for the agent to be unsure of exactly how many black dots there are on the painting after the quick glance. The quick glance of the 137 black dots contained all the information the agent would have needed to be sure that the painting contained 137 black dots, but an agent need not have access to all the information contained in his phenomenal state.

Think about the staggeringly huge number of possible phenomenal states. For each phenomenal state there are myriad other phenomenal states just barely different from it. There are then other phenomenal states just barely different from those (though perhaps more than barely different from the original phenomenal state). And so on. Agents like us are not justified in being certain of which phenomenal state we're in—we're not good enough at distinguishing different phenomenal states.

Now one might think that while we are not justified in being certain of our exact phenomenal state, we can be justified in being certain that our phenomenal state falls into some range of phenomenal states. But I don't think this is plausible. There does not seem to be a sharp cutoff for how mistaken agents like us can be when trying to identify our phenomenal states. Consider the continuum of visible colors. It's obvious that I can mistake one shade of red for another shade of red. But I have no reason to think that it's impossible for me to mistake a shade of red for a shade of red / orange, or to mistake a shade of red for a shade of orange, or to mistake a shade of red for a shade of green. I think I'm unlikely to mistake a

³⁶For example, here's CI Lewis in "The Given Element in Empirical Knowledge": "Nothing is probable unless something is certain."

³⁷I owe the structure of this case to Roderick Chisholm's problem of the speckled hen.

shade of red for a shade of green, but there's no last shade I might mistakenly pick out—I might mistake any shade for any other shade at all.³⁸

Let's take for granted that agents like us are never warranted in being certain of anything non-trivial. The certainty-based conception of evidence would thus leave agents like us insuperably without evidence, and that would be bad. Epistemology should not decree that agents like us never have any reason to change our minds about anything. We therefore need an alternative notion of evidence.

Enter the knowledge-based conception of evidence. Knowledge is *much* easier to come by than certainty.³⁹ A key principle of Williamson's knowledge-first epistemology is that $E = K$, that an agent's evidence is the totality of what that agent knows. While knowledge can play the role of evidence, since knowledge is quite different from certainty, knowledge-based epistemology will be quite different from certainty-based epistemology.

A few differences between the certainty-based conception of evidence and the knowledge-based conception of evidence are notable. Although the role played by an agent's sense certainties in the former case is played by what the agent knows in the latter case, the role changes quite a bit depending on its occupant. Complete certainty that p would warrant betting one's life that p in order to get a nickel—one would think of such a bet as a free nickel. But the knowledge that p does not warrant betting one's life that p in order to get a nickel. Knowing p is not the loftiest epistemic state an agent can have regarding p . An agent could know that p while failing to know that he knows that p , or know that p and know that he knows that p while failing to know that he knows that he knows that p , and so on. The loftiest epistemic state an agent can have regarding a proposition would be all possible iterations of knowing⁴⁰, but the arguments against justified certainties would show that agents never have all possible iterations of knowledge regarding any non-trivial proposition. Knowing that p is easier than having complete certainty that p , which is why the knowledge-based framework is not forced to be too permissive, too restrictive, or arbitrary. But just as knowing that p is easier than having complete certainty that p , knowing that p is not as helpful as having complete certainty that p . It was a hope of certainty-based decision theory that it would provide agents with rules they would always know how to follow. But since agents are not always in a position to know what they know, knowledge-based decision theory cannot possibly provide agents with rules they would always know how to follow. The proponent of knowledge-based decision theory therefore needs some way to distinguish blameless violations of the theory's norms from blameworthy violations

³⁸For more, see Timothy Williamson's *Knowledge and Its Limits*.

³⁹Barring extreme skeptical worries, of course.

⁴⁰An epistemically ideal agent would also have absence-luminosity: if an ideal agent didn't know that p he would know to all possible iterations that he didn't know that p .

of the theory's norms.⁴¹ For example, we might censure an agent for violating the norms when in a position to know that he was violating them but not when he had every reason to think that he was following them.⁴² These complications impose serious costs for a knowledge-based decision theory.⁴³ But deontology is simply stuck with those costs. Deontology needs a decision theory, and there is no adequate alternative.

6 An Inadequate Alternative

Altering the formal structure of standard decision theory in order to accommodate deontological scruples leaves us with a decision theory which is either too permissive, too restrictive, or arbitrary. A knowledge-first decision theory solves these problems, but it does cause a few others. And while I do think that a knowledge-first decision theory is the best option for deontological ethics, it's worth making sure that we can't solve all our problems with a bit of creativity. Perhaps instead of altering the formal structure of standard decision theory we could just impose some new constraints within the formal structure of standard decision theory. Perhaps an agent with deontological scruples is just an agent whose utility function has certain specific features. If killing an innocent person were absolutely forbidden, then the agent would prefer any world in which he does not kill an innocent person to

⁴¹And similarly between praiseworthy conformity to the theory's norms and non-praiseworthy conformity to the theory's norms.

⁴²There are other plausible methods for apportioning decision-theoretic praise and blame. Consider the following passage from John Hawthorne and Amia Srinivasan: "Suppose two tennis players are each in situations where they know they can't reach the ball in time. One player—call him Andi—gives up. Another, call him Raphael, chases the ball though (unsurprisingly) fails to get to it in time. Andi might deride Raphael as a pathetic figure, someone who gives chase while knowing the chase is futile. But we can see that because the absence of knowledge isn't luminous, Andi risks turning himself into a player who fails to chase the ball when he might, in fact, reach it in time. For cases will arise sooner or later where Raphael *believes* he knows he won't reach the ball, but in fact doesn't know this. In such cases, thanks to the habit he has inculcated in himself, Raphael will sometimes end up reaching the ball. On the other hand, Andi doesn't chase the ball both in cases where he knows that he won't reach it, *and* in those cases in which he falsely takes himself to know that he won't reach it. In this way, Andi—in failing to chase the ball even when he knows he won't reach it—inculcates in himself an undesirable habit. . . If we focus on the local scenario, Andi's actions seem unassailable; after all, he gives up when he knows it is futile. By contrast, Raphael's behaviour seems a little lamentable; after all, he keeps chasing when he knows he can't reach the ball. But if we shift to a more global outlook, we can see Andi's course of action as problematic and blameworthy, Raphael's as noble and praiseworthy. It is far from clear that one of these outlooks has any special authority over the other."

⁴³Although such complications could also aid rivals to knowledge-based decision theory, I believe that knowledge-based decision theory makes better use of complications than its rivals do.

any world in which he does kill an innocent person.⁴⁴ If promise-breaking were absolutely forbidden, then the agent would prefer any world in which he does not break a promise to any world in which he does break a promise. And so on.

Unfortunately, this attempt to situate deontological scruples within standard decision theory causes its own insuperable problems.⁴⁵ The less grievous problem is that the agent's deontological scruples involve essentially first-personal facts. The agent prefers any world in which *he* does not kill an innocent person to any world in which *he* does kill an innocent person. Thus two different deontologists would have to have different utility functions—if one of them must be a murderer, then each must prefer that it be the other. A mere general abomination of killing innocents will not do, for such a general abomination would favor murdering the innocent child of a serial killer if by doing so one could demoralize the serial killer enough to stop him from killing, and this is precisely the sort of thing that deontological scruples forbid. The inclusion of first-personal facts is not an unreasonable feature for a decision theory. Any theory meant to allow for self-interest must include first-personal facts, and standard decision theory is meant to allow for self-interest.⁴⁶ I note, however, that the pertinence of first-personal facts for deontological ethics would then represent a decisive departure from consequentialist ethics for which first-personal facts must be irrelevant. One more nail in the coffin for “consequentializing” deontological ethical theories.

