REALISTIC DECISION

A Dissertation Presented

by

JUSTIN DEALY

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 16, 2026

Department of Philosophy

REALISTIC DECISION

A Dissertation Presented

by

JUSTIN DEALY

Approved as to style and content by:

Christopher J. G. Meacham, Chair

Phillip Bricker, Member

Alejandro Pérez Carballo, Member

Sophie Horowitz, Member

Seth Cable, Member

Kevin Klement, Department Chair
Department of Philosophy

DEDICATION

To my younger self, and to my family.

ACKNOWLEDGMENTS

I am grateful to have known and befriended many good people while at UMass. Vince Abruzzo, Elle Benjamin, Lu Chen, Donovan Cox, Andrea Daventry, Thomas Daventry-Shea, Cruz Davis, Jonathan Dixon, Ed Ferrier, Cameron Gibbs, Patrick Grafton Cardwell, Bob Gruber, Scott Hill, Ali Jawin, Casey Knight, Jordan Kroll, Victor Ma, Moisés Macías Bustos, Gray Maddrey, Tricia Magalotti, Ergo Miller, Kee Nahm, José Navarro Talavera, Luis Oliveira, Ryan Olsen, Molly O'Rourke-Friel, Bailie Peterson, Crystal Plahuta, Jillian Powell, Julie Rose, Sam Schechter, Kim Soland, Bex Staneslow, David Turon, Aaron Washington, and Wally Wirchnianski — to you all, and to all members of the UMass faculty whom I've shared time with and learned from, *thank you*. Thank you as well to the wonderful Beth Grybko, and to my dear friends, Patrick Campbell and Louis Gularte, for their support, as well as to my first mentor, Kelly Jolley, and my teachers at NIU, Geoff Pynn, David Buller, Carl Gillett, and Tomis Kapitan. I wish to thank all of my students at UMass, WNE, and CCSU, from whom I've learned much about how to do philosophy. I am most indebted to Chris Meacham. My first pedagogical gig was TA-ing for Chris, a job I signed up for three times. From the start, I was blown away by the discipline and consideration Chris applies to every aspect of designing and leading a philosophy course; my own teaching, I hope, bears a permanent stamp left by him. As my dissertation chair, his meticulous criticism of my work over many years has been a source of joy and foundational guidance. Chris's forensic approach to philosophy — his patience for complexity, his refinement in weighing intuitions, and his dispassionate pursuit of plausible unifying explanations — continually inspires me. I am also very grateful to Phil Bricker, Alejandro Pérez Carballo, and Sophie Horowitz. Phil made doing philosophy at UMass exciting and daunting; he always held my work to the highest standards, and in doing so, helped me to see through irrelevancies to what matters. Beyond just their outstanding philosophical help, I thank Alejandro and Sophie for their friendly camaraderie over the years. Finally, thank you Mom and Dad for believing in me — I love you both.

ABSTRACT

REALISTIC DECISION

MAY 16, 2026

JUSTIN DEALY

B.A., AUBURN UNIVERSITY

M.A., UNIVERSITY OF MASSACHUSETTS AMHERST

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Christopher J. G. Meacham

This is a dissertation in three parts. The parts relate to one another thematically; the theme is realistic rational decision. The first part is a study of a classic challenge to applying formal decision theory to real-world decision-making situations. The second part is about practical reason's relation to desire; it argues that extrinsic desires of a certain kind can be the sole fundamental grounds of subjective practical reasons. The third part presents a new decision theory based on a subjective conception of conditional belief and credence.

TABLE OF CONTENTS

LIST OF FIGURES

Figure **Page**

PART I: THE PROBLEM OF SMALL WORLDS

CHAPTER 1

INTRODUCTION AND DECISION THEORY BACKGROUND

1.1 Introduction

Ordinary real world decision-making has a familiar structure. A few options are present to the mind, and a few possible states of the world. One considers what each option's outcome will likely be in each of the possible states of the world, and in light of this, attempts to choose in such a way as to maximize the likelihood of desire satisfaction. In this first part of the dissertation, I examine a challenge to decision theories that wish to account for such everyday real world decision-making. This challenge is known as the Problem of Small Worlds. Structurally, the challenge has its roots in the work of Leonard Savage (1972) but recent thinking on it has been shaped by the work of James Joyce (1999).[1] The idea is to show that decision theory, which is ordinarily studied under strong idealizing assumptions about the psychology of decision-makers, can be usefully applied to more realistically modeled agents. Following Joyce, a "realistically modeled" decision-maker is given by (1) a pair of arbitrary orthogonal partitions of logical space, corresponding to the agent's conception of their options and the possible states of the world, (2) a pair of functions on the intersections of the cells of those partitions, representing the agent's estimates as to the likelihoods and desirabilities their ideal self would associate with those intersections, and (3) a traditionally-modeled ideal agent, with its own option and state partitions and precise credence and utility functions, representing the agent's ideal self. As Joyce conceives the challenge, decision theory must first be extended to realistic agents modeled in the way just described, and must tell us both, for any such agent, what estimates are correct relative to their ideal self, and which element out of the set modeling the agent's conception of their options is licensed by their correct estimates. Second, decision theory

[1] The quick summary I'm about to give will make use of terms that will be introduced with more care and explication in later sections. Readers who find aspects of this introduction opaque are encouraged to nevertheless read on; explanations are coming.

must be able to show that any such agent who chooses a licensed option will in some appropriate sense have acted in a way that coheres with what decision theory says is rational for their ideal self to do. Joyce argues that in order to meet the challenge, decision theory must be formulated so as to be *partition invariant*. He argues that because it cannot be so formulated, the Causal Decision Theory of David Lewis (1981b) cannot meet the challenge, whereas the Evidential Decision Theory of Richard Jeffrey (1965) and versions of Causal Decision Theory based on imaging like the theory of Jordan Howard Sobel (1978, 1986) can. After presenting essential background in the remainder of this chapter, I begin this first part of the dissertation by pointing out that the Problem of Small Worlds consists of two separate challenges differing in the degree of difference permitted between one of Joyce's realistically modeled agents and its ideal self (§2). The Restricted Problem of Small Worlds considers only those agents whose ideal selves possess the same set of options, and thus differ from their realistic selves only by having credence and utility functions and by possibly having a more detailed set of possible states of the world. The Unrestricted Problem of Small Worlds extends to those realistically modeled agents whose ideal selves can have in addition a more numerous and detailed array of options. I show that partition invariance suffices for a solution to the Restricted Problem (§3) but that the Unrestricted Problem is unsolvable on its most natural construal (§4). I close by surveying some competing answers to the question of what norms govern the relationship between realistic agents and their ideal selves when the latter represent their choice via a more detailed range of options (§5). I give some reasons for thinking that a form of pluralism about such norms is best.

1.2 Propositions

There are different approaches to decision theory. The one I follow builds its models using propositions thought of as sets of possible worlds.[2] The set of all possible worlds is Ω. Propositions

[2]For a quick introduction to the topic of possible worlds see (Jeffrey, 1965, Ch. 12, §8). For more in-depth introductory material, see Kripke (1980); Lewis (1986); Stalnaker (1984b, 2012). A leading alternative approach to decision theory uses actions thought of as functions from situations to outcomes (Savage (1972); Buchak (2017)).

are subsets of Ω—the set of propositions is $\mathcal{P}(\Omega)$.[3] Hence, propositions A and B are identical iff they contain the same possible worlds.

Propositions are the contents of beliefs and desires; they are either true or false. Among the various possible worlds one is actual. A proposition is true at all and only the possible worlds it contains, and is true *simpliciter* iff it contains the actual world.

One proposition $\top$ is true regardless of what the actual world is like; that is, $\top$ is the only necessarily true proposition, i.e., $\top = \Omega$. Every proposition has a unique negation; the negation $\overline{A}$ of a proposition A is a proposition that is true if and only if A is false. The negation of $\top$ is the necessarily false proposition $\bot$ and it is false regardless of what possibility obtains, i.e., $\bot = \varnothing$.

The conjunction of propositions A and B is $AB = A \wedge B \stackrel{\text{df}}{=} A \cap B$. (I will only use '$\wedge$' when clarity seems to require it.) The disjunction of A and B is $A \vee B \stackrel{\text{df}}{=} A \cup B$. The conjunction of a set of propositions $\Gamma \subseteq \mathcal{P}(\Omega)$ is $\bigwedge \Gamma \stackrel{\text{df}}{=} \{w : \forall P \in \Gamma, w \in P\}$. The disjunction of a set of propositions $\Gamma \subseteq \mathcal{P}(\Omega)$ is $\bigvee \Gamma \stackrel{\text{df}}{=} \{w : \exists P \in \Gamma, w \in P\}$.

Two propositions A and B are *compatible* iff their intersection is nonempty, i.e., iff $AB \neq \bot$. A proposition A *entails* another B iff it is a subset of it, i.e., iff $A \subseteq B$.

1.3 Partitions

A *partition* of $\Psi \subseteq \Omega$ is a set Γ of nonempty propositions such that:

1. Propositions in Γ are pairwise incompatible (i.e., for $X, Y \in \Gamma$, $XY = \bot = \varnothing$).

2. The disjunction of all the propositions in Γ is Ψ.

A partition of logical space (just "a partition" for short) is a partition of Ω.[4] Putting it in terms of sets, a partition is a set Γ of nonempty sets of worlds, such that any given world is in at least and at most one member of Γ. The set of all partitions of Ω is $\mathbb{P}(\Omega)$. A *refinement* Γ^* of a partition Γ is a partition such that any cell of Γ is a disjunction of cells of Γ^* — for any $P \in \Gamma$, there is some

[3] Nothing I will say will require the assumption that Ω is infinite, so unless specified otherwise we will tacitly assume Ω and thus $\mathcal{P}(\Omega)$ to be finite but large sets.

[4] Whenever I mean to speak of a partition that is not a partition of logical space, I will be explicit about it.

$\Gamma^{**} \subseteq \Gamma^{*}$ such that $\bigvee \Gamma^{**} = P$. The propositions in a parition are commonly referred to as *cells* of that partition.

1.4 Decision Situations

Decision theory can be usefully viewed as the study of *decision situations*. Decision situations are abstract models of real-world choice situations understood psychologically.[5] Each decision situation is a bundle of formal objects, and each object in a decision situation represents an element of a real-world choice situation essential to its rational analysis. A decision situation contains, in addition to a credence function and a utility function, a set of *options*, a set of *states*, and a set of *outcomes*. The metaphysics of actions and events, etc., are all abstracted away. Options, states, and outcomes are all identified with propositions.

Let's start with options. Each option in a decision situation is a proposition. If one of Zek's options is to play dabo at t, then decision theory identifies this with the proposition $\{w \in \Omega :$ Zek plays dabo in w at $t\}$.[6] The options in a decision situation make up a partition.

Another important partition in a decision situation comprises the various *states* that the agent thinks the world might be in prior to their decision. Each state in a given decision situation is assumed to be logically compatible with each option. When the options are conjoined with the states, the result is a partition of *outcomes*.

Generalizing, a triple of partitions $(\mathcal{O}, \mathcal{S}, \mathcal{X})$ is a *representation base* if and only if (i) for any $O \in \mathcal{O}$ and $S \in \mathcal{S}$ we have $OS \in \mathcal{X}$, and (ii) for any $X \in \mathcal{X}$ there is $O \in \mathcal{O}$ and $S \in \mathcal{S}$ such that $X = OS$. Intuitively, $\mathcal{O}$ contains options, $\mathcal{S}$ contains states, and $\mathcal{X}$ contains outcomes. The set of all representation bases is $\mathbb{B}$.

The set $\mathbb{D}$ of all decision situations (or what we will later sometimes call "orthodox" decision situations) is the set of all triples consisting of a credence function, a utility function[7], and a representation base:

---

[5]Perhaps it is better to say a generalization of a model of a psychological choice situation, as most decision situations never have real-world instances.

[6]This is a *Star Trek: Deep Space Nine* reference, if you're curious — a show I highly recommend!

[7]I will explain what these functions are momentarily.

$$(c, u, (\mathcal{O}, \mathcal{S}, \mathcal{X})). \tag{1.1}$$

To refer to the objects of a decision situation I will often use subscripts in what I trust will be a self-explanatory fashion: $c_s, u_{s_1}, \mathcal{O}_x$, etc.

1.5 Credence and Probability

We represent a belief state with a *credence* function $c : \mathcal{P}(\Omega) \to [0, 1]$ from propositions to real numbers between zero and one, inclusive.[8] As we've just seen, each decision situation s comes with a credence function c_s. Credence functions represents degrees of belief or confidence. For instance, a credence of 0.5 in a proposition P would represent 50% confidence in the truth of P, a credence of 1 would represent perfect certainty of the truth of P, and a credence of 0 would represent perfect certainty of the falsity of P.

We assume credence functions are *probability* functions. For us, a probability function is a function from propositions to real numbers such that:

1. Probabilities are non-negative, i.e., for any proposition P, $p(P) \geq 0$.

2. If P and Q are incompatible, i.e., have an empty intersection, then $p(P \vee Q) = p(P) + p(Q)$.

3. The necessary proposition gets probability one, i.e., $p(\top) = 1$.

These assumptions imply that credences must always fall in the interval between zero and one. They also imply that there is such a thing as *conditional credence*, which is defined in the same way as *conditional probability*. For any given probability function p and propositions P and Q, we define the probability of P conditional on Q like so:

$$p(P \mid Q) \stackrel{\text{df}}{=} \frac{p(PQ)}{p(Q)} \text{ if } p(Q) > 0.$$

We note that for any probability $p(\bullet)$ and proposition P, $p(\bullet \mid P)$ is undefined iff $p(P) = 0$, and that if $p(\bullet \mid P)$ is defined, then it is itself a probability function. We note further that $p(\bullet) = p(\bullet \mid \top)$.

[8] Note that we assume the domain of credence functions is the set of all propositions $\mathcal{P}(\Omega)$. This is one of the things that distinguishes them from the estimate functions we will meet in the next chapter.

The definition of conditional credence is identical to the definition of conditional probability and shares the characteristics of conditional probability just mentioned.[9]

It is natural to think of credence in P conditional on Q as roughly confidence in P on the supposition that Q is true.[10] On this identification of conditional credence with suppositional confidence, if you're 20% confident that tomorrow will be sunny and breezy and you're 30% confident that tomorrow will be breezy, then you're approximately 66.6% confident that it will be sunny tomorrow conditional on its being breezy tomorrow. We will not take the identification of conditional credence with suppositional confidence for granted — at least not in the context of decision-making. Instead, we will make use of a more general notion of *imaging*, characterized below (§1.10), of which conditional credence is an instance.

Two propositions P and Q are *independent* relative to a probability (credence) function when the probability (credence) the function assigns to P conditional on Q equals the probability (credence) assigned to P *simpliciter*.[11] Probability *simpliciter* is sometimes called *unconditional* or *prior* probability to distinguish it from conditional probability.

Most decision theories do not work strictly with credence. They instead use credence to define their own special functions. Accordingly, we will assume only that each decision theory $\mathcal{T}$ comes with its own function $p^{\mathcal{T}}$ taking a decision situation and a non-empty proposition to a probability function: $p^{\mathcal{T}} : s \in \mathbb{D}, P \in \mathcal{P}(\Omega) \setminus \{\varnothing\}, Q \in \mathcal{P}(\Omega) \mapsto p_s^{\mathcal{T},P}(Q) \in [O, 1]$. Sometimes I call these functions "probability" functions, though strictly speaking they are mappings to such functions. I trust this will be no cause for confusion, as when I speak of these functions I'll always speak in terms of the probability function *of some theory*.

1.6 Value and Utility

Each decision theory $\mathcal{T}$ will also have a *value function*, $v^{\mathcal{T}} : s \in \mathbb{D}, P, Q \in \mathcal{P}(\Omega) \setminus \{\varnothing\} \mapsto v_s^{\mathcal{T},P}(Q) \in \mathbb{R}$. A theory's value function is supposed to numerically represent the agent's desire state, or some

[9] This is a non-trivial assumption. See e.g. Hájek (2003) for a lengthy argument against the idea that credence conditional on a proposition is undefined when credence in that proposition is zero.

[10] This will be questioned in Part III of this dissertation. But for now, it's fine.

[11] Notably, when this is the case $p(P \mid Q) = p(P) \times p(Q)$.

significant aspect thereof. A value function formally speaking is just a function taking a decision situation s and a pair of (nonempty) propositions (P, Q) to a real number, where the idea is that the higher $v_s^{\mathcal{T},P}(Q)$ is, the more Q is desired by an agent occupying s given P, according to $\mathcal{T}$.[12]

What are value functions supposed to be defined in terms of? Officially, we leave that up to decision theory, but the natural expectation is that they are defined in terms of a more basic kind of desire representation, which we call a *utility function*. Utility functions represent agents' desires and preferences for maximally detailed possibilities, i.e., possible worlds. So, the set of utility functions is just the set of functions from worlds—or more literally, singleton propositions—to real numbers.

As stated in §1.4, part of what is primitively given[13] at any decision situation s will be a utility function. Strictly speaking, utility functions as we will understand them are undefined for propositions other than possible worlds, but from a given utility function u^* we can derive a family of *utility-level propositions*.[14] For a given utility function u^*, the utility-level proposition $u^*[n]$ is true at a possible world w iff $u^*(w) = n$. So, if a proposition P entails a utility-level proposition $u^*[n]$, every P-world is assigned n by the utility function u^*.

We will make the following assumption about the relationship between utility and $\mathcal{T}$-value:

Assumption 1 (Utility Floor)**.** *For any decision theory $\mathcal{T}$, decision situation $s \in \mathbb{D}$, $n \in \mathbb{R}$, and $P, Q \in \mathcal{P}(\Omega) \setminus \{\varnothing\}$, $v_s^{\mathcal{T},P}(Q) = n$ if PQ entails $u_s[n]$.*

1.7 Expected Utility

We can now broadly characterize *expected utility* (EU) functions. Generally speaking, the EU for a given option is a probability-weighted average of values of the outcomes compatible with that

[12] We leave it entirely up to decision theories to decide exactly what "desiring Q given P" means. Though it will not figure in anything I do, it is typically assumed that utility is only unique up to affine transformations — the zero point and the "distance" in utility between any two propositions are considered arbitrary and correspond to nothing psychologically real.

[13] Many strands of decision theory, going back to Savage (1972), do not take credences and utilities as primitives, and instead think of them as representations of an underlying preference ordering (cf. (Joyce, 1999, §3 and §7)). I side with Meacham and Weisberg (2011) in thinking that approach is not obligatory.

[14] Cf. Lewis (1981b).

option, or as we will represent things, of the states *given* that option.[15] Our assumption will be that any theory's EU function will be definable in accordance with the following basic format:

$$\forall s \in \mathbb{D}, \forall O \in \mathcal{O}_s, EU_s(O) \stackrel{\text{df}}{=} \sum_{S \in \mathcal{S}_s} p_s^{\mathcal{T},O}(S) v_s^{\mathcal{T},O}(S). \tag{1.2}$$

Note that in the definition of expected utility each theory uses its own probability (§1.5) and value (§1.6) functions.

1.8 Maximizing

Different versions of decision theory make different claims about which options in decision situations are rational. Perhaps the leading idea among philosophers of decision theory is that rationality requires expected utility to be maximized:

Expected Utility Maximization (EUM): For any given decision situation s, an option $O \in \mathcal{O}_s$ is rationally permissible if and only if there is no alternative option $O^* \in \mathcal{O}_s$ with higher expected utility in s than O.

Some decision theories[16] reject EUM, but I will not be interested in such theories in this dissertation.

1.9 Decision Theories in the Abstract

For our purposes a decision theory $\mathcal{T}$ is a tuple $\mathcal{T}(s) = (p^{\mathcal{T}}, v^{\mathcal{T}}, \mathcal{I}^{\mathcal{T}}, \mathcal{U}^{\mathcal{T}})$. The first two members of $\mathcal{T}(s)$ are familiar at this point. The third member is what we call the theory's *imaging* function. This member hasn't appeared yet in relation to expected utility, but it will play a major role in our later discussions. The fourth member is what we will call the theory's *utility frame*. This will be relevant in the next chapter, so we will hold off on characterizing it until then. Let us turn now to imaging functions.

---

[15] If it's value that's getting averaged, why not call it expected value? Good question! The traditional terminology is "expected utility," and in generalizing by speaking of theories' value functions, we are already departing somewhat from typical terminology. I'd rather keep one foot in tradition than talk in a way that is entirely idiosyncratic. Also, there's already another kind of function, which will prove important in later chapters and that we will discuss below, known as an "expected value" function.

[16] See, e.g., Buchak (2017).

1.10 Imaging Functions

For given $\mathcal{T}$ and $s \in \mathbb{D}$, $\mathcal{I}_s^{\mathcal{T}}$ is a function taking a proposition P to a probability function $\mathcal{I}_s^{\mathcal{T}}(\bullet \parallel P)$ such that $\mathcal{I}_s^{\mathcal{T}}(P \parallel P) = 1$, and for all Q, $\mathcal{I}_s^{\mathcal{T}}(Q \parallel P) \geq c_s(PQ)$.[17] We call $\mathcal{I}_s^{\mathcal{T}}$ the *imaging* function of $\mathcal{T}$ for s. Note that this definition implies that $\mathcal{I}_s^{\mathcal{T}}(Q \parallel P) = \sum_{w \in Q} \mathcal{I}_s^{\mathcal{T}}(w \parallel P)$, and that in general for any partition Γ, $\mathcal{I}_s^{\mathcal{T}}(Q \parallel P) = \sum_{X \in \Gamma} \mathcal{I}_s^{\mathcal{T}}(PX \parallel P) = \sum_{X \in \Gamma} \mathcal{I}_s^{\mathcal{T}}(X \parallel P)$. It also follows that for all P, $\mathcal{I}_s^{\mathcal{T}}(P \parallel \top) = c_s(P)$. We leave $\mathcal{I}_s^{\mathcal{T}}(P \parallel \bot)$ undefined.

A decision theory's imaging functions embody what according to that theory constitutes the rational way to shift one's credences given a *practical supposition*.[18] We make practical suppositions when considering possible courses of action. They correspond to expressions like,

- "Suppose I choose to do this..."
- "Let us suppose I act as follows..."
- "Suppose I proceed in that way..."
- "If I do this..."
- "If I were to do that..."

I hasten to add, these canonical examples simply pick out a type; we are not to take for granted that this type is as a rule restricted to antecedents about action performance. On a standard sort of formal epistemological explication, what happens when one makes a practical supposition is that one temporarily "shifts" all of one's credence to a particular proposition, and then in light of the resulting distribution of credence over hypotheses compatible with that proposition considers how likely some such hypotheses are.

Clearly this is linked to conditional thought. In some sense when one makes a canonical practical supposition, one is considering what will or would follow *if* one makes a particular choice. There is an ongoing debate over whether practical supposition is best understood as indicative or subjunctive.

---

[17] Compare this characterization to SUP_1-SUP_3 (Joyce, 1999, p. 188).

[18] These will not only be of central importance in this first part of the dissertation. They will take center stage in Part III as well.

We will make no assumption about who is winning this debate. Instead, we will allow that theories can define imaging functions in what they view as an indicative way or in what they view as a subjunctive way.[19] Now, even bracketing the question of how to properly understand the nature of practical supposition (e.g., as indicative or subjunctive), one might think any decision theory should identify the probability of a state imaged on an option in a given decision situation with the probability assigned to that state relative to that option, insofar as the latter occurs in the definition of expected utility. This would posit an especially direct connection between practical supposition and expected utility. This is not a trivial assumption, and as far as I assume in this part of the dissertation, decision theories may or may not make it.[20]

Summing up, for the sake of clarity in thinking about aspects of rational deliberation to be introduced later, we will assume that a decision theory is partly defined by what function it takes as its imaging function. Decision theories that are otherwise identical save for their respective conceptions of imaging will be separate for us. If this assumption means we distinguish as different theories that have often been treated as one, let us hope our penchant for precision will be welcomed, or at least deemed harmless.

1.11 Expected Value

One specific way of defining a value function (§1.6) is known as *expected value*. For a given decision situation s, the expected value $\mathbf{V}_s$ of a proposition P is an average of the utilities assigned in s to worlds where the proposition is true, weighted by the credence in s in those worlds conditional on that proposition. That is, for any $P \in \mathcal{P}(\Omega)$, $\mathbf{V}_s(P) \stackrel{\text{df}}{=} \sum_w c_s(w \mid P) u_s(w)$. When we speak of expected value as "value function," what we mean is as follows. Recall that a value function assigns a value, in a decision situation, to a proposition Q given another proposition P. To use expected value for this purpose, we conjoin the two propositions. That is, if in our theory $\mathcal{T}^*$ we use expected value as our value function, then $v_s^{\mathcal{T}^*,P}(Q) \stackrel{\text{df}}{=} \mathbf{V}_s(PQ)$.

[19] In Part III I will argue that this opposition between moods is misconceived in the context of practical supposition.

[20] In Part III I will defend a theory that does posit such a close tie, but my business at this stage of the dissertation is not in defending any particular theory.

1.12 Newcomb's Problem

A lot of research in philosophical decision theory has its origin in a famous semi-paradoxical puzzle known as Newcomb's Problem.[21] As with many such puzzles, the basic features of Newcomb's Problem are easy to describe, but there is no expert consensus on its solution. Here's how it goes.

Newcomb's Problem: In front of Robert on a table is an opaque box and a transparent box. Robert sees that the transparent box contains one thousand dollars cash. He is certain the opaque box is either empty or contains one million dollars cash. Robert can take whatever is in the opaque box and leave the thousand ("one-box"), or he can take the thousand together with whatever is in the opaque box ("two-box"). He has no other options.[22] Robert is certain of the following:

1. Earlier in the day an extremely reliable (but not infallible) Predictor of Human Behavior predicted whether the agent would one-box or two-box.
2. The Predictor put the million in the box earlier in the day if and only if she predicted that the agent would one-box when the time came to decide.

Robert wants the million more than the thousand, and wants the million together with the thousand most of all. He has no other relevant desires.

Given his desires, what is it rational for Robert to do here? Perhaps both options are acceptable and Robert will be uncriticizable no matter what he does. On the other hand, perhaps one of the two options is irrational. If so, which one? Answering these questions is the puzzle.

The reason this is seen as such a hard problem is that for either option a powerful case can be made to the effect that it is the uniquely rational choice. Take one-boxing. One-boxing is attractive because Robert knows how reliable the Predictor is. We can flesh this out with more storytelling. Suppose the Predictor has a track record of 99.99% reliability, and she's been doing this every day in full public view for decades. Robert is, if not certain that the Predictor's prediction will prove

[21] Newcomb's Problem was first discussed in print by Robert Nozick (1969).

[22] Assume that doing nothing is equivalent to one of these options, e.g., by an unavoidable bank transfer.

correct in his own case, a hair's breadth from such certainty. He has zero reason to think he's going to be an exception to the trend; if he is, he is sure it would be a fluke. Put yourself in Robert's shoes. With that level of confidence in the prediction, would you really feel comfortable reaching for both boxes? Reflect on this: 99.99% of previous one-boxers have walked off millionaires, whereas 99.99% of previous two-boxers have gone home with only a thousand. How could two-boxing be the rational move given such a lousy track record? But now consider the case for two-boxing. Robert knows the prediction was made already; the million is either already sitting on the table in the opaque box or it isn't — nothing he can do now will change that. In light of this simple fact, doesn't it seem foolish not to take everything on the table? Suppose the million is there and Robert one-boxes. He has become a millionaire like so many one-boxers before him — but wait, he would've been a millionaire regardless! All he really did was turn down an extra thousand like all the one-boxers before him. How could one-boxing be rational when it predictably refuses the only valuable thing Robert can control: the thousand?

If only decision theory could solve the puzzle instantly by simply directing Robert to maximize expected utility.[23] Unfortunately, that advice seems to many to be useless. As we will see, expected utility can be defined in a way that implies one-boxing has higher expected utility than two-boxing, or it can be defined to imply precisely the reverse. So, where we might have liked decision theory to cut through everyone's intuitions and give us a clear and independently motivated answer, the opposite seems to be what happens; intuitions about Newcomb's Problem, and about a bevy of spin-off puzzles collectively known as *Newcomblike* puzzles, divide decision theory into competing schools of thought. The two most prominent of these embrace EUM but disagree on its proper interpretation, with one side maintaining it's one-boxing in Newcomb's Problem that is uniquely rational and the other maintaining it's two-boxing.

1.13 Evidential Decision Theory

Evidential Decision Theory (EDT) is the natural home for decision theorists whose considered judgment is that one-boxing is rationally required in Newcomb's Problem. The theory is due to

[23] This is just the sort of solution that I will argue for in Part III.

Richard Jeffrey (1965) and has more recently been notably defended by Arif Ahmed (2013, 2014). EDT defines expected utility[24,25] in terms of expected value:

$$\forall s \in \mathbb{D}, \forall O \in \mathcal{O}_s, \mathrm{EU}_s^{\mathrm{EDT}}(O) = \sum_{S \in \mathcal{S}_s} c_s(S \mid O)\mathbf{V}_s(SO) = \mathbf{V}_s(O). \qquad \textbf{(EDT's Definition of EU)}$$

EDT gets its name because of the assumed link between conditional probability and evidence. This connection is widely recognized and is the subject of intensive study. It comes by way of the norm of rationally updating one's credences in light of receiving new evidence. Undoubtedly the leading theoretical approach to this norm is the Bayesian account on which rational updating means exchanging one's prior credence function for the function that results from conditioning it on one's evidence (see, e.g., Easwaran (2011); Meacham (2015)). On this way of understanding the significance of conditional credence, EDT can be glossed a theory that claims options must be viewed as pieces of potential evidence — or "news," as many[26] like to say — that one can enact at will. Evaluating one's options on this picture means considering what state of the world they evidentially support.

Let's apply EDT to Newcomb's Problem. Let M be the proposition that the million is in the opaque box. For our state partition we'll use the simple set $\{\mathrm{M}, \overline{\mathrm{M}}\}$ and we'll make the natural assumptions about c and u. For simplicity we also leave off the subscripts.

Let's calculate expected utilities:

$$\begin{aligned} \mathrm{EU}^{\mathrm{EDT}}(\mathrm{OneBox}) &= c(\mathrm{M} \mid \mathrm{OneBox})\mathbf{V}(\mathrm{M} \wedge \mathrm{OneBox}) + c(\overline{\mathrm{M}} \mid \mathrm{OneBox})\mathbf{V}(\overline{\mathrm{M}} \wedge \mathrm{OneBox}) \\ &\approx (1 \times 1,000,000) + (0 \times 0) \\ &\approx 1,000,000. \end{aligned} \tag{1.3}$$

[24] Cases where the agent is certain what they will do are cases where EDT does not apply, unless conditional credence is taken as primitive (Hájek (2003)). I will not engage with that approach to extending EDT.

[25] Using this definition of expected value it is easy to show that for any partition $\Gamma \in \mathbb{P}(\Omega)$, proposition $P \in \mathcal{P}(\Omega)$, and decision situation $s \in \mathbb{D}$, $\sum_{Q \in \Gamma} c_s(Q \mid P)\mathbf{V}(QP) = \mathbf{V}(P)$.

[26] The term has come to be associated with Lewis (1981b,d) but it goes back to Jeffrey (1965) who says (§5.7) it first was suggested by Savage.

$$\begin{aligned}\mathrm{EU}^{\mathrm{EDT}}(\mathrm{TwoBox}) &= c(\mathrm{M} \mid \mathrm{TwoBox})\mathbf{V}(\mathrm{M} \wedge \mathrm{TwoBox}) + c(\overline{\mathrm{M}} \mid \mathrm{TwoBox})\mathbf{V}(\overline{\mathrm{M}} \wedge \mathrm{TwoBox}) \\ &\approx (0 \times 1,001,000) + (1 \times 1,000) \\ &\approx 1,000.\end{aligned} \tag{1.4}$$

One-boxing comes out with higher expected utility than does two-boxing, and is therefore is what EDT recommends as the uniquely rational choice.

1.14 David Lewis's Causal Decision Theory

The main competitor to EDT is known as Causal Decision Theory (CDT). There isn't one specific theory that can claim exclusive rights to be called CDT — rather, CDT is more like a family of kindred decision theories. Because it is one of the most prominent members of this family, and because of the attention paid to it by Joyce (1999), we will focus on the formulation of CDT that is due to David Lewis (1981b). The most conspicuous difference between Lewis's theory and EDT is in the special role Lewis gives to what are known as *dependency hypotheses*. A dependency hypothesis for an agent with a given partition of options and utility function is a proposition that makes a maximal claim about how what the agent cares about (per the utility function) does and does not causally depend on what they choose (Lewis, 1981b, §5).

A dependency hypothesis is not to be understood as a claim about an isolated option. Rather, a single dependency hypothesis makes a causal claim about every single one of the agent's options. It's this "menu" aspect that makes the set of all dependency hypotheses relative to a given set of options and utility function into a partition — it ensures that no two dependency hypotheses in such a set can be true at once, and at least one must be true.

The dependency hypotheses in a given decision situation are standardly understood as each entailing maximally specific answers to these questions:

1. What are the things the agent cares about that are *not* under their control?
2. What are the things the agent cares about that *are* under her control?

3. How exactly do, and how exactly will, the things the agent cares about and that are not under their control stand?
4. How exactly would the things they care about that are under their control turn out depending on what they do?[27]

To present the idea more concretely let's return to Newcomb's Problem. For simplicity, assume Robert only assigns non-zero utility to the million and the thousand. One of Robert's dependency hypotheses will say (entail) the following:

- The million is not under Robert's control and it is in the opaque box.
- One-boxing will causally result in not getting the thousand.
- Two-boxing will causally result in getting the thousand.

A second dependency hypothesis will say (entail) this:

- The million is not under Robert's control and is not in the opaque box.
- One-boxing will causally result in not getting the thousand.
- Two-boxing will causally result in getting the thousand.

These two dependency hypotheses are significant in that Robert isn't certain they're false — they receive non-zero credence. Several other dependency hypotheses for the decision situation exist, but get zero credence from Robert. For example, this:

---

[27] This approach to dependency hypotheses is what you might call maximally subjectivist. Other less subjectivist approaches would insist that partitions of dependency hypotheses be less fine grained, in the sense of not allowing for variance in what matters of fact are and are not under the agent's control as represented by various members of the partition. Less abstractly put, one might wish to insist that objective facts about the agent's universe (e.g., what its laws are, what particular matters of fact are causally upstream the agent's choice) should be part of what determines the relevant partition of dependency hypotheses. Alternatively, and a bit more plausibly, one might insist that for any agent there will be certain matters of particular and general fact that are a priori out of the agent's control, namely matters of fact concerning the past leading up to just prior to the agent's moment of decision and general facts about the laws of nature. The former of these two alternatives (i.e., the less plausible one), if sound, would make trouble for my claim in §3.4 that the Lewisian partition extraction function is state blind, as the appropriate partition of dependency hypotheses would vary between possible agents with the same options and utility function. The latter, which seems much more plausible, is, as far as I can see, compatible with everything I say about Lewisian CDT and partition invariance.

- One-boxing will causally result in getting the million but not the thousand.
- Two-boxing will causally result in getting the million and the thousand.

On the typical understanding of Lewis's theory, it can only be applied to decision situations whose state partitions are partitions of dependency hypotheses. It is thought to differ in this regard from EDT, which can be applied to any decision situation so long as the agent isn't certain of what they will do.[28] Now, a partition of dependency hypotheses is determined once a utility function u and partition of options $\mathcal{O}$ is given; hence, we can refer to partitions of dependency hypotheses like so: $\mathcal{K}_u^{\mathcal{O}}$. So the partition of dependency hypotheses for a given decision situation s (i.e., $\mathcal{K}_s$), will be the one determined by the utility function and partition of options in s (i.e., $\mathcal{K}_{u_s}^{\mathcal{O}_s}$). Let the set $\mathbb{K}$ be the subset of $\mathbb{D}$ consisting of all decision situations of the form $(c, u, (\mathcal{O}, \mathcal{K}_u^{\mathcal{O}}, \mathcal{X}))$. Here then is a standard Lewisian definition of expected utility:

$$\forall s \in \mathbb{K} \subset \mathbb{D}, \forall O \in \mathcal{O}_s, \mathrm{EU}_s^{\mathrm{Lew}}(O) = \sum_{K \in \mathcal{K}_s} c_s(K)\mathbf{V}_s(OK). \qquad \textbf{(Lewis's Definition of EU)}$$

Note the use of unconditional credence. Unlike with EDT, options do not function as pieces of evidence ("news") on Lewisian CDT. This is thought to correspond to a core philosophical difference between EDT and CDT.

Later I will argue that Lewis's theory does not need to be understood as restricted to $\mathbb{K}$. Dependency hypotheses are essential to Lewis's theory but they need not, I will claim, be in the foreground when calculating Lewisian expected utility. The details will have to wait until Ch. 3. For the moment let us suppose that once an agent's utilities and options are determinate there is only one situation that can model their choice on the Lewisian view, namely the situation that uses as its state partition $\mathcal{K}_u^{\mathcal{O}}$ where u is the agent's utility function and $\mathcal{O}$ is their option partition.[29]

---

[28] Incidentally, Lewis's theory *can* be applied in cases where agents are certain what they will do. It's true that it uses expected value which is defined in terms of conditional credence, but that is a superficial presentational choice given that conjunctions of dependency hypotheses with options imply utility-level propositions. Cf. note 24.

[29] If we consider how many separate things a realistic agent cares about, and how many logically possible different ways those things could depend or not depend on that agent's options, we can easily begin to see how one might view Lewis's account as rather remote from the real world. Dependency hypotheses that are fit to model realistic cases are liable to be extremely information-laden — perhaps well beyond what we'd consider within a normal human's capacity to grasp.

One noteworthy feature of decision situations in $\mathbb{K}$ is that the outcomes in these situations imply utility levels (§1.6). For any such situation $s \in \mathbb{K}$ and outcome $X \in \mathcal{X}_s$ there is some n such that $X \subseteq u_s[n]$. In other words, the utility of an outcome in a decision situation in $\mathbb{K}$ is not an average of different utilities; rather, such an outcome guarantees a specific utility level. Hence, the use of expected value in the definition of Lewisian EU is really for notational consistency.

Let's apply Lewis's theory to Newcomb's Problem. Letting "MIB" stand for the dependency hypothesis according to which the million dollars is in the opaque box, we can apply Lewisian expected utility to Newcomb's Problem and see that two-boxing comes out as having higher Lewisian expected utility than one-boxing:

$$\begin{aligned} \mathrm{EU}^{\mathrm{Lewis}}(\text{One-Box}) &= c(\mathrm{MIB})\mathbf{V}(\text{One-Box} \wedge \mathrm{MIB}) + c(\overline{\mathrm{MIB}})\mathbf{V}(\text{One-Box} \wedge \overline{\mathrm{MIB}}) \\ &= 1,000,000 \times c(\mathrm{MIB}) + 0 \times c(\overline{\mathrm{MIB}}) \\ &= 1,000,000 \times c(\mathrm{MIB}) \end{aligned} \tag{1.5}$$

$$\begin{aligned} \mathrm{EU}^{\mathrm{Lewis}}(\text{Two-Box}) &= c(\mathrm{MIB})\mathbf{V}(\text{Two-Box} \wedge \mathrm{MIB}) + c(\overline{\mathrm{MIB}})\mathbf{V}(\text{Two-Box} \wedge \overline{\mathrm{MIB}}) \\ &= 1,001,000 \times c(\mathrm{MIB}) + 1,000 \times [1 - c(\mathrm{MIB})] \\ &> 1,000,000 \times c(\mathrm{MIB}) \end{aligned} \tag{1.6}$$

It doesn't matter what Robert's credences are — the Lewisian EU of two-boxing is always going to be higher. Two-boxing "causally dominates" one-boxing in that regardless of which dependency hypothesis obtains, the utility of the outcome of one-boxing is lower (i.e., 1000 points lower) than the utility of the outcome of two-boxing.

1.15 Partition Invariance

Some theories are *partition invariant*. A partition invariant theory $\mathcal{T}$ is one such that any two decision situations s_1 and s_2 with the same credence and utility functions and the same partition of options share the same $\mathcal{T}$-expected utility function (i.e., if $c_{s_1} = c_{s_2}$, $u_{s_1} = u_{s_2}$, and $\mathcal{O}_{s_1} = \mathcal{O}_{s_2}$, then $\mathrm{EU}^{\mathcal{T}}_{s_1} = \mathrm{EU}^{\mathcal{T}}_{s_2}$). That is, such a theory defines expected utility in such a way that once the agent's credence and utility functions are determined together with a partition of options, it doesn't matter what partition we choose as our partition of states — the same expected utilities are assigned to the

options regardless of the states and outcomes. Another way to put this makes use of equivalence classes. Consider $\text{PartInv}(\mathbb{D}) \stackrel{\text{df}}{=} \{\Gamma \subset \mathbb{D} : s_1, s_2 \in \Gamma \rightarrow c_{s_1} = c_{s_2}, u_{s_1} = u_{s_2}, \mathcal{O}_{s_1} = \mathcal{O}_{s_2}\}$. This is a set of equivalence classes. Every situation in any class $\Gamma \in \text{PartInv}(\mathbb{D})$ has the same options and the same credence and utility functions as every other situation in Γ. If a theory is partition invariant, these equivalence classes can be viewed as the "real" decision situations, where what is determinate is just what the options are, and what the agent's credences and utilities are. Given that information, the theory has what it needs to identify the rational options; whatever representation of states the agent is using can be ignored. Partition invariance will play a significant role in the coming chapters. We will introduce a formal definition for it in Chapter 3 and use that concept to prove various things about EDT and Lewisian CDT.

Before we move on, let us take quick note of the fact that EDT is partition invariant. This fact falls out of the definition of Expected Value (§1.11). Here's a quick proof.

Theorem 1. *If* $\Gamma \in \mathbb{P}(\Omega)$*,* $s \in \mathbb{D}$ *and* $O \in \mathcal{O}_s$*, then* $\mathbf{V}_s(O) = \sum_{X \in \Gamma} c_s(X \mid O)\mathbf{V}_s(OX)$.

Proof. $\mathbf{V}_s(O) \stackrel{\text{df}}{=} \sum_w c_s(w|O)u_s(w) = \sum_{X \in \Gamma} \frac{c_s(XO)}{c_s(O)} \sum_w \frac{c_s(wXO)}{c(XO)} u_s(w) = \sum_{X \in \Gamma} c_s(X \mid O)\mathbf{V}_s(OX)$. □

CHAPTER 2

THE PROBLEM

We come now to the central focus of this first part of the dissertation: the Problem of Small Worlds as characterized by James Joyce (1999, 2000).[1] In this chapter, my aim will be to come to a clear view of Joyce's understanding of this problem, provide a formal framework within which the problem can be tractably analyzed, and present a plausible formulation of it. As briefly discussed at the start of the previous chapter (§1.1), Joyce is thinking of the problem as a kind of litmus test for decision theories, in so far as they aspire to apply in realistic decision-making scenarios. The essence of the problem lies in the fact that ordinary everyday decision making seems not to come equipped with the kind of psychological precision demanded by the formalism of standard decision theories. It seems to Joyce rather that instead of having precise credences and utilities associated with our psychological state when we are trying to make a decision, it's more like we are guessing or estimating what we think our confidence levels and utility levels should be or would be if we had sufficient time and energy to think our situation through in an ideal way. Furthermore, Joyce thinks we tend to conceptualize our decision-making situations in rough and ready terms. That is to say, we don't adequately appreciate in most cases the variety of specific options available to us, or the full variety of relevant states that the world could be in. We identify a few rough options and a few rough states that the world might be in, and we proceed under the assumption that the rough outcome space induced by our assumptions is good enough. But as before, if we could ideally deliberate on our predicament, it seems natural to think that we would in most cases find ourselves presented with a much more fine-grained array of options and states than that with which we started. So the Problem of Small Worlds is about taking the formalism of decision theory, which seems suited to the ideal decision-making context with its precise credences and utilities and its

[1] The Problem of Small Worlds takes its name from Savage's name for a parallel problem facing his non-World-Theoretic decision theory. See (Savage, 1972, Ch. 5, Sec. 5).

possibly highly fine-grained options and states, and figuring out how what it says in that idealized context informs the aims, standards, and commitments of rational deliberation, so to speak, in the down-to-earth context, which Joyce calls the *small world*. Since we want decision theory to apply or at least be useful to us in our everyday activity, Joyce sees it as essential to a theory's acceptability that it provably guarantees the right relation between rationality in the small world and in the ideal setting, which he calls the *grand world*.