The more serious problem is that restricting utility functions to accommodate deontological scruples makes them ridiculous. Suppose that an agent prefers any world in which he does not kill an innocent person to any world in which he does kill an innocent person. That agent would therefore prefer for others to commit the worst horrors in the history of the world, spreading unprecedented misery throughout the Earth before ultimately extinguishing life itself than for him to kill an innocent person. Such an agent will, for all practical purposes, be an absolute pacifist. To his mind, the risk of killing an innocent person is too staggering to take given

⁴⁴Provided that the worlds in which he does not kill an innocent person are worlds in which he does not violate any other deontological scruple.

⁴⁵It does not, however, cause the problem ascribed to it by Frank Jackson and Michael Smith. They claim that such an approach violates agglomeration, such that according to this approach it can be the case that an agent ought to ϕ and ought to ψ despite it not being the case that the agent ought to ϕ -and- ψ . But this is not so. It's easy to see that any agent following standard decision theory will obey agglomeration (at least finite agglomeration). We can apply the standard Kratzer semantics for “ought”, taking the agent's available acts as our modal base and taking the expected utilities of those available acts as our ordering source. The Kratzer semantics validates finite agglomeration, so since we can fit this approach into the Kratzer semantics this approach obeys finite agglomeration. Of course, Jackson and Smith might be thinking of some non-standard notion of “ought”, but they give no indication that they have a Kratzer-violating notion of “ought” in mind.

⁴⁶More complicated worries about personal identity and self-locating belief might require even more formalism, but the framework of centered worlds can provide it.

any plausible level of uncertainty. If promise-breaking were so forbidden, the agent would have to consider the mere act of making a promise—thus risking the most horrific possibility that he wind up breaking that promise—to be unconscionably reckless. This restriction on utility functions is too restrictive, and validates the scruples that deontologists ought to accept only by also validating other scruples that deontologists ought to reject.

Restrictions to utility functions warp deontological sensibilities. It's not the case that deontologists think that killing innocents always makes things worse and is therefore wrong. Deontologists think that killing innocents is wrong whatever its consequences may be. Restricting utility functions is just not a viable option for deontologists.

7 Some Formal Analysis

I've argued that the subjective ought of a deontological ethic should be formulated through the restrictive knowledge-based norm. This norm forbids ϕ -ing if and only if the agent does not know that ϕ -ing would not violate a deontological scruple. This subjective ought is not any sort of expected utility maximization, but is instead expected utility maximization subject to a constraint. (We saw in the last section that unconstrained expected utility maximization does not provide a viable context for a deontological decision theory.) Since I'm advocating that deontologists eschew expected utility maximization (albeit for something closely related to it) I owe some explanation for why doing so is sensible. After all, there is a grand tradition in decision theory which considers representability as an expected utility maximizer to be the very hallmark of practical rationality.

Representability is cheap. I can represent my socks as expected utility maximizers if I want to. The significance of representation theorems is that they represent someone as an expected utility maximizer *in some specific, non-trivial way*. Standardly, an agent's preferences may or may not be representable as maximizing expected utility relative to some credence function and some utility function. There are a number of important representation theorems that have been proven, but they all share their basic structure with the first such theorem—the Von Neumann–Morgenstern utility theorem. The Von Neumann–Morgenstern utility theorem is also the easiest to understand, so I'll focus my analysis on it.

Suppose that an agent is faced with some lotteries. For these purposes, we'll define a lottery as being constituted by some set of mutually exclusive possibilities, with each possibility assigned a probability such that the sum of the probabilities is 1. One lottery might be a 50% chance of eating a delicious, ordinary cupcake and a 50% chance of eating a delicious, cyanide-laced cupcake, and another lottery might be a 50% chance of throwing a delicious, ordinary cupcake into the trash

and a 50% chance of throwing a delicious, cyanide-laced cupcake into the trash. As you can see, one can think of an action as providing an agent with a lottery which corresponds to the credences he would have if he performed the action.

Von Neumann and Morgenstern proved that an agent's preferences over lotteries are representable as maximizing expected utility relative to some credence function and some utility function if and only if that agent's preferences satisfy four constraints:⁴⁷

- (1.) Totality: For any lotteries L and M , exactly one of the following holds: (1) $L \succ M$, (2) $M \succ L$, or (3) $L = M$.
- (2.) Transitivity: If $L \succeq M$ and $M \succeq N$, then $L \succeq N$.
- (3.) Independence: If $L \succ M$, then for any N and any probability $p \in (0, 1]$, $pL + (1 - p)N \succ pM + (1 - p)N$.
- (4.) Continuity: If $L \succeq M$ and $M \succeq N$, then there is a probability $p \in [0, 1]$ such that $pL + (1 - p)N = M$.

The deontological decision theory I've proposed satisfies the first three constraints, violating only the fourth. Let's see why.

Totality requires that an agent either strictly prefer one lottery to another or be indifferent between them. Disallowing acts which an agent does not know to be deontologically licit doesn't interfere with this at all.

Transitivity requires that an agent's preferences over lotteries correspond to some sort of overall ranking. Again, disallowing acts which an agent does not know to be deontologically licit doesn't interfere with this at all.

Independence requires an agent's preference of one lottery over another to hold regardless of the independent possibility of some other outcome. Once more, disallowing acts which an agent does not know to be deontologically licit doesn't interfere with this at all. If both of the lotteries being composed with a third alternative are known to be licit, whatever preference the agent has between the two lotteries will transfer over to the higher-order lotteries of which they are a part. And if one of the two lotteries is known to be licit while the other is not, that preference will transfer over as well.^{48,49}

⁴⁷I express an agent having a strict preference for lottery L over lottery M with $L \succ M$, and an agent being indifferent between lottery L and lottery M with $L = M$.

⁴⁸Note that I continue to set aside moral dilemmas, which in this case means that I only consider cases in which an agent has at least one available act which he knows to be deontologically licit. There are a variety of ways to handle moral dilemmas, the most obvious of which are (1.) for the agent to minimize the probability of violating a deontological scruple, and (2.) for the agent to maximize expected utility without regard to the risks of violating his deontological scruples. Approach (1.) obeys Independence, whereas approach (2.) does not.

⁴⁹The permissive knowledge norm would violate independence. The agent could risk nonzero

Continuity requires that there be some mixture of the best of three options and the worst of three options which will be of equal value to the middle of three options. But the restrictive knowledge norm disallows any act with nonzero chance of violating a deontological scruple, and that disallowance cannot be diluted. The risk of violating a deontological scruple poisons a possibility, rendering it unfit for choice.

Continuity is, however, the least important of the Von Neumann / Morgenstern axioms. It has no real theoretical basis. The axiom of continuity requires either that utility functions be bounded or that there be some restriction on the composition of lotteries, but the only motivation for either requirement is that they make the math easier.⁵⁰ I am not troubled that the theory I have proposed violates continuity.⁵¹

On the whole, the formal structure of knowledge-first deontological decision theory is extremely close to that of traditional decision theory.

8 Conclusion

Why should you accept this package of deontological ethics and knowledge-first epistemology? Well, frankly, I don't think you should. I like neither deontological ethics nor knowledge-first epistemology. It's stable to like neither. Similarly, it's stable to like just knowledge-first epistemology and favor some non-deontological ethic. But a predilection for deontological ethics should strongly incline one to knowledge-first epistemology. A major reason to like standard Bayesian epistemology is that it's good for decision theory.⁵² But in fact standard Bayesian epistemology is only good for consequentialist decision theory. The way to make deontological decision theory work is to adopt a knowledge-first epistemology. So if you like deontological ethics,

probabilities of violating deontological scruples so long as the agent maintained a nonzero probability of not violating deontological scruples. The agent would not kill an innocent person to prevent the world from being destroyed, but could nonetheless prefer a 50% chance of killing an innocent person and a 50% chance of nothing happening to a 50% chance of the world being destroyed and a 50% chance of nothing happening. Knowledge-first decision theories do accord knowledge a special significance, but this result is unpalatable nonetheless. One more reason to prefer the restrictive knowledge norm to the permissive knowledge norm.

⁵⁰If utilities are unbounded and there are no restrictions on the composition of lotteries then the St. Petersburg lottery will lead to violations of continuity.

⁵¹A permissive knowledge norm would also violate continuity, but for different reasons. The agent would not kill an innocent person to prevent the world from being destroyed, but could nonetheless prefer any chance less than 1 of killing an innocent person to a corresponding chance of the world being destroyed.

⁵²Alan Hájek puts things well in "Arguments For—Or Against—Probabilism?": "[O]nce we get to probabilism, it provides us with many fruits. Above all, it forms the basis of a unified theory of decision and confirmation—it combines seamlessly with utility theory to provide a fully general theory of rational action, and it illuminates or even resolves various hoary paradoxes in confirmation theory. I consider that to be the best argument for probabilism."

that's a major motivation to go knowledge-first in epistemology. And if you don't like knowledge-first epistemology, that's a major motivation to reject deontological ethics.

The Problems of Transformative Experience

1 Introduction

Decision theory is uncontroversially well suited for artificially simple decision problems. If you want to know how to bet when a randomly-selected marble is drawn from a randomly-selected urn, you need look no further than decision theory. The suitability of decision theory for more complex, realistic decision problems is more controversial.

Laurie Paul has recently argued that transformative experiences pose a problem for decision theory. According to Paul, agents facing transformative experiences do not possess the states required for decision theory to formulate its prescriptions. Agents facing transformative experiences are impoverished relative to their decision problems, and decision theory doesn't know what to do with impoverished agents.

Richard Pettigrew takes Paul's challenge seriously. He grants that decision theory (in its traditional state) cannot handle decision problems involving transformative experiences. To deal with the problems posed by transformative experiences, Pettigrew proposes two alterations to decision theory. The first alteration is meant to handle the problem posed by epistemically transformative experiences, and the second alteration is meant to handle the problem posed by personally transformative experiences.

I argue that Pettigrew's proposed alterations are untenable. Pettigrew's novel decision theory faces both formal and philosophical problems.⁵³ It is doubtful that Pettigrew can formulate the sort of decision theory he wants, and further doubtful that he should want such a decision theory in the first place. The problems posed by transformative experience thus remain unmitigated.

2 Epistemically Transformative Experience

An epistemically transformative experience is one which provides information about the phenomenal character of an experience that can only be had by having that experience. Examples of such experiences abound. You can't know what it's like to see red without seeing red. You can't know what it's like to eat Vegemite without

⁵³I will explicitly distinguish the formal problems from the philosophical ones. The more serious problems are the philosophical ones, as the formal problems can be palliated by formal means. Readers who are uninterested in formal issues should feel free to skip the formal sections of this paper.

eating Vegemite. You can't know what it's like to have a child without having a child. And so on.

According to Paul, an agent who must choose whether or not to have an epistemically transformative experience faces a difficult problem. Since the agent hasn't had the epistemically transformative experience yet, the agent has no idea what he'd be getting himself into by choosing to have the experience. And since the agent has no idea how to evaluate the epistemically transformative experience, the agent has a difficult time making a rationally acceptable decision.

Pettigrew frames the problem posed by epistemically transformative experiences in a plausible way—as a problem concerning an agent's utilities. Pettigrew notes that in order for him to make a decision in accordance with standard decision-theoretic norms he would need utilities defined over the possible outcomes of his decision. When deciding whether or not to undergo an epistemically transformative experience, he would need utilities defined over experiences he cannot fathom.⁵⁴ According to Pettigrew, “The problem is that... I do not have access to these utilities.”⁵⁵ Pettigrew takes such access to be required for decision theory to function. “[P]resupposition: In order for an agent to follow this procedure, she must have access to her probabilities and to her utilities.”⁵⁶

Pettigrew's proposed solution is to redescribe decision problems involving epistemically transformative experiences with a finer grain. Instead of having an agent evaluate possible states of the world, Pettigrew suggests that an agent should consider conjunctions of possible states of the world and possible utilities for those states of the world.

[I]nstead of taking the set of possible states of the world to be $\mathcal{S}$, we instead take it to be $\mathcal{S}'$, a fine-graining of $\mathcal{S}$, where the fine-grained possible states of the world specify not only how the world is, but also what my utility function is over the outcomes relative to $\mathcal{A}$ and $\mathcal{S}$. Thus, let $u_1, \dots, u_n$ be the utility functions I might take have over the outcomes relative to $\mathcal{A}$ and $\mathcal{S}$. Then we reformulate the decision problem by taking the set of possible states of the world to be:

$$\mathcal{S}' = \{S \wedge \text{My utility function is } u_i : S \in \mathcal{S} \text{ and } i = 1, \dots, n\}$$

And we define the utility of the outcomes relative to $\mathcal{A}$ and $\mathcal{S}'$ as follows:
for each $i = 1, \dots, n$,

⁵⁴To make a decision about whether or not to see red, he would need utilities defined over the various ways of seeing red and the various ways of not seeing red. To make a decision about whether or not to eat Vegemite, he would need utilities defined over the various ways of eating Vegemite and the various ways of not eating Vegemite. To make a decision about whether or not to have a child, he would need utilities defined over the various ways of having a child and the various ways of not having a child.

⁵⁵Pettigrew (2015)

⁵⁶Pettigrew (Forthcoming)

$$u(A \wedge S \wedge \text{My utility function is } u_i) = u_i(A \wedge S)$$

Doing this solves the original problem of epistemically inaccessible utilities.⁵⁷

In short, expected values for utilities do the work that would have been done directly by utilities, had those utilities not been problematically inaccessible.

It is not quite clear what Pettigrew has in mind when he talks about the epistemic inaccessibility of an agent's utilities, and (relatedly) it is not quite clear how he intends his proposed solution to work. I see two possible interpretations, one epistemological and one ontological. Given the epistemological interpretation, agents *have* utilities concerning the experiences which they cannot yet fathom; the inaccessibility consists in the agent's being *uncertain* of what those utilities are. The expected value that Pettigrew defines would then be an expected value of the agent's current utility. Given the ontological interpretation, agents *do not have* utilities concerning the experiences which they cannot yet fathom; the inaccessibility consists in the agent's *lack* of such utilities. The expected value that Pettigrew defines would then be an expected value of the utility the agent would have after undergoing the epistemically transformative experience. But under either interpretation Pettigrew's proposal is fatally flawed.

2.1 Epistemological Interpretation: Uncertain Utilities

If the problem posed by epistemically transformative experiences is that agents are uncertain about the relevant utilities, then Pettigrew's proposed solution faces three problems: the first problem concerns the uselessness of utilities given this proposal, the second concerns a misapprehension about the role of clear understanding in decision theory, and the third concerns the difficulties inherent in averaging across different utility functions.

2.1.1 Philosophical Problem: Useless Utilities

Under this interpretation, agents have all the utilities needed for normal decision-theoretic analysis. Agents who have never seen red have utilities for seeing red. Agents who have never eaten Vegemite have utilities for eating Vegemite. Agents who have never had children have utilities for having children. And so on. This would mean that Pettigrew allows that agents making decisions concerning epistemically transformative experiences have everything needed for a normal decision-theoretic approach⁵⁸, but nonetheless wants to depart from the normal decision-theoretic approach. But that's unreasonable. There's no good reason to mess

⁵⁷Pettigrew (2015)

⁵⁸A normal decision-theoretic analysis requires only that an agent has credences and utilities. More on this shortly.

around with the mechanisms of decision theory if they can function properly. As the platitude goes—if it ain’t broke, don’t fix it! Given this interpretation of epistemically transformative experiences, Pettigrew’s departure from decision-theoretic orthodoxy is starkly unmotivated.

Moreover, Pettigrew’s proposed heterodoxy makes decision theory worse. Note first that the utilities that agents have regarding transformative experiences do no work; they cannot influence any decision whatsoever.⁵⁹ Two agents with the same credences but with different utilities regarding transformative experiences will be guaranteed to always make the same decisions. After all, Pettigrew proposes having expectations of utilities serve the role traditionally served by utilities, and those expectations are determined entirely by credences. Pettigrew would make utilities regarding transformative experiences utterly useless, and useless utilities are implausible.