We will begin this chapter by surveying what Joyce says about the problem in his book and in a paper of his that touches on it (§2.1). I will then present a suite of formal concepts for modeling realistic decision situations in a way that seems appropriate to Joyce's thinking (§2.2). Next I present a further set of formal notions that will be useful in presenting the Problem of Small Worlds itself (§2.3). With the framework in place, I use it to present a formulation of the Problem of Small Worlds (§2.4). I close the chapter by clarifying two key ideas in my formulation of the problem: the "locality" of estimate correctness (§2.5) and the sufficiency of small world rational choice for compatibility with grand world rational choice (§2.6).

2.1 Joyce's Characterization

How should we model "realistic decision-making"? As Joyce sees it, an agent represents their choice situation to themselves in terms of the options they must choose between, the various possible states the world could be in, and the outcomes the options would yield given each of the states. We can of course model this nicely with our concept of a representation base, introduced in the last chapter. But a realistic agent's deliberation is rarely limited to a single static representation — realistic agents often *refine* their representation of the situation they're facing, sometimes many times, before deciding. If we can model the content of a deliberative representation at an instant with a representation base, refinement corresponds to progressive replacement representation bases with new and strictly informationally richer ones. Following Joyce, though, we're going to focus on what happens at the tail end of the process. Eventually, Joyce says, realistic agents' refinements end — they settle on a representation and decide:

> The process of deliberation is often one of *refinement*. A decision maker begins by considering one family of possible actions, states of the world, and outcomes and moves

> on to consider others that provide a fuller and more realistic picture of their situation. When it comes to making a definitive choice she will settle on a single decision problem by focusing on a fixed repertoire of acts and defining outcomes in terms of a specific partition of states. The character of the decision problem that she ultimately sets for herself will depend upon how thoroughly she reflects on her predicament. [...] Choosing is really a two-stage process in which the agent first refines her view of the decision situation by thinking more carefully about her options and the world's state until she settles on the "right" problem to solve and then endeavors to select the best available course of action by reflecting on her beliefs and desires in the context of this problem. (Joyce, 1999, 70-71).

Assuming infinite time and ability, can a person ever reach a point where further refinement of their decision situation would be practically pointless? Intuitively, yes. "It is possible," Joyce says, "to theorize about the decision problem that the agent would come to if they could somehow take all their possible options into account and could resolve the partition of states to the highest level of pertinent detail" (73). To refine one's representation of one's choice to this point would be to reach a *grand world decision*.[2] "The defining feature of a grand world decision," Joyce says, "is that *any* consideration that could be relevant to the agent's desires is taken into account in the specification of acts and states." Further refinement beyond such a stage is pointless from a practical standpoint.[3]

Agents who haven't made it to grand world decisions face *small world decisions*. For Joyce, the contrast between these two is also a matter of how *considered* one's credences and desires are:

> The beliefs and desires of an agent who faces a small world decision problem have a different character from those of a grand world decision maker. The latter attitudes are *fully considered*, whereas the former are not. A *fully considered belief* is one that a person holds after taking all the evidence at her disposal into account and drawing inferences from this evidence using the best principles of inductive and deductive logic. A *fully considered desire* is one that is based on complete and explicit reflection on all the contingencies that might affect the values of the prospects the person is considering. Fully considered beliefs and desires, in short, are the ones that an individual would hold if they were able to deliberate, in an entirely rational manner, on all issues relevant to their concerns (without additional input from the world) until further deliberation was superfluous; they are the 'fixed points' of rational deliberations. (74)

It is perhaps obvious that realistic beings like us basically never reach the grand world. Joyce says we "have neither the time nor the cognitive resources to reason ourselves anywhere near a state

[2]This terminology is due to Savage (1972).

[3]Cf. the final paragraph on p. 72.

of reflective equilibrium. We thus never really know what our fully considered attitudes are" (75). "Deliberation, if it is to be any use at all must stop well short of a grand world problem" (73).

Joyce thinks grand world decisions are rationally connected to small world decisions in two ways. First, he thinks deliberating agents' degrees of belief and desire are *implicit estimates* of the degrees of belief and desire they would have in their grand world decisions:

> [W]e can think of a rational agent's attitudes toward the states, outcomes, and acts in a small world decision problem as her *best estimates* of the attitudes that she would hold regarding those states, outcomes, and acts in the grand world context. When a decision maker forms an opinion about the likelihood of some small world state (grand world event) E, or about its likelihood relative to another state, she implicitly estimates the strength of her grand world belief in E and tries to conform her current belief to it.[4] Likewise, when she adopts a view about the desirability of a small world outcome O, or its value relative to another such outcome, she again makes an implicit estimate of the degree to which she would find O desirable in the grand world context and tries to conform her current desire to it. (76-77)

Second, he thinks that when you make a decision in the small world, you implicitly *commit yourself* to the rationality of your choice in the grand world:

> [E]very time we make a decision, no matter how ill considered, we tacitly *commit* ourselves to holding certain views about the nature of our fully considered beliefs and desires and about the evaluative judgments they would sanction. Specifically, when we choose to perform a small world act... on the basis of less than fully considered beliefs and desires we thereby commit ourselves to the view that our fully considered beliefs and desires would sanction the choice of [that option] from among the alternatives [in the small world decision situation we have in mind]. We cannot, in other words, choose to do [a given small world option] without committing ourselves to the claim that we would not rank some other act in [the small world decision situation] above [that option] were we to deliberate on the matter, in an entirely rational way, until all further deliberation became moot. (75)

Puzzlingly, perhaps, Joyce is explicit in saying rationality does not require us to choose an option in a small world decision that would be sanctioned by our fully considered beliefs and desires:

> The agent might well judge that [a given small world option] is optimal among their small world options, be firmly convinced that they would retain this opinion even if they reflected on the matter more fully, and yet it still might be true that they would come to see her choice of [that small world option] as misguided if they actually were

[4]By "grand world event", Joyce means a disjunction of grand world states. He is using "event" in the sense common to probability theory.

> to deliberate further. *It is not a mark of irrationality to have one's less than fully considered judgments conflict with one's fully considered judgments.* The most we can reasonably ask of an agent is that they reflect on the decision problem she faces until she has good reason to think that further reflection would not change their views (at least not change them enough to affect their decision). At that point her job is to stop deliberating and solve the problem. (77, emphasis added)

The idea seems then not to be that there's a rational requirement on agents in the small world to pick options that would be sanctioned by their grand world counterpart, so to speak, but rather that when we make choices in the small world, we commit ourselves to our choices being sanctioned in the grand world. It might turn out that what we commit ourselves to on a given occasion of choice fails to hold, but Joyce clearly doesn't like the idea that such failure necessarily implies irrationality. But then what kind of commitment are we making when we conclude a deliberation? Is it a commitment to ourselves, to everyone, to no one, or to someone else? Are we in any sense rationally blameworthy if we fall short on our commitment? What is an agent in a small world choice rationally required to do? For now let's table these matters. We will take them up in Ch. 5.

Joyce takes the estimation and commitment connections between the small and grand worlds to motivate a certain job for decision theory, *The Problem of Small Worlds*:

> It will not do simply to say that an agent's "true" views are the ones they would hold in the grand world context, and then to go on to provide a theory that pertains only to this highly refined scenario. *What one needs in addition is (a) some explanation of what it takes for a small world decision maker's estimates of their grand world attitudes to be correct, and (b) an account of rationality that applies to both grand- and small world decision making and that guarantees that any small world decision maker who correctly estimates their grand world attitudes, and who adheres to the laws of rationality, will make a small world choice that is rational when viewed from either the grand world or small world perspective. This is the problem of small worlds.* (77, emphasis added)

Having reviewed the key passages in Joyce (1999) that will frame our understanding of the Problem of Small Worlds, our next step is formalization.

2.2 Modeling Realistic Decision Situations

In what follows I will sometimes refer to the decision situations in $\mathbb{D}$, as defined in the previous chapter, as "orthodox" decision situations. This is just to keep straight the difference between them and a new class of entities about to be introduced. Let us turn now to the formal model of small and grand world decision situations.

2.2.1 Small World Decision Situations

We start with small world decision situations. A small world decision situation is a structure analogous to an orthodox decision situation; it is an ordered triple comprising two functions from propositions to real numbers, called *estimate functions*, together with a representation base:

$$(\mathbf{c}, \mathbf{u}, (\mathcal{O}, \mathcal{S}, \mathcal{X})) \qquad \textbf{(Small World Decision Situation)}$$

The set of all small world decision situations is $\mathbb{S}$. To show that a variable ranges over $\mathbb{S}$ we give it a hat: $\hat{s}$. To refer to the elements in a specific small world situation we can use subscripts (e.g., $\mathbf{c}_{\hat{s}}, \mathbf{u}_{\hat{s}}, \mathcal{O}_{\hat{s}}, \mathcal{S}_{\hat{s}}, \mathcal{X}_{\hat{s}}$). Sometimes it will make formalism clearer if we apply the hat to variables as a way of indicating that they range over small world partitions: $\hat{O}$, $\hat{S}_1$, etc. It will also be a useful bit of shorthand — which I will use a lot — to treat the decision situation itself as containing the elements of its partitions, like so: $O_1 \in \hat{s}$, $S \in \hat{s}$, etc.

Formally, what are these small world "estimate functions"? Estimate functions are real-valued functions defined on sets of propositions, with one catch: the values of credence estimate functions are limited to the unit interval whereas utility estimate functions are not. That is, the set of utility estimate functions is $\mathfrak{U} = \{X \to \mathbb{R} : X \subseteq \mathcal{P}(\Omega)\}$ and the set of credence estimate functions is $\mathfrak{C} = \{X \to [0,1] : X \subseteq \mathcal{P}(\Omega)\}$. Hence, $\mathfrak{C} \subset \mathfrak{U}$. We limit credence estimate functions to the unit interval to normalize the numeric representation of confidence levels, which are intuitively doubly bounded; an agent can estimate that her grand world self will have no (=zero) confidence in a proposition or will have complete (=one) confidence, or any amount in between, but it makes no sense to suppose she estimates her grand world self will have less than no confidence or more than complete confidence.

Given a representation base $X \in \mathbb{B}$ we say an estimate function with domain $\Gamma \subseteq \mathcal{P}(\Omega)$ is an estimate function *for* X iff the union of X is a subset of Γ (i.e., iff $(\mathcal{O}_X \cup \mathcal{S}_X \cup \mathcal{X}_X) \subseteq \Gamma \subseteq \mathcal{P}(\Omega)$). A small world decision situation is just any triple where the third member is a base, the first is a credence estimate function for that representation base, and the second is a utility estimate function for that base. The set $\mathbb{S}$ of small world decision situations is thus the set of all triples $(\mathbf{c}, \mathbf{u}, X)$

where $X \in \mathbb{B}$ is a representation base, $\mathbf{c}$ is a credence estimate function for X, and $\mathbf{u}$ is a utility estimate function for X. $\mathbb{D}$ is thus a proper subset of $\mathbb{S}$.

2.2.2 Grand World Decision Situations

So much for small world situations, how about the grand world? Formally, grand world decision situations are just orthodox decision situations. Intuitively, we are to understand a grand world decision situation as one that need not be "refined further" in the sense Joyce had in mind. The set of all grand world decision situations is $\mathbb{G}$. To indicate that a variable ranges over $\mathbb{G}$ we dot it: $\dot{s}$. We will use subscripts in the familiar way (e.g., $u_{\dot{s}}, \mathcal{X}_{\dot{s}}$, etc.). We also use the other conventions mentioned in the previous section, viz., $\dot{O}$, $\dot{S}_3$, $O \in \dot{s}$, $\dot{S}_2 \in \dot{s}_4$, etc.

Since grand world situations are formally just orthodox situations, the question arises: is the converse also true? No. Unlike orthodox decision situations generally, every outcome in a given grand world decision situation $\dot{s}$ entails a specific utility level relative to the utility function $u_{\dot{s}}$ i.e., the utility function in that grand world situation. That is, for any $\dot{s} \in \mathbb{G}$, if $X \in \mathcal{X}_{\dot{s}}$ then there is some n such that X entails $u_{\dot{s}}[n]$.[5] This corresponds to the assumption that all details relevant to the agent's desires are determined by grand world acts and states; there are no possibilities relevant to what the agent cares about that have not already been taken into account, so to speak, by the partitions of acts and states. So, while an agent might in principle refine a grand world outcome space, doing so would be deliberatively pointless.

2.2.3 Idealization

The heart of the Problem of Small Worlds is the deliberative relation between the small world and the grand world. What is this relation? We cannot say that every small world decision situation determines a unique grand world decision situation. A single small world decision situation could be "occupied" by countless possible agents. Suppose I'm deliberating about whether to have spaghetti or pizza. I represent my options and the possible states of the world at some level of detail or other and I make my credence and utility estimates. Joyce's idea is that I, in that state, determine a particular grand world decision situation — the situation, roughly speaking, that I would be in if I

[5] I'm using the notation for utility-level propositions introduced in §1.6.

were to deliberate ideally to a point beyond which further deliberation is superfluous. But surely this rest state of ideal deliberation for me depends not just on my current option-state representation and estimates, but on deeper psychological facts, including but not obviously limited to my current credences and desires. What's more, those deeper facts could have been different in many ways despite my being in exactly the same small world situation, with the same options, states, and estimates. Should we therefore try to bake these deeper facts into our model? No. That would only add a bunch of extra formal doodads whose only theoretical job would be determining grand world situations.[6] We don't care how the grand world situations are determined; we only need to model the fact that they are determined by occupants of small world situations, and that there are many possible occupants for any such situation. Here is how we can do this. We posit a function $\mathbf{I} : \mathbb{S} \to \mathcal{P}(\mathbb{G})$ and say that for any small world decision situation $\hat{s} \in \mathbb{S}$, the set $\mathbf{I}(\hat{s})$ is the set of grand world situations that are the *possible idealizations* of $\hat{s}$, or just *idealizations* of $\hat{s}$, for short. What this means intuitively, if it isn't clear, is just that for each idealization $\dot{s}$ of some $\hat{s}$, there is some possible agent who "occupies" $\hat{s}$ and whose psychology (or whatever) makes it such that if they deliberated ideally they'd wind up in $\dot{s}$.

2.2.4 Recoverability

Implicit in Joyce's discussion is the idea that refinement, *a fortiori* refinement that reaches the grand world, if it alters the representation base, strictly increases informational content. To model this, we will assume the representation base in any small world decision situation $\hat{s}$ is *recoverable* from the representation base of each grand world idealization of $\hat{s}$. What this means, in a nutshell, is that for any grand world decision situation $\dot{s}$ in the set $\mathbf{I}(\hat{s})$ of grand world idealizations of $\hat{s}$, outcomes in the representational base of $\hat{s}$ are disjunctions of outcomes in the representational base of $\dot{s}$.[7] That is,

Assumption 2 (Recoverability)**.** *For any $\hat{s}$, $\bar{Y} \in \mathcal{X}_{\hat{s}}$, and $\dot{s} \in \mathbf{I}(\hat{s})$ there is $X \subseteq \mathcal{X}_{\dot{s}}$ such that $\bar{Y} = \bigcup X$.*

[6] Additionally, it is not at all obvious what should be included beyond just credences and utilities.

[7] Joyce ignores the possibility that this might fail, and I think not without reason. It is hard to understand what it would mean for a credence or utility estimate in a given small world decision situation to be "correct" relative to a given grand world idealization if there is no way to recover within that grand world situation's representation base the propositions for which the estimates were made.

This assumption implies that for every idealization $\dot{s}$ of a given $\hat{s}$, the options (states) of $\hat{s}$ are disjunctions of the options (states) of $\dot{s}$. That is, for any $\hat{s}$ and $\dot{s} \in \mathbf{I}(\hat{s})$, if $O \in \mathcal{O}_{\hat{s}}$ and $S \in \mathcal{S}_{\hat{s}}$, then there are $X \subseteq \mathcal{O}_{\dot{s}}$ and $Y \subseteq \mathcal{S}_{\dot{s}}$ such that $O = \bigcup X$ and $S = \bigcup Y$.

2.2.5 Plenitude

So far we have a model of small world decision situations, grand world decision situations, and a mapping from the former to sets of the latter constrained by Recoverability. Now we will make some assumptions about the plenitudinousness of the set of small worlds that idealize to a given grand world.

First, some more handy notation. For any $\dot{s}$, let the set of small world options recoverable from $\dot{s}$ be $\mathbf{SWO}(\dot{s}) = \{P : \exists \Gamma \subset \mathcal{O}_{\dot{s}} \text{ s.t. } P = \bigvee \Gamma\}$, let the set of small world states recoverable from $\dot{s}$ be $\mathbf{SWS}(\dot{s}) = \{P : \exists \Gamma \subset \mathcal{S}_{\dot{s}} \text{ s.t. } P = \bigvee \Gamma\}$ and let the set of small world outcomes recoverable from $\dot{s}$ be $\mathbf{SWX}(\dot{s}) = \{P : \exists \Gamma_1 \subset \mathcal{S}_{\dot{s}}, \exists \Gamma_2 \subset \mathcal{O}_{\dot{s}} \text{ s.t. } P = (\bigvee \Gamma_1) \wedge (\bigvee \Gamma_2)\}$. Note our assumption that *any* disjunction of a grand world's options counts as a small world option recoverable from that grand world, and ditto for grand world states. The same does not hold for grand world outcomes, though. Arbitrary disjunctions of outcomes in a grand world do not necessarily correspond to conjunctions of small world options and states recoverable from that grand world. Lastly, for a given grand world $\dot{s}$ let the set of small worlds that idealize to $\dot{s}$ be $\mathbf{SW}(\dot{s}) = \{\hat{s} : \dot{s} \in \mathbf{I}(\hat{s})\}$. Here is our plenitude assumption: for any partition $\mathcal{O}^* \subset \mathbf{SWO}(\dot{s})$, partition $\mathcal{S}^* \subset \mathbf{SWS}(\dot{s})$, and partition $\mathcal{X}^* = \{X : \exists O \in \mathcal{O}^*, \exists S \in \mathcal{S}^* \text{ s.t. } X = OS\}$, there is $\hat{s} \in \mathbf{SW}(\dot{s})$ such that $\mathcal{O}_{\hat{s}} = \mathcal{O}^*$, $\mathcal{S}_{\hat{s}} = \mathcal{S}^*$, and $\mathcal{X}_{\hat{s}} = \mathcal{X}^*$.[8] In other words, we assume that $\mathbf{SW}(\dot{s})$ is plenitudinous in the representation bases of its members. Here's the assumption officially stated:

Assumption 3 (Inverse-Idealization Base Plenitude)**.** *Any representation base recoverable from the representation base of a grand world $\dot{s}$ is possessed by some small world that idealizes to $\dot{s}$.*

Lastly, we make the following much stronger assumption: if a small world has a representation base recoverable from the base of a grand world, then that small world idealizes to that grand world.

[8] Since every member of $\mathcal{O}^*$ is compatible with every member of $\mathcal{S}^*$ and both are partitions, $\mathcal{X}^*$ is a partition.

Assumption 4 (Recovery Implies Idealization)**.** *For any $\hat{s}$ and $\dot{s}$, if for any $P \in \hat{s}$ there is $\Gamma \subseteq \dot{s}$ such that $P = \bigvee \Gamma$, then $\dot{s} \in \mathbf{I}(\hat{s})$.*

To contextualize this, two features of our model deserve notice. First, a small world only contains a representation base and a pair of estimate functions. It does not encode *anything else* about an agent's mental state beyond those features. Hence to say that a small world "idealizes to" a specific grand world $\dot{s}$ cannot imply anything very interesting about what sorts of non-ideal psychological states could or would transform into $\dot{s}$ on ideal reflection — it only implies that an agent with this representation base and who makes these estimates *can* be in a state that transforms on ideal reflection into $\dot{s}$, with its precise credence and desire states. Second, and relatedly, a given small world can intuitively idealize to very different grand worlds, inasmuch as that small world can be "instantiated" by diverse possible agents, all of whom see themselves facing the same set of options and possible states of the world.

2.3 Preliminary Concepts

We have our model of realistic decision situations in front of us. We are almost ready to state the Problem of Small Worlds. But there are still a few more preliminary concepts we need. Bear with me, dear reader.

2.3.1 Boolean Algebras and Lattices

A *Boolean algebra*, or just an *algebra* for short, is a partially ordered set such that in that set there can be found for every one of its elements the *negation* of that element, and for every pair of its elements the *join* (least upper bound) of those elements and the *meet* (greatest lower bound) of those elements relative to the order. In the notation we will use, an algebra $\mathcal{A}(P, Q, \dots)$, for any propositions $P, Q, \dots \in \mathcal{P}(\Omega)$, is the smallest set containing $P, Q, \dots$ and closed under the operations of disjunction, conjunction, and negation.[9] The *atoms* of an algebra $\mathcal{A}$ are the cells of the partition

[9] We thus take for granted that the relevant ordering relation is just set inclusion.

$\Gamma \subset \mathcal{A}$ such that for any nonempty $P \in \mathcal{A}$, there is $\Gamma^* \subseteq \Gamma$ such that $P = \bigvee \Gamma^*$.[10] Given some propositions $P, Q, R, \ldots$ let $\Gamma(P, Q, R, \ldots)$ denote the set of atoms of the algebra $\mathcal{A}(P, Q, R, \ldots)$.