It seems very strange for utilities to do crucial work in nearly every decision problem, and then drop into irrelevance when a transformative experience comes up.⁶⁰ It would make more sense for there to be a uniform treatment—for utilities to *always* be irrelevant, and for the work thought to be done with utilities to be done by credences about them. But it seems wrong for what an agent thinks to make what that agent wants be totally irrelevant. Decision theory rightly uses both credences and utilities; we can’t make do with credences alone.

2.1.2 Philosophical Problem: The Role of Clear Understanding

Pettigrew’s proposal is, I think, motivated by two convictions:

- (1.) Only states which an agent clearly understands can feature in decision theory.
- (2.) An agent’s credal state is one he clearly understands.

Both of these convictions are false.

⁵⁹Pettigrew’s proposal may therefore be incoherent if it’s necessary for utilities to serve a functional role.

⁶⁰Note that the scope of the problem posed by epistemically transformative experiences will depend on how decision problems are individuated. If the possibilities under consideration are individuated coarsely, then the problems posed by epistemically transformative experiences will be relatively rare. If the possibilities under consideration are individuated finely, then the problems posed by epistemically transformative experiences will be ubiquitous. Suppose you’re facing a very pedestrian decision: whether or not to have a sip of water. If you have just one utility for having a sip and just one utility for not having a sip, then the problems posed by epistemically transformative experiences won’t arise. On the other hand, if you have utilities for all the worlds in which you have a sip and for all the worlds in which you don’t have a sip, then the problems posed by epistemically transformative experiences will arise, as you have epistemically transformative experiences in many of those worlds.

Decision-theoretic norms are formulated in terms of credences and utilities. The central norm of decision theory—expected utility maximization—is formulated in terms of credences and utilities. Maybe some agent clearly understands what his credences and utilities are, and maybe that agent doesn't. But it doesn't make a difference as far as the norms of decision theory go. If the agent acts in a way which maximizes expected utility then he acts as decision theory says he should, and if the agent acts in a way which does not maximize expected utility then he acts as decision theory says he should not. The very notion of clear understanding is not part of decision theory; it's just credences and utilities.

There may well be agents who have beliefs and desires, but who are not cognitively sophisticated enough to have either beliefs or desires about their own beliefs and desires. This poses no problem whatsoever for decision theory. The agent's credences and utilities suffice to determine what the agent ought to do. Decision theory can easily determine whether or not the agent takes the rational means for his ends. This agent won't understand decision theory, but that's no matter. Professional decision theorists have to understand decision theory in order to be professional decision theorists, but agents do not have to understand decision theory in order to be agents.

This is not to say that it's not wonderful for an agent to clearly understand his own credences and utilities. Reliably accurate credences are, in general, quite good to have. And clear understanding of one's own credences and utilities is plausibly particularly advantageous. When coupled with clear understanding of how decision theory works, an agent with a clear understanding of his own credences and utilities would be in a position to calculate what the norms of decision theory dictate. Such calculations are entirely wonderful. They are not, however, the sole purpose of decision theory. Pettigrew writes,

Now, if I am to use decision theory to help me deliberate about what to do, or to justify my decision once it is made, I must at least have access to my credences over the relevant dependency hypotheses and to the utilities I assign to the outcomes of the available acts. After all, I must use them to calculate the expected utilities of the available acts.⁶¹

Pettigrew is right that perfectly reliable calculation requires perfectly reliably certainty about one's own credences and utilities. But Pettigrew is wrong that such perfectly reliable calculations are essential to decision theory. There's nothing all that wrong with merely moderately reliable calculations, for example. Or decision theory might provide an agent with a model of ideally rational decision-making which appropriately influences his less than ideally rational-decision making, and that without any calculation whatsoever. Or, as the example of agents who lack

⁶¹Pettigrew (2015)

credences about their own credences and utilities shows, decision theory is perfectly able to evaluate agents who cannot evaluate themselves.

This may well seem strange. One might think this vision of decision theory insufficiently “internalist”. Now admittedly, an agent’s credences and utilities plausibly supervene on that agent’s intrinsic properties, so there’s still a measure of internalism there. But more than that just isn’t part of the framework of decision theory. One can, of course, stipulate that all agents clearly understand their own credences and utilities. And much fruitful theorizing has been done on the basis of such stipulations.⁶² But such stipulations are not essential to decision theory.

An agent’s credences and utilities cannot be assumed to be clearly understood by that agent. An agent may be certain or uncertain of his own credences and utilities, just as an agent may be certain or uncertain of most anything else. Moreover, given that Pettigrew’s model is one in which agents can be uncertain of their own utilities, there seems to be no reason to assume that agents cannot be uncertain of their own credences. But if agents can be uncertain about their own credences, then Pettigrew’s proposal is transparently misguided. Pettigrew suggests that agents look to their credences about their utilities instead of their utilities because agents are uncertain of their utilities. But agents can be just as uncertain about their credences about their utilities as they are about their utilities!

Why look to an agent’s credences about his utilities? Why not go another level up, and look to an agent’s credences about his credences about his utilities? Or another level up, and look to an agent’s credences about his credences about his credences about his utilities? And so on. There’s no reason to privilege the 2nd level of credences over the 12th or to privilege the 12th level of credences to the 2nd. The level that should be privileged is the 0th level—the utilities themselves, and not any sort of credences about them.

One might think that there’s an obvious alternative level to privilege: the top one.⁶³ Suppose that the agent has utilities, credences about his utilities, credences about his credences about his utilities, and that’s it. In that case, why not apply Pettigrew’s proposal at the level of credences about credences about utilities? Wouldn’t that more satisfactorily privilege what the agent clearly understands?

No. In fact, entirely to the contrary. If an agent’s credences go up only a finite number of levels, then the credences at the top level are the furthest thing from clearly understood—they are not understood at all. The agent has no credences about his top level of credences. That means that the agent is epistemically blind to them. The agent doesn’t understand them clearly, the agent doesn’t understand

⁶²For example, Robert Aumann’s celebrated agreement theorem makes this stipulation along with several others.

⁶³Note that there is no guarantee that there will be a top level of credences. The levels might just keep going up and up.

them murkily, the agent doesn't understand them at all.⁶⁴

Decision theory need not be predicated on what an agent clearly understands, and an agent need not clearly understand his own credences.

2.1.3 Formal Problem: Averaging Across Different Utility Functions

Pettigrew's proposal involves taking an average across different utility functions. There are different possible utilities an agent may assign to some experience and different possible utility functions doing the assigning. But, standardly, there can be no such thing as an average across different utility functions. An average across different utility functions is not well-defined mathematically. And this is not just an artifact of the mathematics; the math isn't well-defined because the thing isn't well defined.

There are limits to what sorts of mathematical operations can be meaningfully performed. Suppose someone wondered, "What's the height of a tree that's as tall as the temperature of a hot summer day?" One could reply "A hot summer day is around 95 degrees Fahrenheit, so the tree would be around 95 inches tall. It wouldn't be a particularly tall tree, then, but more likely a sapling." But this reply is obvious nonsense. The whole idea of directly equating some height with some temperature is absurd. Note that the numbers don't help even though the same numbers appear when measuring heights and temperatures. There's no equivalence to be had between 95 degrees Fahrenheit and 95 inches. Degrees Fahrenheit and inches are just very different things; the common '95' doesn't help. Of course, the selection of those units was arbitrary. One could have equally well taken degrees Celsius and centimeters and gotten a different bogus answer. But the basic problem is not the plurality of units. Then the problem would be that there were too many viable answers to the question. There are, however, no viable answers to the question. The question itself is deeply misguided.

Numbers (here, I'm think specifically of the natural numbers) can be put to a great many uses. Different uses capitalize on different aspects of the numbers' numerical structure. We can use the numbers 1, 2, and 3 to describe how many coins Alice, Bob, and Carol have in their pockets. In this case, it makes sense to use the additive structure of those numbers. Just as $1 + 2 = 3$, it makes perfect sense to say that Alice and Bob have as many coins in their pockets as Carol does in her pockets. We can also use the numbers 1, 2, and 3 to describe the order in which Alice, Bob, and Carol completed a race. But in this case it makes no sense to use the additive structure of those numbers. There's no sense in which Alice's race result and Bob's race result, taken together, are equivalent to Carol's race

⁶⁴The only way to represent an agent who clearly understands some level of his own credences is for that agent's hierarchy of credences to go up infinitely, but for the credences in that hierarchy to become trivial after that level.

result. But the ordinal structure of the numbers still applies. Just as $1 < 2 < 3$, it makes perfect sense to say that Alice finished before Bob, who in turn finished before Carol.

If you think clearly about what a utility function is, it becomes obvious why averages across different utility functions are not meaningful. The mathematical structure of utility functions is quite modest. Utilities have the structure only of an interval scale. Utilities have an order, and comparisons of differences between utilities can be made, but that's it. Utilities can express that an agent prefers A to B, and utilities can express that an agent prefers A to B by twice as much as he prefers C to D. But utilities can express nothing more. It is, for example, not meaningful to say that A has twice as much utility as B. The reason is that utilities are meant to capture preferences and nothing more. The preferences defined over simple outcomes are purely ordinal. Preferences over mixtures⁶⁵ of simple outcomes get only a little more structure. An agent prefers A to B by as much as he prefers B to C just in case the agent is indifferent between the certainty of B and a 50 / 50 mixture of A and C. And so on. The numerical representation of an interval scale is—as Pettigrew correctly notes—unique only up to positive affine transformation (that is, adding or subtracting any number to all utilities and multiplying all utilities by any positive number leaves those utilities unchanged). All the properties that matter in a utility function will be preserved by any positive affine transformation. Suppose an agent has three options, A, B, and C. Giving them utilities of 1, 2, 3 is the same as giving them utilities of 5, 6, 7 (just add 5), is the same as giving them utilities of 10, 20, 30 (multiply by 10), is the same as giving them utilities of 15, 25, 35 (multiply by 10 and then add 5), is the same as giving them utilities of 50, 60, 70 (add 5 and then multiply by 10). All the properties a utility function has are preserved in those assignments. Those utility functions all just represent preferring C to B to A, and preferring C to B and B to A by the same amount. Those utility functions are the same utility function, just as $(2 + 2)$ and $(1 + 3)$ are the same number. The representations are different, but it's the same thing being represented.

On a mathematical level it makes sense that you can't average across different utility functions. Suppose you're trying to take a straight average of two utility functions, giving them each equal weight. What numerical representations of the two functions should you take? If you average 1, 2, 3 and 3, 2, 1, you get 2, 2, 2. But if you average 100, 200, 300 and 3, 2, 1, you get 51.5, 101, 150.5. The latter representation of the first utility function gives it massively more of an effect than the former representation does. But any representation-dependent operation is nonsense.

On a philosophical level it makes sense that you can't average across different

⁶⁵Such as lotteries.

utility functions. Utilities don't measure absolute amounts of desirability. It's not as though you can average liking X a lot with liking X a little and thereby get liking X moderately. Two agents, one of whom likes every possibility and the other of whom dislikes every possibility, can easily have the same utility function for those possibilities. Utilities show only relative desirabilities, and standings in two different hierarchies of relative desirability can't be averaged together.

There is thus reason to be worried—very worried—about the coherence of a proposal that requires averaging across different utility functions. But the situation is not totally hopeless. If the averages were taken across the utility functions of different people, then the situation would be totally hopeless. But since all relevant utility functions are held by a single agent, some special maneuvering might help.⁶⁶

The crucial move would be to identify particular utilities in different utility functions with one another. Then, and only then, would an average taken across different utility functions be meaningful. Suppose we can stipulate that value X in one utility function is equivalent to value Y in some other utility function. I'll call such a pair of values an *anchor*. In order for comparisons between two utility functions to be meaningful, there must be two distinct anchors between them. With two distinct anchors, both the relative positions and the relative scales of two utility functions could be determined, thus capturing their full structure.⁶⁷ So long as an agent had enough utilities which were not problematized by epistemically transformative experiences, those unproblematic utilities could act as anchors.

2.2 Ontological Interpretation: Utility Gaps

If the problem posed by epistemically transformative experiences is that agents lack the relevant utilities and only gain them (if ever) after having having an epistemically transformative experience, then Pettigrew's proposed solution faces two problems: the first problem concerns the irrelevance of the future, and the second concerns the difficulties inherent in averaging across different utility functions.

2.2.1 Philosophical Problem: The Irrelevant Future

Decision theory mandates that an agent act so as to maximize expected utility. That expectation is calculated with the agent's current credences and the agent's

⁶⁶Speaking for myself, I do not think that special maneuvering helps. I think that averages across different utility functions are insuperably ill-defined. But I appreciate that others may think that the structure of utilities is more flexible than I do. Thus I feel I should explore the uses that such flexibility could be put to.

⁶⁷Given only one anchor it would be impossible to capture the relative scales of the two utility functions. And the scale is always the root of the problem; merely additive differences always wash out on their own.

current utilities. The agent's credences may well change in the future⁶⁸ and the agent's utilities may well change in the future. But those future credences and utilities are of no direct relevance to an agent's decision. In just the same way that an agent can have credences and utilities about what the weather will be like tomorrow, an agent can have credences and utilities about what his credences and utilities will be like tomorrow. There's nothing special about those future credences and utilities—they're decision-theoretically inert. An agent can prefer that his future preferences be satisfied; he should then act so as to satisfy them. An agent can prefer that his future preferences not be satisfied; he should then act so as to not satisfy them. When an agent has occurrent preferences the question of what preferences the agent should aim to satisfy is completely trivial—the agent should aim to satisfy those occurrent preferences. And when an agent does not have occurrent preferences then it is totally unclear what the aim of that agent's actions should be (or even if there's a fact about how agent should act in the first place). It is unreasonable to cast around for something to constrain the actions of an agent without preferences, and there's no reason for that agent's expectations about his future utilities to ground such constraints.

Pettigrew writes,

[P]resupposition: In the description of this procedure, we have not specified which of her utilities the agent should weight by her probabilities in order to obtain her expected utility for a given action—is it her utilities at the time she makes the decision, or at some other time, perhaps when the consequences of her decision will be making themselves felt?... Since this is not specified in standard expected utility theory, if that account is to be well-defined, there must be a presupposition that my utilities remain unchanged over time.⁶⁹

This is simply false. This putative presupposition is deservedly absent from every textbook on decision theory. Decision theory no more makes the assumption that agents' utilities will remain fixed throughout the course of their lives than it makes the presupposition that agents will live forever. Agents do not even have to be *alive* when the effects of their decisions are felt; they certainly do not have to be *bouletically unchanged* when the effects of their decisions are felt. Decision theory does not depend on future utilities in any regard.

Remember that utilities specify preferences, and that preferences are defined over total states of affairs.⁷⁰ It would therefore be wrong to say that a hedonist

⁶⁸And likely will as the agent gets more evidence.

⁶⁹Pettigrew (Forthcoming)

⁷⁰These total states of affairs need not be maximal. Their extent can vary from context to context.

gives high utility to eating when he is hungry and low utility to eating when he is full, high utility to sleeping when he is tired and low utility to sleeping when he is well-rested, and so on. A hedonist has constant utilities—he always prefers according to his own pleasure. Utilities are not defined over general things like eating and sleeping; they are defined over specific states of affairs. At the finest grain, a hedonist looks at whole worlds and then prefers according to the extent of his pleasure in them. For any particular choice of whether to eat or whether to sleep the hedonist can simply check whether his expected pleasure is better eating or not, sleeping or not. His expected happiness conditional on eating is different when he is hungry than when he is full. His expected happiness conditional on sleeping is different when he is tired than when he is not tired. Preparing for a future change of appetite is not preparing for a future change of utilities.