In addition to algebras, we will also need a more general sort of structure known as a *Boolean lattice* or *lattice* for short. A lattice is like an algebra but without negation closure. So, a lattice is a partially ordered set that contains for any two of its members both their join and their meet. In our notation, a lattice $\mathcal{L}(P, Q, \ldots, Z)$ for any $P, Q, \ldots, Z \in \mathcal{P}(\Omega)$ is the smallest set containing P, Q, …, Z that is closed under conjunction and disjunction. Likewise, if $\Psi \subseteq \mathcal{P}(\Omega)$, $\mathcal{L}(\Psi)$ is the smallest superset of Ψ closed under conjunction and disjunction.

2.3.2 Three Set-Theoretic Operations

The following will be handy. First, if Γ is a set and P is a set, let $\Gamma/P \stackrel{\text{df}}{=} \{Q \in \Gamma : Q \cap P \neq \varnothing\}$. Note that $\Gamma/P \subseteq \Gamma$. Second, if Γ is a set and P is a set, let $\Gamma//P \stackrel{\text{df}}{=} \{P \cap Q : Q \in \Gamma\} \setminus \{\varnothing\}$. Note that $\Gamma//P \subseteq \Gamma$ if and only if $\bigvee(\Gamma/P) = P$. Third, if Γ is a set and P is a set, let $\Gamma///P \stackrel{\text{df}}{=} \Gamma//P \cup \{Q \setminus P : Q \in \Gamma\}$. Note that $\Gamma///P$ is a refinement of Γ.

2.3.3 $\mathcal{T}$-imaging Isomorphism

Now we come to the specifically decision-theory-related stuff. The first such thing we will define is an isomorphism concept that will help us specify the features of decision situations that (in whole or part) determine estimate correctness. Say a tuple $(\mathscr{R}_1, \mathscr{I}_1, \dot{s}_1)$ consisting of two sets of propositions and a grand world situation is *$\mathcal{T}$-imaging isomorphic under* a bijection $f : \mathscr{R}_1 \cup \mathscr{I}_1 \to \mathscr{R}_2 \cup \mathscr{I}_2$ to the tuple $(\mathscr{R}_2, \mathscr{I}_2, \dot{s}_2)$ iff

1. $\mathcal{I}^{\mathcal{T}}_{\dot{s}_1}(P \| \bigvee \Psi) = \mathcal{I}^{\mathcal{T}}_{\dot{s}_2}(f(P) \| \bigvee \{f(Q) : Q \in \Psi\})$ for any $P \in \mathscr{R}_1$ and $\Psi \subseteq \mathscr{I}_2$.
2. $\mathcal{I}^{\mathcal{T}}_{\dot{s}_1}(P \| \bigwedge \Psi) = \mathcal{I}^{\mathcal{T}}_{\dot{s}_2}(f(P) \| \bigwedge \{f(Q) : Q \in \Psi\})$ for any $P \in \mathscr{R}_1$ and $\Psi \subseteq \mathscr{I}_2$.[11]
3. $\mathcal{I}^{\mathcal{T}}_{\dot{s}_1}(P \| \top) = \mathcal{I}^{\mathcal{T}}_{\dot{s}_2}(f(P) \| \top)$, and hence $c_{\dot{s}_1}(P) = c_{\dot{s}_2}(f(P))$, for any $P \in \mathscr{R}_1$.

[10] The only algebras we'll be interested in will be based on the set of possible worlds, which we assume is finite.

[11] Here for simplicity let us adopt the convention that $A = B$ if A and B are both undefined and $A \neq B$ if one but not the other is undefined.

Say a tuple $(\Gamma_1, \dot{s}_1)$ is $\mathcal{T}$-imaging isomorphic under a bijection $f : \Gamma_1 \to \Gamma_2$ to the tuple $(\Gamma_2, \dot{s}_2)$ iff $(\Gamma_1, \Gamma_1, \dot{s}_1)$ is $\mathcal{T}$-imaging isomorphic under f to $(\Gamma_2, \Gamma_2, \dot{s}_2)$. Generalizing further, say for $\mathbb{G}^*, \mathbb{G}^{**} \subseteq \mathbb{G}$ that the tuple $(\mathscr{R}_1, \mathscr{I}_1, \mathbb{G}^*)$ is *$\mathcal{T}$-imaging isomorphic under* $f : \mathscr{R}_1 \cup \mathscr{I}_2 \to \mathscr{R}_2 \cup \mathscr{I}_2$ to $(\mathscr{R}_2, \mathscr{I}_2, \mathbb{G}^{**})$ iff for every $\dot{s} \in \mathbb{G}^*, \dot{x} \in \mathbb{G}^{**}$, $(\mathscr{R}_1, \mathscr{I}_1, \dot{s})$ is $\mathcal{T}$-imaging isomorphic under f to $(\mathscr{R}_2, \mathscr{I}_2, \dot{x})$.

Let me state what $\mathcal{T}$-imaging isomorphism amounts to (slightly!) less formally. First, note that the relation is an equivalence relation (symmetric, transitive, and reflexive). It encodes a tight similarity in how credence in grand worlds moves around under $\mathcal{T}$-imaging. Within any tuple in the field of the relation there are two sets of propositions and a grand world. We can call the first set $\mathscr{R}$ the recipient set and the second set $\mathscr{I}$ the imaging set. So, the first and I suppose the most obvious thing that $\mathcal{T}$-imaging isomorphism calls for when it holds between two tuples is equinumerosity between the union of the recipient and imaging sets of the one and the union of the recipient and imaging sets of the other. The lingo will be that we take the $\mathcal{T}$-image *of* propositions in the recipient set *in* the given grand world *on* conjunctions or disjunctions of subsets of propositions in the imaging set (or *on* the necessary proposition $\top$). What a tuple x being $\mathcal{T}$-imaging isomorphic under a bijection f to a tuple y really means is just that if we take the $\mathcal{T}$-image of some $P \in \mathscr{R}_x$ in s_x on $Q \in \mathcal{L}(\mathscr{I}_x \cup \{\top\})$ the value will be equal to that of taking the $\mathcal{T}$-image of $f(P) \in \mathscr{R}_y$ in s_y on $f(Q) \in \mathcal{L}(\mathscr{I}_y \cup \{\top\})$. I will sometimes abbreviate $\mathcal{T}$-imaging isomorphic(/ism) as just '$\mathcal{T}$-imaging-iso'.

2.3.4 Utility Isomorphism

Here we have another tool to use to describe the features of situations that determine estimate correctness. Say a tuple $(\Gamma_1, \dot{s}_1)$ of a set of propositions and a grand world situation is *utility isomorphic under* a bijection $f : \Gamma_1 \to \Gamma_2$ to a tuple $(\Gamma_2, \dot{s}_2)$ iff for every $P \in \Gamma_1$ there is some $n \in \mathbb{R}$ such that $P \subseteq u_{\dot{s}_1}[n]$ and $f(P) \subseteq u_{\dot{s}_2}[n]$. (Remember, for any utility function u, $u[n]$ is the u-utility-level proposition that holds at a world w iff $u(w) = n$.) I will sometimes abbreviate utility isomorphic(/ism) as just 'utility-iso'.

2.3.5 Estimated Expected Utility

We will need a concept of *estimated expected utility* for small worlds. The assumption will be that for any $\hat{s}$ and $O \in \mathcal{O}_{\hat{s}}$ the estimated expected utility of O in $\hat{s}$, $\mathbf{EstEU}_{\hat{s}}(O)$, is the sum $\sum_{S \in \mathcal{S}_{\hat{s}}} \mathbf{c}_{\hat{s}}(O \wedge S) \times \mathbf{u}_{\hat{s}}(O \wedge S)$, that is, a sum of products, one product for each outcome of $\hat{s}$ compatible with O, and the terms in the product for a given outcome being the credence and utility estimates for that outcome.

2.3.6 Utility Coarsenings

We will need to relate collections of propositions to one another based on patterns of sameness in their grand world utility levels. Here's the tool we'll use for this. If Γ_1 is a partition and P is some proposition, Γ_2 is an *improper utility coarsening* of Γ_1 *on* P *in* $\dot{s}$, which we write $\Gamma_2 \sqsubseteq_{\dot{s},P} \Gamma_1$, iff for every $R \in \Gamma_2/P$ there is (1) some $n \in \mathbb{R}$ such that $R \subseteq u_s[n]$ and (2) a subset $\Gamma^* \subseteq \Gamma_1/P$ such that (i) $Q \subseteq u_s[n]$ for every $Q \in \Gamma^*$ and (ii) $\bigvee \Gamma^* = R$. Note that for any $\dot{s} \in \mathbb{G}$ and P, $\sqsubseteq_{\dot{s},P}$ is a partial order on the set of refinements of $\mathcal{X}_{\dot{s}}$.[12]

2.3.7 Utility Frames

The final preliminary concept will be what we use to model how different theories use credence information differently when assessing estimates for correctness. We will consider a class $\mathcal{U}$ of triples consisting of a grand world decision situation, a proposition P, and a set of disjoint propositions whose disjunction is P. Define $\mathcal{U}$ as follows: $(\dot{s}, P, \Gamma) \in \mathcal{U}$ iff for every $X \in \mathcal{X}_{\dot{s}}$ compatible with P, $\bigvee(\Gamma/X) = X$. In other words, $(\dot{s}, P, \Gamma) \in \mathcal{U}$ iff every $X \in \mathcal{X}_{\dot{s}}$ compatible with P is a disjunction of a subset of Γ. Hence, intuitively speaking, for any $(\dot{s}, P, \Gamma) \in \mathcal{U}$ the set Γ is a *refinement* of $\mathcal{X}_{\dot{s}}/P$. We will use a system of set functors based on '$\mathcal{U}$': e.g., $\mathcal{U}_{\dot{s}} = \{(P, \Gamma) : (\dot{s}, P, \Gamma) \in \mathcal{U}\}$ $\mathcal{U}_{\dot{s}}(P') = \{\Gamma : (\dot{s}, P', \Gamma) \in \mathcal{U}\}$, etc. It will be useful to have variables ranging over sets of disjoint propositions based on the functors but with numerical and symbolic superscripts: e.g., $\ulcorner \mathcal{U}^{12}_{\dot{s}}(P') \urcorner$ and $\ulcorner \mathcal{U}^{**}_{\dot{s}}(P') \urcorner$ range over $\mathcal{U}_{\dot{s}}(P')$, etc. A function taking any $\dot{s}$ and $S \in \mathbf{SWS}(\dot{s})$ to a set $\mathcal{U}^*_{\dot{s}}(S)$ is a *utility frame*. The utility frame for theory $\mathcal{T}$ is designated $\ulcorner \mathcal{U}^{\mathcal{T}} \urcorner$, and the set it assigns to $\dot{s}$ and

[12] For any $\dot{s} \in \mathbb{G}$, any proposition P, and any three refinements of $\mathcal{X}_{\dot{s}}$, $\Gamma, \Gamma^*, \Gamma^{**} \in \mathbb{P}(\Omega)$, (1) $\Gamma \sqsubseteq_{\dot{s},P} \Gamma$, (2) if $\Gamma \sqsubseteq_{\dot{s},P} \Gamma^*$ and $\Gamma \neq \Gamma^*$ then $\Gamma^* \not\sqsubseteq_{\dot{s},P} \Gamma$, (3) if $\Gamma \sqsubseteq_{\dot{s},P} \Gamma^*$ and $\Gamma^* \sqsubseteq_{\dot{s},P} \Gamma^{**}$ then $\Gamma \sqsubseteq_{\dot{s},P} \Gamma^{**}$.

$S \in \mathbf{SWS}(\dot{s})$ is $\mathcal{U}_{\dot{s}}^{\mathcal{T}}(S)$. So then what is a utility frame? A utility frame is just a function that a decision theory gives us such that when that function is given a grand world $\dot{s}$ and small world state S (recoverable from the states in $\dot{s}$), it hands back a set of disjoint propositions whose union is S such that any outcome in $\dot{s}$ compatible with S is a disjunction of propositions in the set delivered by the utility frame.

2.4 Presentation

With the preliminary concepts out of the way, we can now present the Problem of Small Worlds. The first major novelty in our approach compared to what can be found in Joyce's book will be that we will distinguish two different levels of constraint — levels of difficulty, if you like — at which the Problem Small Worlds can be posed. The Problem of Small Worlds, as we will understand it, has both *restricted* and *unrestricted* forms.

Suppose we take a representation base at a given grand world and derive a coarsened small world representation base from it. We can do this in one of three ways. We can take disjunctions of grand world states, or of grand world options, or both.[13] The following diagram depicts the different ways grand world representations can be coarsened:

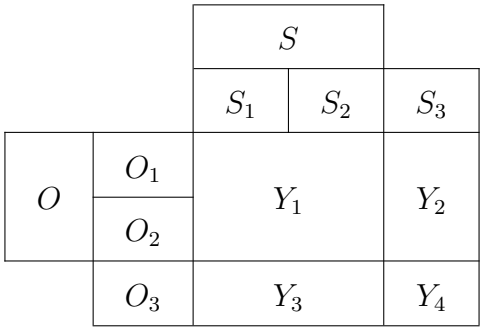

Figure 2.1: Three Ways of Coarsening a grand world Representation Base

Here we have three small world outcomes, Y_1, Y_2, and Y_3, and one grand world outcome, Y_4. The small world outcomes each combine grand world outcomes in different ways. The first, Y_1, involves disjunctions both of grand world states and of grand world options. The second, Y_2, involves a

[13] Of course, some small worlds will have the same representation base as their idealizations; but that's neither here nor there at the moment.

disjunction of grand world options; it's a disjunction of the small world option O and grand world state S_3. The third, Y_3, involves a disjunction of grand world states; it's a disjunction of the grand world option O_3 and the small world state S. Lastly, the grand world outcome Y_4 is just the conjunction of O_3 and S_3. To put the main point another way, if $\dot{s} \in \mathbf{I}(\hat{s})$ is a grand world idealization of a small world $\hat{s}$, it might be the case that $\mathcal{O}_{\hat{s}} = \mathcal{O}_{\dot{s}}$ or $\mathcal{S}_{\hat{s}} = \mathcal{S}_{\dot{s}}$ or both or neither.[14] Now, in light of this, as mentioned, the job of solving the Problem of Small Worlds comes at two separate levels of difficulty. At the first and much easier level, we solve it for pairs of small world situations and idealizations whose representation bases contain the same set of options — that is, we solve it for the case where idealizing to the grand world does not alter the agent's option representation, but at most refines her representation of the possible states of the world (and only thereby refines her representation of the possible outcomes). I call this solving the *Restricted* Problem of Small Worlds. At the second and more difficult level, we solve the problem for cases where going to the grand world refines one or more of the agent's options (regardless of whether states are also refined). I call this solving the *Unrestricted* Problem of Small Worlds. Now, it turns out to be easiest to state the two different forms of the problem in reverse order of difficulty. So we begin with the Unrestricted Problem. Keep in mind that you're about to see a schema; the $\mathcal{T}$ needs to be filled in with some specific decision theory. Technically, therefore, each decision theory faces *its* version of the Problem of Small Worlds.

2.4.1 The Unrestricted Problem of Small Worlds, First Formulation

Prove there are *correctness functions* $\mathbf{C}^{\mathcal{T}} : \dot{s} \mapsto \big(\mathbf{SWO}(\dot{s}) \times \mathbf{SWS}(\dot{s}) \to [0,1]\big)$, $\mathbf{U}^{\mathcal{T}} : \dot{s} \mapsto \big(\mathbf{SWO}(\dot{s}) \times \mathbf{SWS}(\dot{s}) \to \mathbb{R}\big)$, and a utility frame $\mathcal{U}^{\mathcal{T}} : \dot{s} \mapsto \{\Gamma \subset \mathcal{P}(\Omega) : \bigwedge\Gamma = \bot, \bigvee\Gamma \in \mathbf{SWX}(\dot{s})\}$ such that:

(PSW1) For any $\hat{s}_1, \hat{s}_2$, $O, S \in \hat{s}_1$, $O', S' \in \hat{s}_2$, and for any $\dot{s}_1, \dot{s}_2$, if there is a bijection f such that

(a) $f(OS) = O'S'$

(b) $f(O\overline{S}) = O'\overline{S'}$

[14] In the "both" case, the situations are still distinct because their ingredient functions will differ.

(c) $f(\overline{O}S) = \overline{O'}S'$

(d) $f(\overline{O}\,\overline{S}) = \overline{O'}\,\overline{S'}$

and $(\{OS, O\bar{S}, \bar{O}S, \bar{O}\bar{S}\}, \dot{s}_1)$ is $\mathcal{T}$-imaging-iso under f to $(\{O'S', \overline{O'}S', O'\overline{S'}, \overline{O'}\,\overline{S'}\}, \dot{s}_2)$, $\mathbf{C}^{\mathcal{T},O}_{\dot{s}_1}(S) = \mathbf{C}^{\mathcal{T},O'}_{\dot{s}_2}(S')$.

(PSW2) For any $\hat{s}_1, \hat{s}_2, O, S \in \hat{s}_1, O', S' \in \hat{s}_2$, if $\exists \mathcal{U}^{*}_{\dot{s}_1}(S) \sqsubseteq_{\dot{s}_1,O} \mathcal{U}^{\mathcal{T}}_{\dot{s}_1}(S)$, and $\exists \mathcal{U}^{**}_{\dot{s}_2}(S') \sqsubseteq_{\dot{s}_2,O'} \mathcal{U}^{\mathcal{T}}_{\dot{s}_2}(S')$, and for some bijection $f : \mathcal{U}^{*}_{\dot{s}_1}(S)///O \cup \mathcal{L}(O,S) \to \mathcal{U}^{**}_{\dot{s}_2}(S')///O' \cup \mathcal{L}(O',S')$ such that

(a) $f(OS) = O'S'$

(b) $f(O) = O'$

(c) $f(S) = S'$

(d) $f(O \vee S) = O' \vee S'$,

both (i) $(\mathcal{U}^{*}_{\dot{s}_1}(S)///O, \{O,S\}, \dot{s}_1)$ is $\mathcal{T}$-imaging-iso under f to $(\mathcal{U}^{**}_{\dot{s}_2}(S')///O', \{O',S'\}, \dot{s}_2)$ and (ii) $(\mathcal{U}^{*}_{\dot{s}_1}(S)//O, \dot{s}_1)$ is utility-iso under f to $(\mathcal{U}^{**}_{\dot{s}_2}(S')//O', \dot{s}_2)$, then $\mathbf{U}^{\mathcal{T},O}_{\dot{s}_1}(S) = \mathbf{U}^{\mathcal{T},O'}_{\dot{s}_2}(S')$.

(PSW3) For any $\hat{s}$ and $\dot{s} \in \mathbf{I}(\hat{s})$, if $\mathbf{c}_{\hat{s}}$ and $\mathbf{u}_{\hat{s}}$ are such that for all $O, S \in \hat{s}$, we have $\mathbf{c}_{\hat{s}}(OS) = \mathbf{C}^{\mathcal{T},O}_{\dot{s}}(S)$ and $\mathbf{u}_{\hat{s}}(OS) = \mathbf{U}^{\mathcal{T},O}_{\dot{s}}(S)$, then any small world option $O \in \mathcal{O}_{\hat{s}}$ that maximizes $\mathbf{EstEU}_{\hat{s}}$ is *compatible with* a grand world option $O \in \mathcal{O}_{\dot{s}}$ that maximizes $\mathrm{EU}^{\mathcal{T}}_{\dot{s}}$.

2.4.2 The Restricted Problem of Small Worlds

Prove for your theory $\mathcal{T}$ that there are *correctness functions* $\mathbf{C} : \dot{s} \mapsto \big[O \in \mathcal{O}_{\dot{s}} \mapsto \big(S \in \mathbf{SWS}(\dot{s}) \mapsto \mathbf{C}^{O}_{\dot{s}}(S)\big)\big]$ and $\mathbf{U} : \dot{s} \mapsto \big[O \in \mathcal{O}_{\dot{s}} \mapsto \big(S \in \mathbf{SWS}(\dot{s}) \mapsto \mathbf{U}^{O}_{\dot{s}}(O \wedge S)\big)\big]$, and a utility frame $\mathcal{U}^{\mathcal{T}}$ such that:

(PSW1) Same as before.

(PSW2) Same as before.

(PSW3*) For any $\hat{s}$ and $\dot{s} \in \mathbf{I}(\hat{s})$ *such that* $\mathcal{O}_{\hat{s}} = \mathcal{O}_{\dot{s}}$, if for all $O \in \mathcal{O}_{\hat{s}}, S \in \mathcal{S}_{\hat{s}}$, we have $\mathbf{c}_{\hat{s}}(O \wedge S) = \mathbf{C}^{\mathcal{T},O}_{\dot{s}}(S)$ and $\mathbf{u}_{\hat{s}}(O \wedge S) = \mathbf{U}^{\mathcal{T},O}_{\dot{s}}(S)$, then any small world option $O \in \mathcal{O}_{\hat{s}}$ that maximizes $\mathbf{EstEU}_{\hat{\mathbf{s}}}$ *is identical to* a grand world option $O \in \mathcal{O}_{\dot{s}}$ that maximizes $\mathrm{EU}^{\mathcal{T}}_{\dot{s}}$.

2.4.3 Weak and Strong

We thus have these two different versions of the problem of small worlds, the Restricted Problem and the Unrestricted Problem. Each consists of three conditions which are satisfied by providing an estimate correctness functions and a utility frame in addition of course to a definition of expected utility in the grand world. We're going to make a further subdivision, yielding four total divisions of the problem. This further subdivision will be based on which of the three conditions are solved. We're going to call the *Weak* version of the problem the version that you solve just by satisfying the third condition, which is the condition that requires that correct estimates be such that they guarantee the appropriate linkage between correctly estimated expected utility and grand world rationality. The *Strong* version of the problem will be the Weak version plus satisfaction of the first two conditions, which pertain not only to the link between the small world and the grand world, but to the fact that the estimate functions provided are suitably "atomic" in nature. By this I mean, the functions are sensitive only to the relevant isolated portion of the grand world associated with the small world outcome for which the estimated is being made. In summary we have four problems, the Weak Restricted Problem the Strong Restricted Problem the Weak Unrestricted Problem and the Strong Unrestricted Problem.

2.5 Estimate Correctness

Joyce stipulates that to solve the Problem of Small Worlds decision theories must give an account of what it takes for estimates of grand world attitudes, made in the small world, to be correct. This corresponds to the requirement in our version of the problem to provide estimate correctness functions. The idea with estimate correctness functions is to encapsulate correctness conditions in a way similar to a truth function in possible world semantics. As truth functions determine truth values relative to worlds, estimate correctness functions determine correct estimate values for small world option-state pairs relative to grand worlds.[15] Now, the formal definition of a correctness function does not require that correctness values be determined by any specific aspect of grand worlds. It just dictates that for each small world option-state pair and each grand world

[15]Not just any small world option-state pair; they must be recoverable from the grand world representation base.

there has to be some correct estimate value. But as correctness is to be understood as part of the Problem of Small Worlds, it seems that we can't be so liberal as to leave it at that. Otherwise, it might turn out that in order to settle whether a particular estimate is correct, all the information contained in the grand world might be called upon, and this seems unintuitive in that when an agent is estimating their grand world attitude toward a small world outcome, the information in the grand world that fixes whether their estimate is correct should be information associated in some sense by their grand world self with the isolated portion of logical space that they are focused on. That is, the information in the grand world that determines the correctness value for a given small world option-state pair should be confined in an intuitively appropriate sense to parts of logical space carved out by the state and option the agent has in mind. Ensuring that one's estimate correctness functions meet these constraints is something we intuitively regard as part of solving the Problem of Small Worlds. Hence these constraints, specified using the isomorphism concepts we introduced, constitute the first two parts of our version of the problem, i.e., PSW1 and PSW2.

We assume thus that in order to solve the Strong Problem Small Worlds (Restricted or Unrestricted) one must not only show that on one's theory performing an act in the small world that maximizes correctly estimated expected utility connects in the right way to rational grand world choice. One must also show that one's characterizations of correctness are "local" in the right way. But what is this concept of "locality"? How do our specifications of PSW1 and PSW2 capture it? Let's look at each condition briefly.

2.5.1 Explicating PSW1

The idea behind PSW1 is that when one is estimating the credence value for a given small world option-state pair relative to a given grand world, all that's intuitively necessary to determine whether one's estimate is correct or incorrect should be recoverable from certain facts about how credence is *imaged* in the grand world. Now, each decision theory supplies its own imaging function; so, PSW1 does not "care," so to speak, exactly what a given decision theory says it means to image credence from one proposition into another. What it requires is that the correctness relative to a given grand world of a credence estimate for a given small world option-state pair must be determinanble if we know how credence in that grand world images — however imaging happens to be defined — into

the small world outcome on that small world state and on that small world option, and if we know how what the prior credences in the state, option, and outcome are in the grand world, and for good measure, if we know how credence images into Boolean combinations of the state and outcome on such combinations. We think this is so because we assume, in essence, that what one is doing when one makes a credence estimate is making a practical supposition of some kind, and imaging is the formal concept that explicates such supposition. Again, we don't dictate what exactly the content of the practical supposition has to be, we leave that to decision theory to decide; we just assume that it has to be limited to propositions in the Boolean algebra determined by the small world option and state — no other propositions should be relevant.[16] PSW1 thus requires that any two grand worlds where credence "images around" in an identical way in corresponding algebras based on state-pairs must have the same correctness value for a credence estimate for the corresponding state-option pairs.

2.5.2 Explicating PSW2

PSW2 imposes a similar if slightly more complex constraint — more complex simply because we're dealing with utility as well as credence. The basic intuition is that whether a utility estimate for a given option-state pair is correct in a grand world should depend on features of the utilities in the grand world at worlds compatible with the small world option-state conjunction — the small world outcome — in question. Utilities of worlds incompatible with the state or with the option that the agent is considering should not be determinative with respect to whether their estimate is correct. Now just as with credence estimation, it seems that when an agent is making a utility estimate, there's also a practical supposition taking place, and so imaging should be involved in determining estimate correctness once again. But since what is at issue is not a grand world credence value for the outcome as a whole (or the state or the option), but instead something more fine-grained — roughly, how credence is imaged into practically relevant logical parts of the relevant state or small world outcome — we need to determine how to divide a given small world state

[16] It's also important to bear in mind that we assume that prior credence represents a kind of null supposition; in other words, a decision theory could just say, for instance, that a credence estimate for a small world state-option pair is correct in a grand world if it equals the credence assigned in that grand world to the state — hence this unconditional credence is one of the features preserved by the isomorphism specified in PSW1.

into logical parts such that how credence images into those parts is relevant to determining the correctness of utility estimate. Two thoughts suggest themselves, namely, we could just use the partition of the small world state consisting of the set of grand world outcomes compatible with it, or else, the partition of the state into worlds (i.e., singletons). But neither of these ultimately proves adequate. The former runs into the problem that you could have two different grand worlds from which the same small world state is recoverable, but where the respective sets of grand world outcomes are not equinumerous — such differences are in part what differences in grand worlds amount to. The latter maximally fine-grained partitioning avoids this problem, but is too "one-size-fits-all." It may be that according to some decision theory utility estimation pertains to certain submaximal possibilities compatible with the small world state that are both more fine grained than the partition by grand world outcomes and less fine grained than worlds. We need some kind of compromise between these two first try ideas, and the happy medium is to allow each decision theory to determine its own "utility frame," which (as explained in §2.3.7) is a function that provides a tailored partition for a proposition relative to each grand world that is at least as refined as the partition of that proposition into grand world outcomes, and possibly coarser than the partition of that proposition into worlds. One might reasonably ask, how can this be a happy medium, if it just runs into the same problem that the partitions of grand world outcomes ran into: couldn't it happen that the utility frame for the same small world state could differ from one grand world to another, despite the fact that we want to say that the estimate for that state (given a small world option) should be the same based on the underlying utility facts? Indeed, this seems conceivable. For this reason we have our improper utility coarsening relation. The idea with this is as follows. The difference between the utility frame partition of the same small world state in two different grand worlds will be superficial from the point of view of the correctness of a utility estimate made for that small world state together with some small world outcome if we can recover the same sort of utility-cum-imaging structure in each of the two grand worlds for that part of logical space. What I mean by this is that, if we can conjoin cells in the utility frame partition of that small world state in one of the two grand worlds, and do so based on their having the same utility level in that grand world, and we can map members of the resulting coarsened partition one-to-one into the cells of the utility frame partition for the other grand world for the same small world state, or

into another set of thusly coarsened cells in another grand world for the same small world state, and this mapping preserves utility levels and preserves the pattern of imaging, then it would seem that whatever differences there are between the utility frame partitions of that small world outcome in those two grand worlds must be superficial, and therefore the correct utility estimate for that outcome should be the same in those two grand worlds. Now, this description remains a bit rough. There is an important final nuance. We are interested in the utilities of worlds compatible with the particular small world *outcome* not necessarily worlds that are compatible only with the state (or with the option); but utility frames are partitions of states.[17] So what we in fact do is conjoin cells in the utility frame partition of the state as described, but then use our operator '//' to obtain a cross-section containing the maximal logical pieces of that partition that *entail the outcome*, and preservation of utility levels of *these* is what intuitively counts. In summary, the correctness of a utility estimate for a small world state-option pair relative to a given grand world idealization should be determined by a pattern of utility levels in elements derived (in the way described) from a utility frame partition of the relevant small world state together with an accompanying pattern of imaging into those elements. This is how we explicate the intuitive idea that the only utility information that should be relevant to determining the correctness of the utility estimate is information "local" to the small world outcome that the agent has before their mind.

2.6 Connecting Small and Grand Worlds

The Problem of Small Worlds is not just about stating when a deliberating agent's credence and utility estimates are correct, but about showing how decision theory guarantees that if an agent performs a small world option that maximizes estimated expected utility when calculated using correct credence and utility estimates, their choice will, to use Joyce's words, be rational from both the small world and the grand world perspective. What can this "guaranteed to be rational from

[17]Utility estimation also involves shifting credence into the possibilities that are in the agent's view, and while we do insist that the only utility information that should be relevant to the correctness of a utility estimate is utility information local to the pertinent small world outcome, we do not need to insist that the only credence information (or imaging information) that is relevant is information that is local to that small world outcome. We will permit, therefore, decision theories to make use of imaging information pertaining to possibilities within the state that are consistent with other options than the one the agent has in view.

both perspectives" idea amount to? The most natural way to understand this is to say that a theorist must prove that any time a small world option maximizes correctly estimated expected utility, it will entail a grand world option that maximizes expected utility. But it is easy to see on a little reflection that while this is fine in the Restricted Problem — where options stay the same when one idealizes — it just will not work in the Unrestricted Problem, insofar as as any time a small world option that maximizes correctly estimated EU is compatible with more than one grand world option, it will fail to imply whichever one maximizes EU — assuming not all of them do. What then is the next most natural way of capturing Joyce's idea? My answer to this question is expressed by PSW3: it must be shown that any small world option that maximizes correctly estimated expected utility relative to a grand world is implied by an option of that grand world that maximizes expected utility.[18]

[18]In conversation, Jim affirmed that this was indeed more or less what he had in mind. We will look into the intuitive motivation for this proposal more carefully in Ch. §5.

CHAPTER 3

THE RESTRICTED PROBLEM

3.1 Introduction

In his book, Joyce states that EDT is "the only [then extant] theory that does not fall victim to the problem of small worlds" (Joyce, 1999, 113). In §1.15 we got acquainted with the concept of partition invariant decision theories, and we observed in a quick and general way that Evidential Decision Theory (EDT) is partition invariant. According to Joyce, this feature enables EDT to solve the Problem of Small Worlds:

> Partition invariance makes it possible to employ expected utility maximization in small-world decision making. Since the acts, states, and outcomes of a small-world decision problem are disjunctions of the acts, states, and outcomes of the grand-world problem, it follows from partition invariance that any assignment of probabilities and utilities [i.e., expected values] to the later entities determines a unique assignment of probabilities and utilities to the former. Moreover ... an agent who *correctly* estimates her fully considered desires and beliefs regarding the states and outcomes in [a given small-world] can be sure that the policy of maximizing [correctly estimated expected utility] in [that small-world] will not lead her astray because, no matter how she chooses to subdivide the possibilities, the use of [$\mathrm{EU}^{\mathrm{EDT}}$] will lead her to the same results that she would obtain were she facing the grand-world decision itself. In [EDT], then, the basic laws of rationality are the same for both kinds of decision making, and there is guaranteed agreement between grand- and small-world representations [...]. The partition invariance of Jeffrey's theory should thus be seen as one of its main advantages [...]. (Joyce, 1999, pp. 121-22)

In another work, Joyce seems to imply that he thinks partition invariance is not only sufficient but necessary for solving the Problem of Small Worlds: "Partition invariance, in my view, is not an optional virtue for an expected utility theory to have, since only a partition invariant theory can be applied to the kinds of 'small world' decisions that people face in real life" (Joyce, 2000, S8). Joyce's view seems not to have been that all versions of CDT extant at the time of his writing fail to solve the Problem of Small Worlds. He seems rather to leave it open whether at least some of these

theories can be reformulated in partition invariant ways. But he clearly thinks partition-dependent formulations are inadquate:

> The major shortcoming of CDT is that it only tells us how to calculate the utilities of actions relative to the "grand-world" partition [of possible worlds]. [Lewis's] formulation does not do any better since it too uses a special partition of propositions to compute efficacy values. ... [It] is unwise to formulate decision theory in this way because it makes it impossible to apply it to the kinds of "small-world" decision problems people face in real life. The only way to obtain a theory with such "small-world" applicability is by finding a partition-invariant equation for computing expected utilities. (Joyce, 2000, S8)

Is partition invariance the non-negotiable key to solving the Problem of Small Worlds? Well, as we have it presented, there are really four problems, the Restricted Problem in its Weak and Strong forms, and the Unrestricted Problem in its Weak and Strong forms. The appropriate question is therefore whether partition invariance enables solutions to any of these. I will not argue that partition invariance is necessary for solving any of these; I will, however, show that it is sufficient for some of them.

In this chapter I define a suitable concept of partition invariance within our decision situation framework (§3.2). Using this concept, I show EDT can solve the Weak and Strong Restricted Problem (§3.3). I show that Lewisian CDT conforms to this concept of partition invariance (§3.4) and that it too can thus solve the Weak and Strong Restricted Problem (§3.5). I respond to an objection to the solution I offer on behalf of the Lewisian (§3.6) and address a particular qualm Lewisians themselves may have with the solution (§3.7).

3.2 Regular Partition Invariance

Let $\mathbf{P}$ be the set of all functions taking a decision situation to a probability function in the sense given in the final paragraph of §1.5. Let $\mathbf{V}$ be the set of all functions taking a decision situation to a value function in the sense given in §1.6.

Recall the set of equivalence classes $\mathrm{PartInv}(\mathbb{D})$ given in §1.15. For a given $s \in \mathbb{D}$, let $\mathrm{PartInv}(s)$ denote the unique $\Gamma \in \mathrm{PartInv}(\mathbb{D})$ such that $s \in \Gamma$. Recall that $\mathbb{P}(\Omega)$ is the set of all partitions.

Say that a decision theory $\mathcal{T}$ is *regularly partition-invariant under* $p \in \mathbf{P}$ *and* $v \in \mathbf{V}$ iff

$$\begin{aligned} &\forall s \in \mathbb{D}, \\ &\quad \forall O \in \mathcal{O}_s, \\ &\qquad \forall s' \in \mathrm{PartInv}(s), \quad \mathrm{EU}^{\mathcal{T}}_{s}(O) = \mathrm{EU}^{\mathcal{T}}_{s'}(O) = \sum_{S \in s'} p^{O}_{s'}(S) v^{O}_{s'}(S). \end{aligned} \tag{3.1}$$

Say that a decision theory $\mathcal{T}$ is *regularly partition-invariant* iff there are $p \in \mathbf{P}$ and $v \in \mathbf{V}$ such that $\mathcal{T}$ is regularly partition invariant under p and v.[1]

Theorem 2 (Regular Partition Invariant Theories Solve the Weak Restricted POSW)**.** *If a theory* $\mathcal{T}$ *is regularly partition invariant under* $p \in \mathbf{P}$ *and* $v \in \mathbf{V}$*, then the Weak Restricted Problem of Small Worlds (First Formulation) can be solved for* $\mathcal{T}$ *with* $\dot{s}, O \in \mathbf{SWO}(\dot{s}), S \in \mathbf{SWS}(\dot{s}) \mapsto p^{O}_{\dot{s}}(S)$ *and* $\dot{s}, O \in \mathbf{SWO}(\dot{s}), S \in \mathbf{SWS}(\dot{s}) \mapsto v^{O}_{\dot{s}}(S)$*.*

Proof. Consider $\hat{s}$ such that $O^* \in \hat{s}$ and $\dot{s} \in \mathbf{I}(\hat{s})$ such that $\mathcal{O}_{\hat{s}} = \mathcal{O}_{\dot{s}}$. Note that $O^* \in \dot{s}$. Ex hypothesi there are $p \in \mathbf{P}$ and $v \in \mathbf{V}$ such that $\forall O, S \in \hat{s}$, $\mathbf{c}_{\hat{s}}(OS) = p^{O}_{\dot{s}}(S)$ and $\mathbf{u}_{\hat{s}}(OS) = v^{O}_{\dot{s}}(S)$. Hence, we have

$$\mathbf{EstEU}_{\hat{s}}(O^*) = \sum_{S \in \hat{s}} p^{O^*}_{\dot{s}}(S) v^{O^*}_{\dot{s}}(S). \tag{3.2}$$

Given the regular partition-invariance of $\mathcal{T}$ under p and v,

$$\mathbf{EstEU}_{\hat{s}}(O^*) = \sum_{S \in \hat{s}} p^{O^*}_{\dot{s}}(S) v^{O^*}_{\dot{s}}(S) = \sum_{S \in \dot{s}} p^{O^*}_{\dot{s}}(S) v^{O^*}_{\dot{s}}(S) = \mathrm{EU}^{\mathcal{T}}_{\dot{s}}(O^*). \tag{3.3}$$

Hence, therefore any option in $\mathcal{O}_{\hat{s}} = \mathcal{O}_{\dot{s}}$ that maximizes $\mathbf{EstEU}_{\hat{s}}$ must also maximize $\mathrm{EU}^{\mathcal{T}}$. This completes PSW3*. □

3.3 An EDT Solution to the Strong Restricted Problem

It's easy to show that EDT is regular partition invariant:

Lemma 1 (EDT is Regular Partition Invariant)**.** *EDT is regular partition invariant under*

[1] Note that a decision theory can be regular partition invariant under a $p \in \mathbf{P}$ and $v \in \mathbf{V}$ even if $p \neq p^{\mathcal{T}}$ and $v \neq v^{\mathcal{T}}$. Regular partition invariance thus need not be thought of as a relation involving the specific probability and value functions of a given decision theory. Cf. §1.15.

- $s \in \mathbb{D}, P, Q \in \mathcal{P}(\Omega) \mapsto c_s(Q \mid P)$, *and*
- $s \in \mathbb{D}, P, Q \in \mathcal{P}(\Omega) \mapsto \mathbf{V}_s(PQ)$.

Proof. Choose $s \in \mathbb{D}, O \in s$, and $s' \in \text{PartInv}(s)$,

$$\text{EU}_s^{\text{EDT}}(O) \stackrel{\text{df}}{=} \sum_{S \in s} c_s(S \mid O)\mathbf{V}_s(OS) = \sum_{S' \in s'} c_{s'}(S' \mid O)\mathbf{V}_{s'}(OS').$$

□

Hence, we have

Theorem 3 (EDT Solves the Weak Restricted Problem)**.** *EDT solves the Weak Restricted Problem of Small Worlds with*

- $\dot{s} \in \mathbb{G}, O \in \mathbf{SWO}(\dot{s}) \mapsto [S \in \mathbf{SWS}(\dot{s}) \mapsto c_{\dot{s}}(S \mid O)]$, *and*
- $\dot{s} \in \mathbb{G}, O \in \mathbf{SWO}(\dot{s}) \mapsto [S \in \mathbf{SWS}(\dot{s}) \mapsto \mathbf{V}_{\dot{s}}(OS)]$.

Proof. From Theorem 2 and Lemma 1. □

Now let us show that EDT solves the Strong Restricted Problem. For this we need to show that the functions with which it solves the Weak Restricted Problem can function as full-fledged "local" estimate correctness functions (cf. §2.5). That is, we need to show that these functions, together with an appropriate utility frame, satisfy PSW1 and PSW2. To do this we will need to make an assumption about EDT's imaging function (cf. §1.10). It is safe to assume that the appropriate imaging function is just the conditional credence function.[2] That is $\mathcal{I}_s^{\text{EDT}}(Q \parallel P) \stackrel{\text{df}}{=} c_s(Q \mid P)$ for all $s \in \mathbb{D}$ and $P, Q \in \mathcal{P}(\Omega)$.[3]

[2]One might question this, but I see no other ready alternative. It is hard to imagine a proponent of EDT denying that practical supposition goes via conditional probability.

[3]Except for when $P = \varnothing$, in which case the function goes undefined.

Lemma 2 (EDT Solves PSW1)**.** *EDT solves PSW1 with*

$$\mathbf{C}^{\mathrm{EDT}}_{\dot{s}} \stackrel{\mathrm{df}}{=} O \in \mathbf{SWO}(\dot{s}), S \in \mathbf{SWS}(\dot{s}) \mapsto c_{\dot{s}}(S \mid O).$$[4]

Proof. Choose $\dot{s}_1, \dot{s}_2$, $O, S \in \dot{s}_1$, $O', S' \in \dot{s}_2$, and bijection f such that (1) $f(OS) = O'S'$, $f(O\overline{S}) = O'\overline{S'}$, $f(\overline{O}S) = \overline{O'}S'$, $f(\overline{O}\,\overline{S}) = \overline{O'}\,\overline{S'}$, and (2) $(\{O\overline{S}, S\overline{O}, OS, \overline{O}\,\overline{S}\}, \dot{s}_1)$ is EDT-imaging isomorphic under f to $(\{O'\overline{S'}, S'\overline{O'}, O'S', \overline{O'}\,\overline{S'}\}, \dot{s}_2)$.

Given (2) above,

$$\mathcal{I}^{\mathrm{EDT}}_{\dot{s}_1}(S \parallel O) = \mathcal{I}^{\mathrm{EDT}}_{\dot{s}_1}(OS \vee \overline{O}S \parallel OS \vee O\overline{S}) = \mathcal{I}^{\mathrm{EDT}}_{\dot{s}_2}(O'S' \vee \overline{O'}S' \parallel O'S' \vee O'\overline{S'}) = \mathcal{I}^{\mathrm{EDT}}_{\dot{s}_2}(S' \parallel O'). \tag{3.4}$$

Now, for any $\dot{s}, O \in \mathbf{SWO}(\dot{s}), S \in \mathbf{SWS}(\dot{s})$, $\mathbf{C}^{\mathrm{EDT},O}_{\dot{s}}(S) \stackrel{\mathrm{df}}{=} c_{\dot{s}}(S \mid O)$ and $\mathcal{I}^{\mathrm{EDT}}_{\dot{s}}(S \parallel O) \stackrel{\mathrm{df}}{=} c_{\dot{s}}(S \mid O)$. Hence,

$$\mathbf{C}^{\mathrm{EDT},O}_{\dot{s}_1}(S) = \mathbf{C}^{\mathrm{EDT},O'}_{\dot{s}_2}(S'). \tag{3.5}$$

□

To show that EDT can solve PSW2, we need to supply it with a utility frame. For this, it is sufficient to use the simple frame that delivers the partition of states by grand world outcomes.