A utility gap is a far-reaching thing. The only way for an agent not to have a utility for any experience he cannot fathom is for the agent to have no utility for any state of affairs in which he has an experience he cannot fathom. In the limit, the agent would need to have no utility for any world in which he has such an experience. It's not as though the agent can have utilities for the things in a world with which he's acquainted and leave out the utilities for the things in the world with which he isn't acquainted. Utilities don't apply to things in a world, but to worlds themselves. An agent's future utilities after having had an epistemically transformative experience are then necessarily something about which the agent has no preferences. To say that the agent prefers that his future preferences be satisfied even though he has no current preferences is contradictory—asserting that the agent both does and does not have current preferences. The only way to make the case of a utility gap consistent is for the agent to have no utilities regarding his future utilities, to not care about them in any way at all. But in that case the future utilities are obviously devoid of immanent relevance.

2.2.2 Formal Problem: Averaging Across Different Utility Functions (Reprise)

Under this interpretation, averaging across different utility functions poses the same problems as before. In this regard there is no difference between an agent averaging different utilities he thinks he may have and an agent averaging different utilities he thinks he may come to have. Just as before, the only hope for the meaningfulness of such averages is a plurality of anchor utilities. Only through anchor utilities can such averages be mathematically and philosophically meaningful.

3 Personally Transformative Experience

A personally transformative experience is one which alters an agent's values—which things are valued and the extent to which they are valued. In a decision-theoretic framework, transformative experiences are ones which alter an agent's utilities. While I think that Pettigrew is right that personally transformative experiences can pose a problem for decision theory, I think he thinks misunderstands what that problem is. And since Pettigrew's understanding of the problem is intimately connected with his proposed solution, I'll leave my own thoughts about transformative experiences aside for the time being.

According to Pettigrew, decision theory faces a problem in any situation in which an agent's utilities are not guaranteed to remain fixed. In light of this conviction, it is obvious why Pettigrew takes personally transformative experiences to pose a problem—personally transformative experiences guarantee a change of utility.

Suppose I face a choice between a range of acts; and suppose that one of those acts has a possible outcome that involves a [personally transformative experience]. For instance, suppose I must choose whether or not to become a parent for the first time. Thus, in this decision problem $\mathcal{A} = \{Adopt; Don't\ adopt\}$. . . And let us say, very crudely and ignoring the fact that becoming a parent may be an [epistemically transformative experience] as well as a [personally transformative experience], that I know that, if I become a parent, I will come to value time spent with friends less and time spent with family more. What's more, I know that the outcome *Adopt* involves a lot more time with family and a lot less with friends, while this is reversed in *Don't adopt*. Thus, my current utility for *Adopt* is lower than for *Don't adopt*; and this is true of my future utilities as well in the outcome in which I don't adopt. But, if I do adopt, then my future utilities will be the reverse of my current utilities: I will value the family time entailed by *Adopt* more than the time with friends entailed by *Don't adopt*. How, then, am I to make this choice? This is Paul's second challenge to decision theory: it is ambiguous in cases in which my utilities change over time.⁷¹

Pettigrew's proposed solution is to require that agents have a new sort of utility function. A traditional utility function represents how an agent values complete scenarios (in the limit, how the agent values entire worlds). This new sort of utility function, which Pettigrew dubs a local utility function, instead represents how an agent values possible present time slices. Pettigrew's proposal is to aggregate local utility functions together to make what he calls a global utility function, and then

⁷¹Pettigrew (2015)

to use that global utility function as traditional decision theory uses a traditional utility function.

A natural response is this. First, introduce the notion of a local utility function: my local utility function at a given time is the function that measures how much I value outcomes at that time. Next, let us demand that, as in the previous section, the possible states of the world are fine-grained enough to include a specification of the relevant facts about my utilities: in this case, where my utilities may change over time, that will include a specification of my local utility function for each time during my life in that state of the world. Now, let us fix attention on a particular outcome that results from conjoining an act with a state of the world that is fine-grained in that way: thus, this outcome will include a specification A of the act that is performed and a specification S of how the world is (call the conjunction $A \wedge S$ its worldly component); and it will include a specification of my local utility functions (call this its utility component). Let us simplify and assume that, in this outcome, there are just finitely many moments in my life, $t_1, \dots, t_n$. And let $lu_{t_1}, \dots, lu_{t_n}$ be my local utility functions at those moments. Thus, the outcome in question is:

$$A \wedge S \wedge \bigwedge_{i=1}^n \text{My local utility at } t_i \text{ is } lu_{t_i}$$

Then we might say that the utility function to which I appeal when I make a decision at one of those moments (say t_i) is my global utility function at t_i (which we write u_{t_i}), where the utility that u_{t_i} assigns to the outcome just described in some way aggregates the local utilities assigned to the worldly component of that outcome ($A \wedge S$) by each of the local utility functions $lu_{t_1}, \dots, lu_{t_n}$. The natural means of aggregation in this case is the weighted sum. Thus, my global utility at t_i for the outcome described above is determined by a series of weights $\beta_{t_1}, \dots, \beta_{t_n} \geq 0$. These specify how much each of my local utilities for the worldly portion of that outcome contribute to the global utility for the whole outcome, which is the utility I use when I make a decision or justify it once made. So, for each $j = 1, \dots, n$, the weight β_{t_j} specifies the extent to which I take into account the values I had or have or will have at time t_j . Thus:

$$u_{t_i}(A \wedge S \wedge \bigwedge_{i=1}^n \text{My local utility at } t_i \text{ is } lu_{t_i}) = \sum_{i=1}^n \beta_{t_i} lu_{t_i}(A \wedge S)$$

This solves the formal problem.⁷²

Unfortunately, Pettigrew's proposal is untenable. Whatever malady Pettigrew envisions, the proposed treatment is far worse. Pettigrew's proposal faces three problems: the first problem concerns the irrelevance of the future, the second concerns local utilities themselves, and the third concerns the difficulties inherent in averaging across different utility functions.

3.0.1 Philosophical Problem: The Irrelevant Future (Reprise)

It is simply false that there is any presupposition in decision theory that an agent's utilities will not change over time. It is easy to say what an agent should do in the face of potentially changing utilities—maximize expected utility relative to his current credences and current utilities.

The decision-theoretic primacy of an agent's current credences and current utilities is not some arcane technicality of the formalism—it makes perfect sense. Suppose that you currently value the happiness and wellbeing of your family. Nonetheless, you know that tomorrow you will undergo a horrifying (albeit temporary) change, and will come to instead value the misery and wretchedness of all those you currently hold dear. Note that the only thing that will change is what you value; your psychological reactions will remain the same. Even tomorrow you wouldn't take an iota of pleasure from the suffering of your loved ones. You'd just choose that they suffer, and be utterly horrified that you had so chosen.

What should you do? It would be madness to suggest any sort of compromise between your current utilities and your future utilities. You needn't feel compelled to help your future self get what your future self will want. In fact, since what your future self will want is so diametrically opposed to what you want, you should feel compelled to *prevent* your future self from getting what your future self will want. You should take steps to make sure that you won't be in a position to do any harm when tomorrow takes its strange toll on your utility function. You might consider locking yourself in a cell. If you can get yourself into a medically induced coma, so much the better.⁷³

There is absolutely no reason to give any special status to utilities you may come to have. Concomitantly, there is absolutely no problem posed by utilities changing over time.

3.0.2 Philosophical Problem: Trouble with Local Utilities

Pettigrew's local utility functions do great violence to the notion of the utility function. They are not a little different than regular utility functions. They are

⁷²Pettigrew (2015)

⁷³For more on this sort of decision problem, see Elster (1979).

staggeringly, calamitously different.