Lemma 3 (EDT Solves PSW2)**.** *EDT can solve PSW2 with*

- $\mathbf{U}^{\mathrm{EDT}}_{\dot{s}} \stackrel{\mathrm{df}}{=} O \in \mathbf{SWO}(\dot{s}), S \in \mathbf{SWS}(\dot{s}) \mapsto \mathbf{V}_{\dot{s}}(OS)$*, and*
- $\mathcal{U}^{\mathrm{EDT}}_{\dot{s}} \stackrel{\mathrm{df}}{=} S \in \mathbf{SWS}(\dot{s}) \mapsto \mathcal{X}_{\dot{s}}/S$*.*

Proof. Choose $\hat{s}_1, \hat{s}_2, O, S \in \hat{s}_1, O', S' \in \hat{s}_2$ such that $\exists \mathcal{U}^{*}_{\dot{s}_1}(S) \sqsubseteq_{\dot{s}_1, O} \mathcal{U}^{\mathrm{EDT}}_{\dot{s}_1}(S)$, and $\exists \mathcal{U}^{**}_{\dot{s}_2}(S') \sqsubseteq_{\dot{s}_2, O'} \mathcal{U}^{\mathrm{EDT}}_{\dot{s}_2}(S')$, and for some bijection $f : \mathcal{U}^{*}_{\dot{s}_1}(S)/\!/\!/O \cup \mathcal{L}(O, S) \to \mathcal{U}^{**}_{\dot{s}_2}(S')/\!/\!/O' \cup \mathcal{L}(O', S')$ such that

1. $f(OS) = O'S'$

[4] The term on the right-hand-side of this definition is a function designator — it designates the function that maps a given $O \in \mathbf{SWO}(\dot{s})$ and $S \in \mathbf{SWS}(\dot{s})$ to the conditional probability $c_{\dot{s}}(S \mid O)$. Note that $\dot{s}$ is occurring free. I use function notation like this frequently.

2. $f(O) = O'$

3. $f(S) = S'$

4. $f(O \vee S) = O' \vee S'$,

both (i) $(\mathcal{U}^{*}_{\dot{s}_1}(S)///O, \{O, S\}, \dot{s}_1)$ is EDT-imaging isomorphic under f to $(\mathcal{U}^{**}_{\dot{s}_2}(S')///O', \{O', S'\}, \dot{s}_2)$ and (ii) $(\mathcal{U}^{*}_{\dot{s}_1}(S)//O, \dot{s}_1)$ is EDT-utility isomorphic under f to $(\mathcal{U}^{**}_{\dot{s}_2}(S')//O', \dot{s}_2)$.

We have:

$$\begin{aligned}
\mathbf{U}^{\mathrm{EDT},O}_{\dot{s}_1}(S) \stackrel{\mathrm{df}}{=} \mathbf{V}_{\dot{s}_1}(OS) &= \textstyle\sum_{X \in \mathcal{X}_{\dot{s}_1}} c_{\dot{s}_1}(X \mid OS)\mathbf{V}_{\dot{s}_1}(X) = \sum_{P \in \mathcal{U}^{*}_{\dot{s}_1}(S)/O} c_{\dot{s}_1}(P \mid OS)\mathbf{V}_{\dot{s}_1}(P) \\
&\stackrel{\mathrm{df}}{=} \textstyle\sum_{P \in \mathcal{U}^{*}_{\dot{s}_1}(S)/O} \mathcal{I}^{\mathrm{EDT}}_{\dot{s}_1}(P \parallel OS)\mathbf{V}_{\dot{s}_1}(P) \\
&= \textstyle\sum_{P \in \mathcal{U}^{*}_{\dot{s}_1}(S)/O} \mathcal{I}^{\mathrm{EDT}}_{\dot{s}_2}(f(P) \parallel f(OS))\mathbf{V}_{\dot{s}_1}(f(P)) \\
&= \textstyle\sum_{P \in \mathcal{U}^{**}_{\dot{s}_2}(S')/O'} \mathcal{I}^{\mathrm{EDT}}_{\dot{s}_2}(P \parallel O'S')\mathbf{V}_{\dot{s}_2}(P) \\
&\stackrel{\mathrm{df}}{=} \textstyle\sum_{P \in \mathcal{U}^{**}_{\dot{s}_1}(S')/O'} c_{\dot{s}_2}(P \mid O'S')\mathbf{V}_{\dot{s}_2}(P) \\
&= \textstyle\sum_{X \in \mathcal{X}_{\dot{s}_2}} c_{\dot{s}_2}(X \mid O'S')\mathbf{V}_{\dot{s}_1}(X) = \mathbf{V}_{\dot{s}_2}(O'S') \stackrel{\mathrm{df}}{=} \mathbf{U}^{\mathrm{EDT},O'}_{\dot{s}_2}(S').
\end{aligned} \tag{3.6}$$

□

Theorem 4 (EDT Solves the Strong Restricted POSW)**.** *EDT can solve the Strong Restricted Problem of Small Worlds with*

- $\mathbf{C}^{\mathrm{EDT}}_{\dot{s}} \stackrel{\mathrm{df}}{=} O \in \mathbf{SWO}(\dot{s}), S \in \mathbf{SWS}(\dot{s}) \mapsto c_{\dot{s}}(S \mid O)$,
- $\mathbf{U}^{\mathrm{EDT}}_{\dot{s}} \stackrel{\mathrm{df}}{=} O \in \mathbf{SWO}(\dot{s}), S \in \mathbf{SWS}(\dot{s}) \mapsto \mathbf{V}_{\dot{s}}(OS)$, *and*
- $\mathcal{U}^{\mathrm{EDT}}_{\dot{s}} \stackrel{\mathrm{df}}{=} S \in \mathbf{SWS}(\dot{s}) \mapsto \mathcal{X}_{\dot{s}}/S$.

Proof. Follows from Lemmas 2, 3, and Theorem 3. □

3.4 Lewisian CDT is Regular Partition Invariant

As alluded to previously (§1.14, §1.15) Lewis's theory is thought not to be partition invariant. This is because it is assumed that in his theory expected utility is only defined for decision situations where states are dependency hypotheses. Joyce writes, "The major shortcoming of [any extant version of] CDT, is that it only tells us how to calculate the utilities of actions relative to the

'grand-world' partition... A $\boldsymbol{K}$-partition formulation [i.e., Lewisian CDT] does not do any better since it too uses a special partition of propositions [*viz.*, dependency hypotheses] to compute efficacy values [*viz.*, EU's]" (Joyce, 1999, 176). But the role that dependency hypotheses play in Lewis's definition of expected utility for a given decision situation can be severed from the role of states in that situation. Lewisian CDT can be given a regular partition invariant formulation and can therefore solve the Weak Restricted Problem. I'll show in the next section that it solves the Strong Restricted Problem.

For any function from decision situations to partitions $\Theta : \mathbb{D} \to \mathbb{P}(\Omega)$ such that $(\dot{s} \in \mathbb{G}, S \in \mathbf{SWS}(\dot{s})) \mapsto \Theta_{\dot{s}}//S$ is a utility frame, let $\mathbf{M}^{\Theta}$ be the function

$$(s \in \mathbb{D}, O \in s) \mapsto \big(P \in \mathcal{P}(\Omega) \mapsto \sum_{X \in \Theta_s} c_s(X \mid P)\mathbf{V}_s(OX)\big).$$

Call these *center of mass functions*. When given a decision situation s and an option O in s, a center of mass function thereafter treats any proposition you give it as a piece of evidence as to which cells in an underlying partition — determined by s in accordance with Θ — obtain, and averages the expected values assigned to conjunctions of those cells with the option, according to how likely the cells are given the proposition.

Let a *center of mass decision theory* be any theory $\mathcal{T}$ such that for some $\Theta : \mathbb{D} \to \mathbb{P}(\Omega)$,

$$\forall s \in \mathbb{D}, \forall O \in \mathcal{O}_s, \mathrm{EU}_s^{\mathcal{T}}(O) = \sum_{S \in \mathcal{S}_s} c_s(S)\mathbf{M}_s^{\Theta,O}(S)$$

$$= \sum_{S \in \mathcal{S}_s} c_s(S) \sum_{X \in \Theta_s} c_s(X \mid S)\mathbf{V}_s(OX) = \sum_{X \in \Theta_s} c_s(X)\mathbf{V}_s(OX).$$

Using the center of mass function $\mathbf{M}^{\mathcal{K}}$ we get the following expression of Lewisian CDT:

$$\forall s \in \mathbb{D}, \forall O \in \mathcal{O}_s, \mathrm{EU}_s^{\mathrm{Lew}}(O) = \sum_{S \in s} c_s(S)\mathbf{M}_s^{\mathcal{K},O}(S) = \sum_{S \in s} c_s(S) \sum_{K \in \mathcal{K}_s} c_s(K \mid S)\mathbf{V}_s(OK). \tag{3.7}$$

The underlying partitions the Lewisian center of mass function extracts from decision situations are thus just partitions of dependency hypotheses.

In proving things about theories that use center of mass functions, we'll obtain results that apply to Lewisian CDT as a corollary. Let us say that a partition extraction function $\Theta : \mathbb{D} \to \mathbb{P}(\Omega)$ is

state-blind when for any $s \in \mathbb{D}$ and $s' \in \text{PartInv}(s)$, $\Theta_s = \Theta_{s'}$. Such partition extraction functions do not depend on the state partitions of decision situations, and are therefore fixed by the equivalence classes in $\text{PartInv}(\mathbb{D})$ (cf. note **??** in §1.15).

Lemma 4 (Some Center of Mass Theories Are Regular Partition Invariant)**.** *A center of mass decision theory $\mathcal{T}(\Theta^*)$ that uses a state-blind partition extraction function Θ^* is regular partition invariant under $(s \in \mathbb{D}, P, Q \in \mathcal{P}(\Omega)) \mapsto c_s(Q)$ and $(s \in \mathbb{D}, P, Q \in \mathcal{P}(\Omega)) \mapsto \mathbf{M}_s^{\Theta^*,P}(Q)$.*

Proof. Letting $s \in \mathbb{D}$, $O \in s$, and $s' \in \text{PartInv}(s)$, we have:

$$\begin{aligned}
\sum_{S \in s} c_s(S)\mathbf{M}_s^{\Theta^*,O}(S) &= \sum_{S \in s} c_s(S) \sum_{X \in \Theta_s^*} c_s(X \mid S)\mathbf{V}_s(OX) \\
&= \sum_{S \in s} c_s(S) \sum_{X \in \Theta_s^*} \frac{c_s(SX)}{c_s(S)}\mathbf{V}_s(OX) \\
&= \sum_{S \in s} \sum_{X \in \Theta_s^*} c_s(SX)\mathbf{V}_s(OX) \\
&= \sum_{X \in \Theta_s^*} c_s(X)\mathbf{V}_s(OX) \\
&= \sum_{X \in \Theta_{s'}^*} c_{s'}(X)\mathbf{V}_{s'}(OX) \\
&= \sum_{S \in s'} \sum_{X \in \Theta_{s'}^*} c_{s'}(SX)\mathbf{V}_{s'}(OX) \\
&= \sum_{S \in s'} c_{s'}(S) \sum_{X \in \Theta_{s'}^*} \frac{c_{s'}(SX)}{c_{s'}(S)}\mathbf{V}_{s'}(OX) \\
&= \sum_{S \in s'} c_{s'}(S) \sum_{X \in \Theta_{s'}^*} c_{s'}(X \mid S)\mathbf{V}_{s'}(OX) \\
&= \sum_{S \in s'} c_{s'}(S)\mathbf{M}_{s'}^{\Theta^*,O}(S).
\end{aligned}$$

□

Corollary 1 (Some Center of Mass Decision Theories Solve the Weak Restricted POSW)**.** *Any center of mass decision theory $\mathcal{T}(\Theta)$ that uses a state blind partition extraction function and expected value solves the Weak Restricted Problem of Small Worlds with*

- $\mathbf{C}^{\mathcal{T}(\Theta,\mathbf{V})} \stackrel{\text{df}}{=} \dot{s} \in \mathbb{G}, O \in \mathbf{SWO}(\dot{s}), S \in \mathbf{SWS}(\dot{s}) \mapsto c_{\dot{s}}(S)$*, and*
- $\mathbf{U}^{\mathcal{T}(\Theta,\mathbf{V})} \stackrel{\text{df}}{=} \dot{s} \in \mathbb{G}, O \in \mathbf{SWO}(\dot{s}), S \in \mathbf{SWS}(\dot{s}) \mapsto \mathbf{M}_{\dot{s}}^{\Theta,O}(S)$.

Proof. From Lemma 4 and Theorem 2. □

We get the result for Lewisian CDT in virtue of the fact that as we saw, Lewisian CDT extracts partitions of dependency hypotheses from each decision situation in a state-blind way (cf. note 27).

Corollary 2 (Lewisian CDT Solves the Weak Restricted POSW)**.** *Lewisian CDT solves the Weak Restricted Problem of Small Worlds with*

- $\mathbf{C}^{\text{Lew}} \stackrel{\text{df}}{=} \dot{s} \in \mathbb{G}, O \in \mathbf{SWO}(\dot{s}), S \in \mathbf{SWS}(\dot{s}) \mapsto c_{\dot{s}}(S)$*, and*
- $\mathbf{U}^{\text{Lew}} \stackrel{\text{df}}{=} \dot{s} \in \mathbb{G}, O \in \mathbf{SWO}(\dot{s}), S \in \mathbf{SWS}(\dot{s}) \mapsto \mathbf{M}_{\dot{s}}^{\mathcal{K},O}(S)$.

Proof. From Corollary 1. □

3.5 A Lewisian Solution to the Strong Restricted Problem

Given Corollary 2, proving that Lewisian CDT can solve the Strong Restricted Problem of Small Worlds amounts to showing that it can solve PSW1 and PSW2. That is, we need to show that the functions under which Lewisian CDT is general partition invariant have the appropriate "locality" to function as genuine estimate correctness functions (cf. §2.5). This is immediate with the Lewisian probability function, as it's just unconditional credence, which is trivially preserved by imaging isomorphism. It's a bit less obvious that Lewisian Centers of Mass will conform to the locality requirement of PSW2, given the way they seem to pertain to all of logical space, averaging as they do over all dependency hypotheses. Lucky for Lewisians, this whiff of holism ends up being illusory.

Lemma 5 (Lewisian CDT Solves PSW1)**.** *Lewisian CDT can solve PSW1 with* $\mathbf{C}^{\text{Lewis}} \stackrel{\text{df}}{=} \dot{s} \mapsto O \in \mathbf{SWO}_{\dot{s}} \mapsto S \in \mathbf{SWS}_{\dot{s}} \mapsto c_{\dot{s}}(S)$.

Proof. Choose $\hat{s}_1, \hat{s}_2, O, S \in \hat{s}_1$, $O'S' \in \hat{s}_2$, $\dot{s}_1$, $\dot{s}_2$ where $f(OS) = O'S'$, $f(\overline{O}S) = \overline{O'}S'$, $f(O\overline{S}) = O'\overline{S'}$, $f(\overline{O}\,\overline{S}) = \overline{O'}\,\overline{S'}$ and $(\{OS, \overline{O}S, O\overline{S}, \overline{O}\,\overline{S}\}, \dot{s}_1)$ is LCDT-imaging-iso to $(\{O'S', \overline{O'}S', O'\overline{S'}, \overline{O'}\,\overline{S'}\}, \dot{s}_2)$ under f.

$$\mathbf{C}_{\dot{s}}^{\text{Lewis},O}(S) \stackrel{\text{df}}{=} c_{\dot{s}}(S) = \mathcal{I}_{\dot{s}_1}^{\text{Lewis}}(OS \vee \overline{O}S \parallel \top) = \mathcal{I}_{\dot{s}_2}^{\text{Lewis}}(O'S' \vee \overline{O'}S' \parallel \top) = c_{\dot{s}}(S) \stackrel{\text{df}}{=} \mathbf{C}_{\dot{s}}^{\text{Lewis},O}(S). \quad (3.8)$$

□

Lemma 6 (Some Center of Mass Theories Solve PSW2)**.** *A center of mass decision theory* $\mathcal{T}(\Theta)$ *can solve PSW2 with* $\mathbf{U}^{\mathcal{T}} \stackrel{\text{df}}{=} \mathbf{M}^{\Theta}$*,* $\mathcal{U}^{\mathcal{T}} \stackrel{\text{df}}{=} \dot{s} \in \mathbb{G}, S \in \mathbf{SWS}(\dot{s}) \to \Theta_{\dot{s}}//S$*.*

Proof. Choose $\hat{s}_1, \hat{s}_2, O, S \in \hat{s}_1, O', S' \in \hat{s}_2$ such that there is $\mathcal{U}^*_{\dot{s}_1}(S) \sqsubseteq_{\dot{s}_1,O} \mathcal{U}^{\mathcal{T}}_{\dot{s}_1}(S)$, and $\mathcal{U}^{**}_{\dot{s}_2}(S') \sqsubseteq_{\dot{s}_2,O'} \mathcal{U}^{\mathcal{T}}_{\dot{s}_2}(S')$, as well as a bijection $f : \mathcal{U}^*_{\dot{s}_1}(S)///O \cup \mathcal{L}(O,S) \to \mathcal{U}^{**}_{\dot{s}_2}(S')///O' \cup \mathcal{L}(O',S')$, such that both

(i) $(\mathcal{U}^*_{\dot{s}_1}(S)///O, \{O,S\}, \dot{s}_1)$ is $\mathcal{T}$-imaging-iso under f to $(\mathcal{U}^{**}_{\dot{s}_2}(S')///O', \{O',S'\}, \dot{s}_2)$ and

(ii) $(\mathcal{U}^*_{\dot{s}_1}(S)//O, \dot{s}_1)$ is utility-iso under f to $(\mathcal{U}^{**}_{\dot{s}_2}(S')//O', \dot{s}_2)$.

Note the following credence identities that come by way of the $\mathcal{T}$-imaging isomorphism:

$$c_{\dot{s}_1}(S) = \sum_{P \in \mathcal{U}^*_{\dot{s}_1}} \mathcal{I}^{\mathcal{T}}_{\dot{s}_1}(P \parallel \top) = \sum_{Q \in \mathcal{U}^{**}_{\dot{s}_2}} \mathcal{I}^{\mathcal{T}}_{\dot{s}_2}(Q \parallel \top) = c_{\dot{s}_2}(S'), \tag{3.9}$$

and

$$\forall P \in \mathcal{U}^*_{\dot{s}_1}, c_{\dot{s}_1}(P) = \mathcal{I}^{\mathcal{T}}_{\dot{s}_1}(P \parallel \top) = \mathcal{I}^{\mathcal{T}}_{\dot{s}_2}(f(P) \parallel \top) = c_{\dot{s}_2}(f(P)). \tag{3.10}$$

We will define a function $f^* : \mathcal{U}^*_{\dot{s}_1} \to \mathcal{U}^{**}_{\dot{s}_2}$ such that $f^*(P) \stackrel{\text{df}}{=} \bigvee\{f(Q) : Q \in (\mathcal{U}^*_{\dot{s}_1}///S)/P\}$. Here is the complete proof in a series of equations:

$$\begin{aligned}
\mathbf{U}^{\mathcal{T},O}_{\dot{s}_1}(S) \stackrel{\text{df}}{=} \mathbf{M}^{\Theta,O}_{\dot{s}_1}(S) &= \textstyle\sum_{P \in \Theta_{\dot{s}_1}} c(P \mid S)\mathbf{V}_{\dot{s}_1}(OP) = \sum_{P \in \Theta_{\dot{s}_1}} \frac{c_{\dot{s}_1}(PS)}{c_{\dot{s}_1}(S)}\mathbf{V}_{\dot{s}_1}(OP) \\
&= \textstyle\sum_{P \in \Theta_{\dot{s}_1}} \frac{c_{\dot{s}_1}(PS)}{c_{\dot{s}_2}(S')}\mathbf{V}_{\dot{s}_1}(OP) = \sum_{P \in \mathcal{U}^*_{\dot{s}_1}(S)} \frac{c_{\dot{s}_1}(P)}{c_{\dot{s}_2}(S')}\mathbf{V}_{\dot{s}_1}(OP) \\
&= \textstyle\sum_{P \in \mathcal{U}^*_{\dot{s}_1}(S)} \frac{c_{\dot{s}_2}(f^*(P))}{c_{\dot{s}_2}(S')}\mathbf{V}_{\dot{s}_2}(f(OP)) = \sum_{P \in \mathcal{U}^{**}_{\dot{s}_2}(S')} \frac{c_{\dot{s}_2}(P)}{c_{\dot{s}_2}(S')}\mathbf{V}_{\dot{s}_2}(O'P) \\
&= \textstyle\sum_{P \in \Theta_{\dot{s}_2}} \frac{c_{\dot{s}_2}(PS')}{c_{\dot{s}_2}(S')}\mathbf{V}_{\dot{s}_2}(O'P) = \sum_{P \in \Theta_{\dot{s}_2}} c_{\dot{s}_2}(P \mid S')\mathbf{V}_{\dot{s}_2}(O'P) \\
&= \mathbf{M}^{\Theta,O'}_{\dot{s}_2}(S') \stackrel{\text{df}}{=} \mathbf{U}^{\mathcal{T},O'}_{\dot{s}_2}(S').
\end{aligned} \tag{3.11}$$

Let's look at the two key steps in this series. The first key step comes with this:

$$\sum_{P \in \Theta_{\dot{s}_1}} \frac{c_{\dot{s}_1}(PS)}{c_{\dot{s}_2}(S')}\mathbf{V}_{\dot{s}_1}(OP) = \sum_{P \in \mathcal{U}^*_{\dot{s}_1}(S)} \frac{c_{\dot{s}_1}(P)}{c_{\dot{s}_2}(S')}\mathbf{V}_{\dot{s}_1}(OP). \tag{3.12}$$