Given Pettigrew's proposal, there is literally no way to care about anything that happens after you die. Agents have no local utility functions covering times after their deaths, so those times are utterly irrelevant to the global utility function which is stitched out of local utility functions. It is absurd to have a decision theory in which it is not so much as *possible* to work for the wellbeing of the family one will leave behind, for the good of future generations, or for the grim purpose of posthumous vengeance against one's enemies. A decision theory in which the future cannot matter is a decision theory that isn't worth having.

Pettigrew's local utility functions only value particular time slices. There is therefore no way for local utility functions to value or disvalue the various ways that time slices fit together in a world. Thus Pettigrew would force agents to evaluate each moment in isolation. There would be no way, for example, for one to care about the persistence of one's spouse. Being a newlywed is pretty spectacular; it would therefore be hard to avoid valuing most highly a weird, patchwork world in which one gets married (apparently for the first time) over and over again. The weird, patchwork world doesn't make sense, but there is no way to disvalue that lack of sense.

Local utility functions are not even particularly well-defined. Since local utility functions are operant only over possible present moments, it's tremendously hard for them to make the sorts of comparisons they need to. Suppose you're deciding whether or not to have a child. At any time in which you know you have a child, your local utility for not having a child will be undefined. At any time in which you know you don't have a child, your local utility for having a child will be undefined. Agents can only have local utilities for epistemic possibilities, and that's a big problem. Even if you don't know whether or not you will have a child in the future, your normal-looking futures are ones in which you know whether or not you wound up having a child.

Local utility functions are untenable.⁷⁴ Decision theory cannot countenance them.

3.0.3 Formal Problem: Averaging Across Different Utility Functions (Finale)

The problems of averaging across different utility functions present themselves again, but more starkly. Personally transformative experiences straightforwardly change an agent's utility function, making comparisons across utility functions more difficult. As before, a plurality of anchor utilities is required to make the

⁷⁴The weighting function over the local utilities is also something of a blank check, and it is unclear how deal with different possible life-spans.

comparisons mathematically and philosophically meaningful. The problem is that the existence of such anchor utilities is even more doubtful.

The only hope for getting anchor utilities is a specification of which utilities were changed by the personally transformative experience. Unchanged utilities might then act as anchors, as before. If some personally transformative experience leaves no utility unchanged, then there is simply no hope for a comparison. Note that the specification of which utilities changed and which remained the same would have a practical effect. Changing these utilities and leaving those the same would be very different from changing those utilities and leaving these the same. Suppose an agent has utilities defined over four options: A, B, C, and D. Suppose that the agent's utilities for these options can be given the following numerical representation: 2, 4, 6, and 8. (Remember, the agent's utilities are unique only up to positive affine transformation; other numerical representations for the agent's utilities are also available.) Consider a change of the agent's utilities such that A and B become worse and C and D remain the same. The agent's utilities are now 0, 2, 6, and 8. Now consider a change of the agent's utilities such that C and D become better and A and B remain the same. The agent's utilities are now 2, 4, 8, 10. These two utility functions, 0, 2, 6, 8 and 2, 4, 8, 10, would have different effects on an average. But note that, standardly, 0, 2, 6, 8 and 2, 4, 8, 10 are two numerical representations of *the same utility function*. Utilities are unique only up to positive affine transformation, and the latter utilities are each just 2 larger than the former utilities. This means that Pettigrew needs to invoke more finely-grained utility functions than are employed in traditional decision theory. Averaging across different utility functions always requires some violence against the standard framework of decision theory. This sort of averaging across different utility functions requires more violence than usual.

4 Conclusion

To my mind, the problems of transformative experience are rooted in utility gaps. The most serious problem that an epistemically transformative experience could pose is that agents would lack utilities for experiences they could not fathom. The most serious problem that a personally transformative experience could pose is that agents would lack utilities for being radically different than they are.⁷⁵ Pettigrew's proposals for revising decision theory are untenable, and I am not optimistic that an alternative proposal will fare better. Decision theory is a well thought-out thing.

⁷⁵Note that this is not the problem that Pettigrew takes transformative experiences to pose. It remains to be seen how radical a transformation would need to be in order to induce a utility gap. It need not be the case that any change in utilities would induce a utility gap, and yet some changes do.

If decision theory cannot offer good guidance for agents who lack utilities—and it can't—then there is simply no good guidance in the offing. The problems of transformative experience stand.

Probabilities Cannot Be Rationally Neglected⁷⁶

Nicholas Smith argues that the value that an ideally rational agent places on a gamble need not equal that gamble's expected value.⁷⁷ Instead, Smith argues that outcomes of sufficiently low probability may fail to influence a gamble's valuation—that some probabilities are rationally negligible.

Rationally Negligible Probabilities (RNP): For any lottery featuring in any decision problem faced by any agent, there is an $\epsilon > 0$ such that the agent need not consider outcomes of that lottery of probability less than ϵ in coming to a fully rational decision (Smith 2014).⁷⁸

RNP allows problematic infinite gambles (like the St. Petersburg game or the Pasadena game) to be reduced to unproblematic finite gambles. RNP's contextual sensitivity makes it a flexible principle; the value of ϵ can change from lottery to lottery. Yet despite that flexibility RNP has untenable implications. (And it's a good thing that RNP has untenable implications—if low probabilities can be ignored, then my fine-grained treatment of agents' lack of self-knowledge would be unworkable.)

Smith's employment of RNP is not totally unsystematic. He considers various constraints for the values that ϵ can take, and endorses the weakest of them. This principle is

Weak Consistency: If a single decision problem d involves two lotteries L_1 and L_2 over the same set of outcomes, then if one truncates L_1 at ϵ for purposes of addressing d , one must also truncate L_2 at ϵ for purposes of addressing d .

But a troubling result easily follows from Weak Consistency: RNP must license taking arbitrarily much risk for arbitrarily little reward.

Let us consider a fair coin which will be tossed until it lands Heads.⁷⁹ We will define some lotteries over that set of outcomes, and an agent will choose whichever of them he likes best. One available lottery is Status Quo, which pays off \$0

⁷⁶Published in *Mind* in May 2016.

⁷⁷When such a value exists.

⁷⁸According to Smith, the probabilities of the remaining outcomes should be renormalized. The value of the lottery would then be the expected value according to these renormalized probabilities.

⁷⁹The coin may continue landing Tails forever, but this unusual scenario is of no moment for what follows.

however the coin lands.⁸⁰ Another available lottery is Bad News, which is a bit more complicated.

Bad News: The payoff of the game is $-\$(2^{3n-2})$, where n = the total number of tosses performed.⁸¹

Bad News bears some resemblance to the classic St. Petersburg game. Each gamble guarantees a finite payoff, yet the expected value of each gamble is infinite. But in Bad News, unlike in the St. Petersburg game, the payoffs are negative and the terms make increasingly significant contributions to expected utility.⁸² Also available are lotteries of the form Penny or Doom _{n} , for all $n > 0$. These pay one penny so long as the coin is not flipped exactly n times, and pay $-\$(2^{3n-2})$ if the coin is flipped exactly n times.

Smith must permit an agent to prefer infinitely many versions of Penny or Doom to Status Quo.⁸³ Smith must permit an agent to evaluate infinitely many versions of Penny or Doom as being worth a penny, despite the fact that they have enormous (and growing) negative expected payouts. The proof is simple. RNP mandates that some of the outcomes in Bad News may be ignored.⁸⁴ Weak Consistency mandates that the same outcomes be negligible in all of the lotteries.⁸⁵ Thus outcomes of sufficiently low probability will be negligible in all of the lotteries. Therefore infinitely many versions of Penny or Doom will be such that their negative outcomes may be ignored, thus making it rational to value them at a penny. And a penny is better than nothing.

So lotteries with a very high probability of producing a penny and a very low probability of producing a calamitous loss may be rationally evaluated as though they were sure to produce the penny. The arbitrarily calamitous risk may rationally

⁸⁰For simplicity, I will make the standard assumption that an agent's utilities over money are linear. This simplification could be dispensed with so long as the agent's utilities remained unbounded.

⁸¹If the coin never lands Heads the payoff of the game is 0.

⁸²The St. Petersburg game can be defined over the same setup, but with payoffs of $\$2^n$. Ordered according to their probabilities, the outcomes of each lottery have probabilities of $\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}$, etc. In Bad News, these outcomes have values of $-\$2, -\$16, -\$128, -\$1024$, etc., thus making their contributions to expected utility $-\$1, -\$4, -\$16, -\$64$, etc. In the St. Petersburg game, these outcomes have values of $\$2, \$4, \$8, \$16$, etc., thus making their contributions to expected utility $\$1, \$1, \$1, \$1$, etc.

⁸³A version of Penny or Doom being a lottery of the form Penny or Doom _{n} for some $n > 0$.

⁸⁴The probabilities of the outcomes in Bad News approach 0 asymptotically. Therefore whatever the value for ϵ set for Bad News, some (indeed, infinitely many) of its outcomes will have probabilities less than ϵ .

⁸⁵Weak Consistency implies that in a decision problem involving any number of lotteries (even infinitely many) if an agent truncates one of the lotteries at ϵ that agent must truncate all of the lotteries at ϵ . Repeated applications of the stated principle easily secure this result.

be ignored. But it seems deeply wrong to ignore the obviously salient possibility of calamitous loss.

Given Weak Consistency, RNP must license taking arbitrarily much risk for arbitrarily little reward. But this untenable result is hard to avoid even if one gives up on Weak Consistency. On pain of violating dominance reasoning, RNP must still license taking arbitrarily much risk for arbitrarily little reward. And Smith cannot easily give up on dominance reasoning. Not only does Smith explicitly endorse dominance reasoning, he represents RNP's consistency with dominance reasoning as one of its major strengths. "One important upshot of the approach will be that it allows us to *retain* dominance" (Smith 2014).

Let us again consider an agent who must choose between Status Quo, Bad News, and all versions of Penny or Doom. Given only dominance reasoning Smith must still permit an agent to prefer infinitely many versions of Penny or Doom to Status Quo. Here's the proof: Again, RNP mandates that only finitely many of the terms in Bad News cannot be ignored, thus making it possible to assign Bad News a finitely negative value. But the expected values of the versions of Penny or Doom decrease without limit. Therefore, if the versions of Penny or Doom must be valued according to their expectations, Bad News will be preferable to infinitely many of them. But any version of Penny or Doom is, by dominance, preferable to Bad News—each version of Penny or Doom is just Bad News with all but one of its bad outcomes replaced by a nice penny. Therefore infinitely many versions of Penny or Doom need not be valued according to their expectations. These versions of Penny or Doom must have bad outcomes which may be ignored, thus these versions of Penny or Doom may be rationally valued at a penny. And once again, a penny is better than nothing.

RNP allows an agent evaluating the versions of Penny or Doom to ignore the very low probabilities of the very, very bad outcomes. RNP licenses looking at such gambles and thinking, "Ooh! Free penny!". But the versions of Penny or Doom do not offer a free penny. They offer pennies that come with terrible risk. Such gambles should not be taken blithely. Such gambles should not be taken at all.⁸⁶

RNP is an apparently flexible principle, yet it is subject to serious constraint. On pain of violating dominance reasoning, RNP must license taking arbitrarily much risk for arbitrarily little reward. RNP makes for a tractable decision theory, but not a sound one. Probabilities cannot be rationally neglected.

⁸⁶This judgement is widely, but not universally shared. See Alexander 2011 for a dissenting view.

References

- Aboodi, Ron, Borer, Adi, and Enoch David, “Deontology, Individualism, and Uncertainty: A Reply to Jackson and Smith,” *Journal of Philosophy* 105 (2008): 259-72.
- Alexander, J. McKenzie, 2011. “Expectations and Choiceworthiness” *Mind*, 120 (479), pp. 803-817.
- Aumann, Robert, 1976. “Agreeing to Disagree” *The Annals of Statistics*, 4 (6), pp. 1236-1239.
- Brown, Campbell. “Consequentialize This.” *Ethics* 121 (2011): 749 - 71.
- Chisholm, Roderick. “The Problem of the Speckled Hen.” *Mind* 51 (1942): 368 - 73.
- Egan, Andy, 2010. “Some Counterexamples to Causal Decision Theory.” *Philosophical Review* 116, pp. 93-114.
- Elster, Jon. *Ulysses and the Sirens*. Cambridge: Cambridge University Press, 1979.
- Gibbard, Allan & Harper, William. 1978. “Counterfactuals and Two Kinds of Expected Utility.” In *Foundations and Applications of Decision Theory*, vol. 1, ed. Clifford Hooker, J. J. Leach, and Edward McClennen, pp. 123-62.
- Gustafsson, Johan, 2011. “A Note in Defence of Ratificationism.” *Erkenntnis* 75(1), pp. 147-150.
- Hájek, Alan. “Arguments For—Or Against—Probabilism?” *British Journal for the Philosophy of Science* 59 (2008): 793 - 819.
- Harman, Gilbert. *Change in View: Principles of Reasoning*. Cambridge: MIT Press, 1988.
- Hawthorne, John and Srinivasan, Amia. “Disagreement Without Transparency: Some Bleak Thoughts,” in *The Epistemology of Disagreement: New Essays* edited by David Christensen and Jennifer Lackey, 9 - 30. Oxford: Oxford University Press, 2013.
- Hintikka, Jaakko. *Knowledge and Belief: An Introduction to the Logic of the Two Notions*. Ithaca: Cornell University Press, 1962.

Jackson, Frank, 1991. "Decision-Theoretic Consequentialism and the Nearest and Dearest Objection." *Ethics* 101, pp. 461-82.

Jackson, Frank and Smith, Michael. "Absolutist Moral Theories and Uncertainty." *The Journal of Philosophy* 103 (2006): 267 - 83.

Kant, Immanuel. *Perpetual Peace: A Philosophical Essay*. London: George Allen and Unwin, 1917.

Kemeny, John, 1955. "Fair Bets and Inductive Probabilities." *Journal of Symbolic Logic* 20 (3), pp. 263-273.

Lehman, R. Sherman, 1955. "On Confirmation and Rational Betting." *Journal of Symbolic Logic* 20 (3), pp. 251-262.

Lewis, David, 1981. "Causal Decision Theory." *Australasian Journal of Philosophy* 59, pp. 5-30.

Lewis, David. "Elusive Knowledge." *Australian Journal of Philosophy* 74 (1996): 549 - 67.

Lewis, Clarence Irving. "The Given Element in Empirical Knowledge." *The Philosophical Review* 61 (1952): 168 - 75.

Mahtani, Anna, 2015. "Dutch Books, Coherence, and Logical Consistency." *Noûs* 49: pp. 522-537.

Neumann, John von and Morgenstern, Oskar. *Theory of Games and Economic Behavior*. Princeton: Princeton University Press, 1953.

Paul, Laurie. *Transformative Experience*. Oxford: Oxford University Press, 2014.

Paul, Laurie, 2015. "What You Can't Expect When You're Expecting" *Res Philosophica*, 92 (2), pp. 1-23.

Pettigrew, Richard, 2015. 'Transformative Experience and Decision Theory' *Philosophy and Phenomenal Research*, 91 (3), pp. 766-774.

Pettigrew, Richard, Forthcoming. "Review: L. A. Paul, *Transformative Experience*" *Mind*.

Smith, Nicholas J. J., 2014. "Is Evaluative Compositionality a Requirement of Rationality?" *Mind*, 123 (490), pp. 457-502.

Weirich, Paul. “Causal Decision Theory”, *The Stanford Encyclopedia of Philosophy* (Winter 2012 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/entries/decision-causal/>.

Williamson, Timothy. *Knowledge and Its Limits*. Oxford: Oxford University Press, 2002.

Timothy Williamson. 2008. Why epistemology cannot be operationalized. In *Epistemology: New Essays*, ed. Quentin Smith, pp. 277-300.