Here we substitute members of $\mathcal{U}^*_{\dot{s}_1}(S)$ for conjunctions of S with $\Theta_{\dot{s}_1}$-cells. Each $P \in \mathcal{U}^*_{\dot{s}_1}(S)$ is either just such a conjunction or else is a disjunction of such conjunctions which are each compatible with O and whose conjunctions with O share the same utility level. Notably, the sum on the left has a

term for every $\Theta_{\dot{s}_1}$-cell, even those disjoint from S, whereas the sum on the right omits cells disjoint from S, since, of course, $\mathcal{U}^*_{\dot{s}_1}(S)$ is (a coarsening of) a utility frame for S in $\dot{s}_1$. This substitution makes for no difference in the sums because, on the left, the expected values of conjunctions of $\Theta_{\dot{s}_1}$-cells with O get zero weight when the cell is disjoint from S. So when we transition to the right hand side, we effectively just toss out some zero terms. Neither will the transition effect the expected value multipliers, though the weights get regrouped; this is ensured by the fact that we only take disjunctions of S-conjunctions with $\Theta_{\dot{s}_1}$-cells when the intersections of the disjuncts with O have the same utility level.[5]

The second key step comes with this equation:

$$\sum_{P \in \mathcal{U}^*_{\dot{s}_1}(S)} \frac{c_{\dot{s}_1}(P)}{c_{\dot{s}_2}(S')} \mathbf{V}_{\dot{s}_1}(OP) = \sum_{P \in \mathcal{U}^*_{\dot{s}_1}(S)} \frac{c_{\dot{s}_2}(f^*(P))}{c_{\dot{s}_2}(S')} \mathbf{V}_{\dot{s}_2}(f(OP)) \tag{3.13}$$

Here the $\mathcal{T}$-imaging and utility isomorphisms under f enable us to use f and the derived function f^* to map the sum for OS in $\dot{s}_1$ into an equal sum for $O'S'$ in $\dot{s}_2$. □

How does the center of mass theorist — in particular, the Lewisian — ultimately explain away the worry that centers of mass will fail to be "local" in the way required in order for them to play the role of bona fide utility estimate correctness conditions? Well as we can see if we've gone through the preceding proof, what happens is that the center of mass is a weighted average of utility-levels, namely utility-levels of logical parts of the relevant small world outcome. That is to say, the only utilities that get any weight in the average that constitutes the center of mass are utility levels of logical parts of the small world outcome. This is not accomplished directly but rather indirectly, in that the arguments that are delivered to the expected value function to obtain the terms in the sum are not themselves logical parts of the outcome necessarily, but rather conjunctions of the option with cells in the center of mass function's underlying partition. This is what perhaps gives rise to

[5] What about when for some $\Theta_{\dot{s}_1}$-cell P the conjunction OPS is empty? This isn't a problem for the proof; it's a question for the given center of mass theory to answer. Presumably such theories will have a story to tell about the expected value in a decision situation of the empty set. The issue doesn't arise, incidentally, for Lewisian CDT, since the partition extraction function always extracts dependency hypothesis partitions, and every dependency hypothesis for a given decision situation is, by definition, compatible with every option in that situation. (And as mentioned, if a dependency hypothesis is disjoint from the state, then it receives zero weight in the center of mass for that state—but will crop up in some other state.)

the suspicion as to there being some average over all of logical space or at least including parts of logical space disjoint from the state. In a way, this isn't completely an illusion because there are such parts of logical space that are getting assigned expected values and occurring in the sum; it is just that those parts of logical space are receiving zero weight because the expected value of each conjunction of a cell in the center of mass's underlying partition with the option gets its weight from the credence assigned to that cell conditional on the state. So any cells that are disjoint from the state simply receive zero credence in the sum, though the expected values of their conjunctions with the option do occur in it in a literal sense. But because they receive zero weight we see that when we make use of a utility frame to transition between separate small world outcomes that are isomorphic in the desired way, we only need to deal with expected values of literal logical parts of the small world outcome in question, which get mapped to parts of another small world outcome in a way that preserves utility level.

3.6 Demandingness

One might worry that the Lewisian solution I've just given for the Restricted Problem implausibly characterizes rational deliberation as involving humanly impossible feats of calculation. This objection has bite insofar as the point of solving the Problem of Small Worlds is in part to show how decision theory can be applied to real world deliberation. To be sure, literally mentally calculating a Lewisian center of mass in a realistic deliberative context seems superhuman. But the proposed Lewisian solution does not entail that deliberation involves such calculation. On the Joycean picture of deliberation, the deliberating agent does not need to calculate the utilities and credences they use in their deliberation — they only estimate these values, and estimation is quite a different mental activity from calculation. But what is this "estimation"? In particular, how hard will it be to estimate *correctly* on the Lewisian account? A solution to the Problem of Small Worlds which makes it nigh impossible to ever deliberate correctly is *ipso facto* not a plausible one. This worry is well taken, but I believe it to be ill-founded.

The most direct way to see through it is to consider what utility estimation means on EDT and compare that to the Lewisian picture. A common backbone is perceptible between the two if we squint a little. Utility estimation on both accounts can be viewed (perhaps a bit tendentiously)

as follows. For a given small world option O and state S, we are estimating a weighted average of utility levels of conjunctions of O with possible configurations of the causal background (i.e., dependency hypotheses) consistent with S. In both cases the weights are derived from ratios of prior credence. The only difference is in precisely what ratios we consider. On EDT we let the option itself narrow our focus; we look at the credence ratios between conjunctions of the option, the state, and the various dependency hypotheses. With Lewisian centers of mass, we look at ratios between conjunctions of the state and the configurations; we don't let the option function evidentially. But on both approaches, in any case, the utilities we are averaging are the same — there is no difference between $\mathbf{V}_s(OSK)$ and $\mathbf{V}_s(OK)$, where $O, S \in s \in \mathbb{D}$ and $K \in \mathcal{K}_s$, since OK is a utility level proposition in s. Stepping back, we see that estimating a Lewisian center of mass for S given O is equivalent to estimating the expected value for OS, provided only that O adds nothing evidential to S — that is, provided only that the state, considered as a piece of evidence as to the causal background, screens off whatever the option itself, considered as such a piece of evidence, indicates. Mathematically, if $c_s(K \parallel S) = c_s(K \parallel OS)$ for every $K \in \mathcal{K}_s$, then $\mathcal{M}_s^{\mathcal{K},O}(S) = \mathbf{V}_s(OS)$. Cases where this fails are Newcomblike, which is precisely where one would expect the divergence between the two accounts to arise.

Will Newcomblike cases make it harder to make correct utility estimates on the Lewisian account as opposed to the Evidentialist account? Let's examine an example. Consider Newcomb's Problem and suppose Robert, the agent in that case, is deliberating about whether to two-box or one-box. Suppose Robert is thinking in terms of two possible states: (1) The Predictor's prediction is (or will turn out) factually correct, and (2) The Predictor's prediction is (or will turn out) factually incorrect.[6][7] Robert is estimating the utility associated with two-boxing given the former state. On EDT, Robert should incorporate two-boxing qua piece of evidence into his conception; his estimate will be right if he assigns the utility equivalent of \$1000 to the outcome. On the Lewisian account, Robert is not to let the option enter into his conception of the likely outcome; the center of mass

[6]By "turn out factually correct" we just mean the proposition that obtains iff the predictor is correct. We don't mean, e.g., anything that supports counterfactuals.

[7]Of course, which of these two "states" obtains — or will soon obtain — is partly under Robert's control, and he is well aware of this. But so what — this is how Robert is thinking.

should be dictated solely by his prior credence as to what he expects himself to ultimately do. E.g., if his prior credence tends to cluster on the hypothesis that he'll ultimately one-box, then the Lewisian center of mass for the state (i.e., the correct prediction according to Lewisian CDT) given the option of two-boxing should approach the utility equivalent of $1,001,000. To put it another way, on the Lewisian account, if we are estimating the Lewisian center of mass for an option we do not consider what we will ultimately do as being evidentially settled by the option we are contemplating, but as being rather merely on a par with any other hypothesis about future events. This seems no more inherently difficult to get right than the competing Evidentialist requirement. The two just give diverging advice as to how to think when deliberating in Newcomblike situations. In all other situations their advice is identical.[8]

3.7 Getting "News" Out of Centers of Mass

Looking ahead to the Unrestricted Problem, where the grand world can refine not just a small world's state partition but its option partition as well, a Lewisian might have a certain qualm with approach set out in the preceding sections. The partition of dependency hypotheses our center of mass function extracts from a given grand world is based on the options in that grand world rather than on small world options. The concern is that this might open the door to the kind of "good-news-chasing" deplored by causalists.

Take a given small world $\hat{s}$ and $\dot{s} \in \mathbf{I}(\hat{s})$ where $\mathcal{O}_{\hat{s}} \neq \mathcal{O}_{\dot{s}}$, and choose an option $O \in \hat{s}$ such that $O \notin \dot{s}$. The expected value of the conjunction of O with a grand world dependency hypothesis $K \in \mathcal{K}_{\dot{s}}$ will be an average of utility levels assigned by K to grand world options that imply O. That is, $\mathbf{V}_{\dot{s}}(OK) = \sum_{O' \in \dot{s}} c_{\dot{s}}(O' \mid OK)\mathbf{V}_{\dot{s}}(O'K)$. This doesn't matter in the context of the Restricted Problem where small world options just are grand world options. But in the Unrestricted setting it will allow small world options to be evidence ("news") about what grand world options will be performed in just the sort of way Lewisians eschew.

[8] Though, to be sure, they also differ as to how to estimate the credence value for a given state. There too the evidentialist says we must let the option function as evidence, whereas the Lewisian says we must not. And as ever, this difference hides itself except in Newcomblike cases.

There is a remedy: change what partition the center of mass function gets from grand worlds. Instead of extracting $\mathcal{K}_{u_{\dot{s}}}^{\mathcal{O}_{\dot{s}}}$ from $\dot{s}$, we can instead extract $\mathcal{K}_{u_{\dot{s}}}^{\mathcal{O}_{\hat{s}}}$, i.e., we can use the small world partition of options together with the grand world utilities to determine a set of dependency hypotheses for the grand world that use small world options. Once we do this, the problem about "news" is averted, since a small world option together with one of these dependency hypotheses will imply a grand world utility level.

In fact, we can do better. It will prove even more convenient to think of the relevant set of dependency hypotheses in a given case as being determined by a specific small world option, rather than by a set of small world options. That is, rather than speak of $\mathcal{K}_{\dot{s}}^{\mathcal{O}_{\hat{s}}}$ for a given $\hat{s}$ and $\dot{s} \in \mathbf{I}(\hat{s})$, it is simpler and more convenient to speak of $\mathcal{K}_{\dot{s}}^{O}$ where $O \in \hat{s}$ and $\dot{s} \in \mathbf{I}(\hat{s})$. Thus far we've assumed we need two things to determine a set of dependency hypotheses: an option partition and a utility function; but strictly speaking, this is more than we need. The set of conjunctions of a given option with its dependency hypotheses under a particular utility function will be the same regardless of how (or if) the negation of that option is partitioned into alternative options. That is, each option itself is partitionable via the full range of maximally specific consistent claims compatible with that option concerning what is and is not under the agent's control when it comes to what they desire. What this (sub)partition is for a given option is, so to speak, a matter of logic, settled by the option itself together with the agent's utility function. Further information about what the agent's alternatives are (whether and how the option's negation is divided into separate alternatives) cannot affect that (sub)partition.

3.8 Summing Up, Looking Ahead

In this chapter we have seen that the Restricted Problem of Small Worlds is manageable. Both the evidentialists and the causalists have access to regular partition invariance, and so far as the Restricted Problem is concerned, partition invariance is killer. The only complication has been in demonstrating, in the case of Lewisian CDT, that the functions under which the theory is partition invariant also furnish duly local estimate correctness functions — as we saw, they do. So far, so good, then. But the rosy situation is about to change. In the next chapter we address ourselves

to the Unrestricted Problem, where as we shall see, it will no longer be the case that partition invariance is a silver bullet.

CHAPTER 4

THE UNRESTRICTED PROBLEM

4.1 Introduction

The Restricted Problem of Small Worlds deals only with small-worlds under the assumption that they share their option partition with the grand world. The Unrestricted Problem of Small Worlds brings in cases where idealizing from the small-world to the grand-world can involve not only adding a credence and utility function and refining the small world's partition of states, but also refining the option partition. The first difficulty of the Unrestricted Problem is just how to pose it. As briefly explained previously (§2.6), on its most flatfooted formulation, the problem is trivially unsolvable. No theory can guarantee that performing a small-world option that maximizes correctly estimated expected utility will suffice for maximizing expected utility in the grand world. Why not? Consider some small-world situation $\hat{s}$ and a grand-world idealization of it $\dot{s}$. Any option in $\hat{s}$ might be compatible with *many* options in $\dot{s}$, and we cannot assume these will all have the same expected utility. For instance, if my small-world choice is between going for a run or not and I maximize correctly estimated expected utility by going for a run, no theory can guarantee that I will thereby maximize expected utility in the grand world, as I might run in suboptimal running shoes or choose a suboptimal route. That is, the credence and utility function of my grand-world self might induce expected utilities on grand-world options that are sensitive to the shoes I wear and the route I choose, and per those expected utilities, I might've chosen suboptimally despite having gone for a run and thereby maximized correctly estimated expected utility vis-à-vis my small-world choice. It doesn't matter how correctness for estimates is defined; the problem comes from small world options being supersets of grand-world options. In short, we toss out the flatfooted idea and replace it with the prima facie more promising PSW3.

As often happens, here the prima facie is indeed only facie. In this chapter I show that the Strong Unrestricted Problem of Small Worlds, First Formulation, is unsolvable except by implausible,

abnormal theories — more on this later (§4.6). We can get an easy foretaste of this via the following simple counterexample to the flatfooted EDT solution attempt:

		State 1	**State 2**	**State 3**
Option* 1	Option 1	5	5	5
	Option 2	0	0	0
Option* 2	Option 3	3	3	3
	Option 4	3	3	3

Figure 4.1: Simple Counterexample to EDT Solution Attempt

If we let the numbers in this table be grand world expected values and assume that credence is evenly distributed over all the grand world outcomes, then the EDT-correctly estimated expected utilities are 2.5 for Option* 1 and 3 for Option* 2. But lo, while Option* 2 maximizes correctly estimated expected utility, it is not compatible with the grand world option that maximizes — Option 1. What I prove in this chapter is that there's no way to define correctness for credence and utility estimates to avoid this sort of thing without having to reject something more or less axiomatic. I start by proving a little mathematical theorem about functions on matrices (§4.2). I then use this to prove two unsolvability results (§4.3, §4.5). That done, I clarify what I mean by saying decision theories that escape these results are "abnormal" (§4.6). I close by making an argument about the range of possible decision situations where the kind of failure exhibited above in Fig. 4.1 can occur (§4.7).

4.2 A Core Theorem

I will give two separate arguments for the unsolvability of the First Formulation of the Unrestricted Problem of Small Worlds based on the following theorem.[1]

[1]Understanding the role the theorem plays in the arguments does not require understanding this theorem's proof, though I suppose it can't hurt.

Theorem 5 (Maximal Sums of Functions on Matrix Blocks versus Maximal Sums of Matrix Rows)**.** *Let $\mathcal{M}$ be the set of real matrices. Let $\boxplus : \bigcup_{n\in\mathbb{N}} \mathbb{R}^n \to \mathbb{R}$ and $\oplus : \bigcup_{n\in\mathbb{N}} \mathbb{R}^n \to \mathbb{R}$ be any polyadic functions on the reals such that*

- $\oplus(x_1, \ldots, x_n) \leq \oplus(y_1, \ldots, y_m) \Leftrightarrow \sum_{i=1}^{n} x_i \leq \sum_{i=1}^{m} y_i$
- $\boxplus(x_1, \ldots, x_n) \leq \boxplus(y_1, \ldots, y_m) \Leftrightarrow \sum_{i=1}^{n} x_i \leq \sum_{i=1}^{m} y_i$

for any $x_1, \ldots, x_n, y_1, \ldots, y_m \in \mathbb{R}$.

For any function $g : \mathcal{M} \to \mathbb{R}$, there is a 2×3 real block matrix[2]

$$\begin{bmatrix} A_1 & A_2 & A_3 \\ B_1 & B_2 & B_3 \end{bmatrix} = \begin{bmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \\ d_1 & d_2 & d_3 \end{bmatrix}$$

such that both:

1. $g(A_1) \oplus g(A_2) \oplus g(A_3) \geq g(B_1) \oplus g(B_2) \oplus g(B_3)$.[3]
2. $c_1 \boxplus c_2 \boxplus c_3 > \max\{(a_1 \boxplus a_2 \boxplus a_3), (b_1 \boxplus b_2 \boxplus b_3)\}$.

Proof. Consider some 4×3 real matrix satisfying the following:

$$\mathbf{A} = \begin{bmatrix} \begin{bmatrix} a_1 \\ b_1 = x \end{bmatrix} = A_1 & \begin{bmatrix} a_2 = x \\ b_2 = x \end{bmatrix} = A_2 & \begin{bmatrix} a_3 = x \\ b_3 = x \end{bmatrix} = A_3 \\ \begin{bmatrix} c_1 = x \\ d_1 = x \end{bmatrix} = B_1 & \begin{bmatrix} c_2 = x \\ d_2 = x \end{bmatrix} = B_2 & \begin{bmatrix} c_3 = x \\ d_3 = x \end{bmatrix} = B_3 \end{bmatrix} \tag{4.1}$$

Let us assume $x < a_1$. Hence

1. $(a_1 \boxplus a_2 \boxplus a_3) > \max\{(b_1 \boxplus b_2 \boxplus b_3), (c_1 \boxplus c_2 \boxplus c_3), (d_1 \boxplus d_2 \boxplus d_3)\}$.
2. $B_1 = A_2 = B_2 = A_3 = B_3$.

---

[2] A real block matrix is a matrix of real numbers that is treated as having been partitioned into submatrices.

[3] For clarity, we will write $\boxplus$ and $\oplus$ infix, so that $\boxplus(a, b, c) \stackrel{\text{df}}{=} a \boxplus b \boxplus c$ and $\oplus(a, b, c) \stackrel{\text{df}}{=} a \oplus b \oplus c$.

Now, either:

(**I**) $g(A_1) \leq g(B_1)$.

or

(**II**) $g(A_1) > g(B_1)$

Suppose (**I**). Then $g(A_1) \oplus g(A_2) \oplus g(A_3) \leq g(B_1) \oplus g(B_2) \oplus g(B_3)$ but $(a_1 \boxplus a_2 \boxplus a_1) > \max\{(b_1 \boxplus b_2 \boxplus b_3), (c_1 \boxplus c_2 \boxplus c_3), (d_1 \boxplus d_2 \boxplus d_3)\}$, and this is *QED*.

Suppose (**II**). Let $\mathcal{M}^{4\times1}$ be the set of 4×1 matrices. Consider the set $\mathcal{F}$ of functions on this set that never alter the first or third entries of their arguments:

$$\mathcal{F} = \left\{ f \in (\mathcal{M}^{4\times1} \to \mathcal{M}^{4\times1}) : \forall M \in \mathcal{M}^{4\times1}, f(M)_1 = M_1 \wedge f(M)_3 = M_3 \right\}. \tag{4.2}$$

We will use some shorthand for applying our function g to the outputs of functions in $\mathcal{F}$ when these are applied to columns in our matrix $\mathbf{A}$. Letting $f \in \mathcal{F}$:

- We let $f(A_i)$ be short for X and $f(B_i)$ be short for Y where $f\left(\begin{bmatrix} A_i \\ B_i \end{bmatrix}\right) = f\left(\begin{bmatrix} a_i \\ b_i \\ c_i \\ d_i \end{bmatrix}\right) = \begin{bmatrix} X \\ Y \end{bmatrix}$.[4]
- We let $gf(A_i)$ be short for $g(f(A_i))$ and $gf(B_i)$ be short for $g(f(B_i))$.
- We let $f(a_i) = k$ and $f(b_i) = l$ and $f(c_i) = m$ and $f(d_i) = n$ where $f\left(\begin{bmatrix} a_i \\ b_i \\ c_i \\ d_i \end{bmatrix}\right) = \begin{bmatrix} k \\ l \\ m \\ n \end{bmatrix}$.

We are going to mess with the second and fourth entries in the second and third columns of $\mathbf{A}$. Initially, we're focused specifically on the fourth entry in the second column.

Now either there is $f \in \mathcal{F}$ such that both

(†) $gf(B_2) - gf(A_2) \geq g(A_1) - g(B_1)$

[4]Here X and Y are both 2×1 matrices.

(‡) $(d_1 \boxplus f(d_2) \boxplus d_3) < a_1 \boxplus a_2 \boxplus a_3$

or there is not.

Suppose there is such $f \in \mathcal{F}$ and choose it. Then consider:

$$\left[\begin{array}{ccc} \begin{bmatrix} a_1 \\ b_1 = x \end{bmatrix} = A_1 & \begin{bmatrix} a_2 = x \\ f(b_2) \end{bmatrix} = f(A_2) & \begin{bmatrix} a_3 = x \\ b_3 = x \end{bmatrix} = A_3 \\ \\ \begin{bmatrix} c_1 = x \\ d_1 = x \end{bmatrix} = B_1 & \begin{bmatrix} c_2 = x \\ f(d_2) \end{bmatrix} = f(B_2) & \begin{bmatrix} c_3 = x \\ d_3 = x \end{bmatrix} = B_3 \end{array}\right] \tag{4.3}$$

In this matrix the values of g when applied to the B_i sum to a maximal amount despite the fact that $a_1 \boxplus a_2 \boxplus a_3$ is greater than both $c_1 \boxplus c_2 \boxplus c_3$ (as both rows are unchanged by f) and $d_1 \boxplus d_2 \boxplus d_3$ (which may have been changed, but ex hypothesi not by enough to make it the first row sum).[5] Note that this is *QED*.

Now suppose there is no f satisfying both (†) and (‡) above. Then one of the following two possibilities must obtain:

(**Possibility 1**) There is some upper bound n on the extent to which $f \in \mathcal{F}$ can make the value of g applied to $f(B_2)$ exceed the value of g applied to $f(A_2)$, and this upper bound is less than $g(A_1) - g(B_1)$, i.e., there is n such that

$$gf(B_2) - gf(A_2) \leq n < g(A_1) - g(B_1), \forall f \in \mathcal{F}.$$

(**Possibility 2**) Any $f \in \mathcal{F}$ that switches this inequality (i.e., any f such that $gf(B_2) - gf(A_2) \geq g(A_1) - g(B_1)$) makes it the case that $(d_1 \boxplus f(d_2) \boxplus d_3) \geq (a_1 \boxplus a_2 \boxplus a_3)$.

If neither of these obtains, then (†) and (‡) both hold of some $f \in \mathcal{F}$ and we have *QED* as observed above. We will show that if (1) the first of these two possibilities obtains, or if (2) the second but

[5] It doesn't matter what f does to b_2.

not the first obtains, then we get *QED* anyway, despite the fact that (as we have supposed) there is no function $f \in \mathcal{F}$ satisfying both (†) and (‡).

Suppose (**Possibility 1**) *obtains.*[6] Consider:

$$\left[\begin{array}{lll} \begin{bmatrix} a_1 \\ b_1 = x \end{bmatrix} = A_1 & \begin{bmatrix} a_2 = x \\ b_2 = x \end{bmatrix} = f(A_2) = A_2 & \begin{bmatrix} a_3 = x \\ b_3 = x \end{bmatrix} = A_3 \\ \begin{bmatrix} c_1 = x \\ d_1 = x \end{bmatrix} = B_1 & \begin{bmatrix} c_2 = x \\ d_2 + y \end{bmatrix} = f(B_2) & \begin{bmatrix} c_3 = x \\ d_3 = x \end{bmatrix} = B_3 \end{array}\right] \tag{4.4}$$

We pick f such that

$$(d_1 \boxplus (d_2 + y) \boxplus d_3) > (a_1 \boxplus a_2 \boxplus a_3). \tag{4.5}$$

But ex hypothesi $g(A_1) - g(B_1) > g(B_2) - g(A_2)$, and of course $g(A_3) = g(B_3)$. Hence,

$$g(A_1) \oplus gf(A_2) \oplus g(A_3) > g(B_1) \oplus gf(B_2) \oplus g(B_3). \tag{4.6}$$

Thus we have *QED*.[7]

Suppose (**Possibility 2**) *obtains but* (**Possibility 1**) *does not.* Consider:

$$\left[\begin{array}{lll} \begin{bmatrix} a_1 \\ b_1 = x \end{bmatrix} = A_1 & \begin{bmatrix} a_2 = x \\ b_2 - y \end{bmatrix} = f_1(A_2) & \begin{bmatrix} a_3 = x \\ b_3 + z \end{bmatrix} = f_2(A_3) = f_2(B_3) \\ \begin{bmatrix} c_1 = x \\ d_1 = x \end{bmatrix} = B_1 & \begin{bmatrix} c_2 = x \\ d_2 = x \end{bmatrix} = f_1(B_2) = B_2 & \begin{bmatrix} c_3 = x \\ d_3 + z \end{bmatrix} = f_2(B_3) = f_2(A_3) \end{array}\right] \tag{4.7}$$

We pick f_1 and f_2 such that

$$d_1 \boxplus d_2 \boxplus (d_3 + z) > \begin{cases} a_1 \boxplus a_2 \boxplus a_3 > c_1 \boxplus c_2 \boxplus c_3 \\ b_1 \boxplus (b_2 - y) \boxplus (b_3 + z) \end{cases} \tag{4.8}$$

[6] It doesn't matter whether in addition (**Possibility 2**) obtains vacuously or not.

[7] To meet the letter of the theorem, we just need to swap A_1 for B_1 and $f(A_2)$ for $f(B_2)$.

Since reducing b_2 by y does not alter the first row (the a's) or the fourth row (the d's), (‡) is true of f_1. Given our supposition that (†) and (‡) are not both true of any $f \in \mathcal{F}$, it follows that $g(A_1) - g(B_1) > gf_1(B_2) - gf_1(A_1)$. Since $f_2(A_3) = f_2(B_3)$, it follows that

$$g(A_1) \oplus gf_1(A_2) \oplus gf_2(A_3) > g(B_1) \oplus gf_1(B_2) \oplus gf_2(B_3). \tag{4.9}$$

Thus, we have *QED*.[8] □

4.3 First Unsolvability Result

Each decision theory provides both a probability function and a utility function for the purposes of defining expected utility (§1.5–§1.7). Consider some small world outcome $\bar{X} = \bar{O}\bar{S}$ compatible with some range of grand world outcomes $O^1S^1, \ldots, O^nS^n$. The key premise of this first argument is roughly this: the function $O^iS^j \mapsto p_{\dot{s}}^{\mathcal{T}}(O^iS^j) \times v_{\dot{s}}^{\mathcal{T}}(O^iS^j)$ must determine the value $\mathbf{C}_{\dot{s}}^{\mathcal{T},\bar{O}}(\bar{S}) \times \mathbf{U}_{\dot{s}}^{\mathcal{T},\bar{O}}(\bar{S})$. Assuming this holds, we can use Thm. 5 to show that no theory satisfying PSW1 and PSW2 can satisfy PSW3, provided that two other mild assumptions hold.

Some terminology will be useful. For a grand world outcome X, call $p_{\dot{s}}^{\mathcal{T}}(X) \times v_{\dot{s}}^{\mathcal{T}}(X)$ the $\mathrm{EU}^{\mathcal{T}}$*-term for* X *in* $\dot{s}$. Call $\mathbf{C}_{\dot{s}}^{\mathcal{T},O}(S) \times \mathbf{U}_{\dot{s}}^{\mathcal{T},O}(S)$, for a given small world outcome $X = OS$, the $\mathcal{T}$*-correctness term for* X *in* $\dot{s}$. Using these terms, the key premise can be restated like so:

Correctness Fixity (Rough Version): The assignment of $\mathrm{EU}^{\mathcal{T}}$-terms to outcomes in $\dot{s}$ compatible with a given small world outcome must determine the $\mathcal{T}$-correctness term in $\dot{s}$ for that small world outcome.

Now for a bit more mathematical concreteness. For a given $\dot{s}$, $X \in \mathbf{SWX}(\dot{s})$, and theory $\mathcal{T}$, let $\mathbf{EUTermMat}(\mathcal{T}, \dot{s}, X) \subseteq \mathcal{M}$ be the set of matrices such that $M^{n\times m} \in \mathbf{EUTermMat}(\mathcal{T}, \dot{s}, X)$ iff for some enumeration $O_k : \{1 \ldots n\} \to \mathcal{O}_{\dot{s}}$ of the options in $\dot{s}$ compatible with X and some enumeration $S_k : \{1 \ldots m\} \to \mathcal{S}_{\dot{s}}$ of the states in $\dot{s}$ compatible with X, we have $M_{ij} = p_{\dot{s}}^{\mathcal{T}}(O_iS_j) \times v_{\dot{s}}^{\mathcal{T}}(O_iS_j)$ for every $i \leq n$ and $j \leq m$. Just as $\mathbf{EUTermMat}(\mathcal{T}, \dot{s}, X)$ records the $\mathrm{EU}^{\mathcal{T}}$-terms in $\dot{s}$ for outcomes

[8] Again, to meet the letter of the theorem, we need to swap A_1 for B_1 and $f_1(A_2)$ for $f_2(B_2)$.

compatible with X, $\mathbf{EUTermMat}(\mathcal{T}, \dot{s}, \top)$ records $\mathrm{EU}^{\mathcal{T}}$-terms for *all* outcomes in $\dot{s}$. Correctness Fixity thus can be expressed as follows:

Constraint 1 (Correctness Fixity:)**.** *There is a function* $\mathscr{C}^{\mathcal{T}} : \mathcal{M} \to \mathbb{R}$ *such that for any grand world* $\dot{s}$, $O \in \mathbf{SWO}(\dot{s})$, *and* $S \in \mathbf{SWS}(\dot{s})$, *we have:*

$$\mathscr{C}^{\mathcal{T}}(M) = \mathbf{C}_{\dot{s}}^{\mathcal{T},O}(S) \times \mathbf{U}_{\dot{s}}^{\mathcal{T},O}(S) \text{ iff } M \in \mathbf{EUTermMat}(\mathcal{T}, \dot{s}, OS).$$

Now if a theory $\mathcal{T}$ satisfies Constraint 1, then given Theorem 5 we may immediately infer:

$\mathcal{T}$-problem Matrix: There is a 2×3 real block $\mathcal{T}$-*problem matrix*

$$\begin{bmatrix} A_1 & A_2 & A_3 \\ B_1 & B_2 & B_3 \end{bmatrix} = \begin{bmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \\ d_1 & d_2 & d_3 \end{bmatrix}$$

such that both:

1. $\mathscr{C}^{\mathcal{T}}(A_1) + \mathscr{C}^{\mathcal{T}}(A_2) + \mathscr{C}^{\mathcal{T}}(A_3) \geq \mathscr{C}^{\mathcal{T}}(B_1) + \mathscr{C}^{\mathcal{T}}(B_2) + \mathscr{C}^{\mathcal{T}}(B_3)$.
2. $c_1 + c_2 + c_3 > \max\{(a_1 + a_2 + a_3), (b_1 + b_2 + b_3)\}$.

Lemma 7. *Any theory* $\mathcal{T}$ *that satisfies Constraint 1 also satisfies* $\mathcal{T}$*-problem Matrix.*

Proof. Trivial. □

If **$\mathcal{T}$-problem Matrix** holds of a given theory $\mathcal{T}$, we can prove that the Unrestricted Problem of Small Worlds, First Formulation, is unsolvable for $\mathcal{T}$ provided that the $\mathcal{T}$-problem matrix records the $\mathrm{EU}^{\mathcal{T}}$-terms in some grand world situation which is the idealization of some small world with $\mathcal{T}$-correct estimates, provided, that is, that $\mathcal{T}$ satisfies both of the following additional constraints:

Constraint 2 ($\mathrm{EU}^{\mathcal{T}}$-term Richness)**.** *For any* 4×3 *matrix* M, *there is some grand world* $\dot{s}$ *such that* $M \in \mathbf{EUTermMat}(\mathcal{T}, \dot{s}, \Omega)$.

Constraint 3 ($\mathcal{T}$-correctness Richness)**.** *For any $\dot{s}$ and representation base $\mathcal{B}$ recoverable from the representation base of $\dot{s}$, there is a small world $\hat{s}$ with representation base $\mathcal{B}$ such that $\dot{s} \in \mathbf{I}(\hat{s})$ and $\mathbf{c}_{\hat{s}}(O) = \mathbf{C}_{\dot{s}}^{\mathcal{T},O}(S)$ and $\mathbf{u}_{\hat{s}}(O) = \mathbf{U}_{\dot{s}}^{\mathcal{T},O}(S)$ for all $O, S \in \hat{s}$.*

A theory $\mathcal{T}$ that is both $\mathrm{EU}^{\mathcal{T}}$-term rich and $\mathcal{T}$-correctness rich is one where, first, $\mathrm{EU}^{\mathcal{T}}$-terms in grand worlds with four options and three states can be any pattern of reals we like. Second, such a theory is one where for any grand world with representation base $\mathcal{B}$ and any representation base $\mathcal{B}^*$ recoverable from $\mathcal{B}$ it is possible to make correct $\mathcal{T}$-estimates for the outcomes in $\mathcal{B}^*$, meaning there is some small world having $\mathcal{B}^*$ as its representation base that idealizes to the grand world in question and that has $\mathcal{T}$-correct estimate functions.

Theorem 6 (First Unsolvability Result)**.** *No theory meeting Constraints 1, 2, & 3 solves PSW1–3.*

Proof. Choose some such theory $\mathcal{T}$. Suppose for reductio that there are correctness functions $\mathbf{C}^{\mathcal{T}}$ and $\mathbf{U}^{T}$ that satisfy PSW3. Given **$\mathcal{T}$-problem Matrix** there is a 4×3 problem matrix

$$\begin{bmatrix} A_1 & A_2 & A_3 \\ B_1 & B_2 & B_3 \end{bmatrix} = \begin{bmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \\ d_1 & d_2 & d_3 \end{bmatrix}$$

such that both:

(1) $\mathscr{C}^{\mathcal{T}}(A_1) + \mathscr{C}^{\mathcal{T}}(A_2) + \mathscr{C}^{\mathcal{T}}(A_3) \geq \mathscr{C}^{\mathcal{T}}(B_1) + \mathscr{C}^{\mathcal{T}}(B_2) + \mathscr{C}^{\mathcal{T}}(B_3)$.

(2) $c_1 + c_2 + c_3 > \max\{(a_1 + a_2 + a_3), (b_1 + b_2 + b_3)\}$.

Call this problem matrix $M^{\mathcal{T}}$. Given that $\mathcal{T}$ satisfies Constraint 2, there is some grand world decision situation $\dot{s}$ such that $M^{\mathcal{T}} \in \mathbf{EUTermMat}(\mathcal{T}, \dot{s}, \Omega)$. Hence there is an enumeration $O_k : \{1,2,3,4\} \to \mathcal{O}_{\dot{s}}$ of the options of $\dot{s}$ and an enumeration $S_k : \{1,2,3\} \to \mathcal{S}_{\dot{s}}$ of the states of $\dot{s}$ such that $M_{ij}^{\mathcal{T}} = p_{\dot{s}}^{\mathcal{T}}(O_i S_j) \times v_{\dot{s}}^{\mathcal{T}}(O_i S_j)$. Using this enumeration, it follows that $A_i \in \mathbf{EUTermMat}(\mathcal{T}, \dot{s}, (O_1 \vee O_2) \wedge S_i)$ and $B_i \in \mathbf{EUTermMat}(\mathcal{T}, \dot{s}, (O_3 \vee O_4) \wedge S_i)$ for $i \in \{1,2,3\}$. Given Constraint 1 and (1) above and letting $O_A = O_1 \vee O_2$ and $O_B = O_3 \vee O_4$, it follows that $\sum_{i=1}^{3} \mathbf{C}_{\dot{s}}^{\mathcal{T},O_A}(S_i) \times \mathbf{U}_{\dot{s}}^{\mathcal{T},O_A}(S_i) \geq \sum_{i=1}^{3} \mathbf{C}_{\dot{s}}^{\mathcal{T},O_B}(S_i) \times \mathbf{U}_{\dot{s}}^{\mathcal{T},O_B}(S_i)$. Given that $\mathcal{T}$ satisfies Constraint 3, we may choose $\hat{s}$ such that $\dot{s} \in \mathbf{I}(\hat{s})$, $\mathcal{O}_{\hat{s}} = \{O_A, O_B\}$,

and $\mathcal{S}_{\hat{s}} = \{S_1, S_2, S_3\}$, and for $i \in \{1,2,3\}$, $\mathbf{c}_{\hat{s}}(O_A S_i) = \mathbf{C}_{\hat{s}}^{\mathcal{T},O_A}(S_i)$ and $\mathbf{u}_{\hat{s}}(O_A S_i) = \mathbf{U}_{\hat{s}}^{\mathcal{T},O_A}(S_i)$. Since O_A maximizes correct $\mathbf{EstEU}_{\hat{s}}$, it follows that O_A is compatible with a grand world option $O \in \mathcal{O}_{\dot{s}}$ that maximizes $\mathrm{EU}_{\dot{s}}^{\mathcal{T}}$. But given (2) above, $\mathrm{EU}_{\dot{s}}^{\mathcal{T}}(O_3) > \max\{\mathrm{EU}_{\dot{s}}^{\mathcal{T}}(O_1), \mathrm{EU}_{\dot{s}}^{\mathcal{T}}(O_2)\}$. Contradiction. □

Summing up, our first argument relies on two mild constraints (viz., Constraints 2 and 3) and the more substantive Constraint 1. Later (§4.6) I'll say a few words about Constraint 1. There is a stronger argument we can make that relies on weaker constraints. We turn to that now.

4.4 Preliminary Concepts

As with the preceding argument we will need to define some new concepts. But there will be more this time so we proceed even more deliberately.

4.4.0.1 $\mathcal{T}$-imaging Homogeneity

Say that a set Γ of propositions is *$\mathcal{T}$-imaging homogenous in $\dot{s}$* when for any subset $S \subseteq \Gamma$ and any $P, Q \in S$, $\mathcal{I}_{\dot{s}}^{\mathcal{T}}(P \parallel \bigvee S) = \mathcal{I}_{\dot{s}}^{\mathcal{T}}(Q \parallel \bigvee S)$. Next, say that a grand world decision situation $\dot{s}$ is *$\mathcal{T}$-imaging homogenous* when its set of outcomes $\mathcal{X}_{\dot{s}}$ is $\mathcal{T}$-imaging homogenous in $\dot{s}$.

When a grand world $\dot{s}$ is $\mathcal{T}$-imaging homogenous, this implies that $c_{\dot{s}}(X) = c_{\dot{s}}(Y)$ for any $X, Y \in \mathcal{X}_{\dot{s}}$, but the converse does not hold. This is because imaging into a set of outcomes each having equal prior credence to the other need not result in these outcomes each receiving equal imaged credence — theories can interpret imaging in ways other than conditionalization. $\mathcal{T}$-imaging homogeneity in a grand world is strictly stronger than there being equal credence in the outcomes of that grand world.

4.4.0.2 Grand World Outcome Uniformity

Say a set of propositions Γ is *grand world outcome uniform in $\dot{s}$* when there is $n > 0$ such that for any $X \in \Gamma$, there is $S \subseteq \mathcal{X}_{\dot{s}}$ such that $X = \bigvee S$ and $|S| = n$.

In what follows I'll abbreviate 'grand world outcome uniform' as 'GWO uniform'. Say that a small world $\hat{s}$ is GWO uniform in a grand world $\dot{s}$ iff $\mathcal{X}_{\hat{s}}$ is GWO uniform in $\dot{s}$. Last, for any grand

world $\dot{s}$, let $\mathbf{Uniform}^{i,j}_{\dot{s}}$ be the set containing every small world $\hat{s}$ that has i options and j states and that is GWO uniform in $\dot{s}$.

4.4.0.3 UniHom Sets

For any $i, j > 0$, let $\mathbf{UniHom}^{\mathcal{T}}_{i,j}$ be the union of all $\mathbf{Uniform}^{i,j}_{\dot{s}}$ where $\dot{s}$ is $\mathcal{T}$-imaging homogenous. $\mathbf{UniHom}^{\mathcal{T}}_{i,j}$ is thus a set of small worlds, i.e., the set containing every small world $\hat{s}$ that has i options and j states and that is GWO uniform in at least one $\mathcal{T}$-imaging homogenous grand world.

4.4.0.4 A Base Function

For each $i, j > 0$ we define a function $\underline{h}^{\mathcal{T},i,j} : \mathbf{UniHom}^{\mathcal{T}}_{i,j} \to \mathcal{P}(\mathbb{G})$ such that $\underline{h}^{\mathcal{T},i,j}(\hat{s})$ is the set of all $\mathcal{T}$-imaging homogenous grand worlds that $\hat{s}$ is GWO uniform in.[9]

4.4.0.5 A Function on Boolean Algebras

For any $x, y > 0$ and $X, Y \in \mathcal{P}(\Omega)$ we define a function $\xi^{x,y}_{X,Y}$ on the algebra[10] $\mathcal{A}(X, Y)$ as follows. First, for the partition $\Gamma(X, Y)$ we set the following values:

1. $\xi^{x,y}_{X,Y}(XY) = 1$
2. $\xi^{x,y}_{X,Y}(X\overline{Y}) = y - 1$
3. $\xi^{x,y}_{X,Y}(\overline{X}Y) = x - 1$
4. $\xi^{x,y}_{X,Y}(\overline{X}\,\overline{Y}) = (x - 1)(y - 1)$.

We extend $\xi^{x,y}_{X,Y}$ to the rest of the algebra $\mathcal{A}(X, Y)$ through summation of values of atoms $\Gamma(X, Y)$ like so: $\xi^{x,y}_{X,Y}(P) \stackrel{\text{df}}{=} \sum_{P \supseteq Q \in \Gamma(X,Y)} \xi^{x,y}_{X,Y}(Q)$.[11] Don't worry now about what this function is supposed to be doing. It will become clear later.

---

[9] I will leave the superscripts tacit hereafter.

[10] For the definition of a Boolean algebra, see §2.3.1.

[11] We leave $\xi^{x,y}_{X,Y}$ undefined for $\varnothing$. For our characterization of atoms, cf. §2.3.1.

4.4.0.6 UMatrices

For a given $\dot{s}$ and $X \in \mathbf{SWX}(\dot{s})$, let $\mathbf{UMat}(\dot{s}, X) \subseteq \mathcal{M}$ be the set of matrices such that $M^{n\times m} \in \mathbf{UMat}(\dot{s}, X)$ iff for some enumeration $O_k : \{1 \dots n\} \to \mathcal{O}_{\dot{s}}$ of the options in $\dot{s}$ compatible with X and some enumeration $S_k : \{1 \dots m\} \to \mathcal{S}_{\dot{s}}$ of the states in $\dot{s}$ compatible with X, $M_{ij} = u_{\dot{s}}(O_i S_j)$ for every $i \leq n$ and $j \leq m$.[12] **UMat** is analogous to **EUTermMat**, except its values represent utilities.

4.4.0.7 Two Constraints on Decision Theories

In the same way that the previous unsolvability result was restricted to theories abiding by three constraints, this second result will be restricted to theories obeying two constraints stated below. The first constraint is that for any real matrix regularly partitioned into a block matrix, there is some $\mathcal{T}$-imaging homogenous grand world having that matrix as a utility matrix, and some small world that idealizes to that grand world whose outcomes correspond to the blocks in the block matrix. The second constraint requires of a theory $\mathcal{T}$ that if a grand world is $\mathcal{T}$-imaging homogenous, then the probabilities $\mathcal{T}$ assigns to each outcome relative to each option in that grand world — i.e., the probabilities used to calculate $\mathcal{T}$-expected utilities (§1.5) — be identical. This does not require theories to identify their imaging functions with the probability functions they use to calculate expected utilities, but rather that such probabilities at least supervene on the behavior of the imaging function.

Constraint 4 (**UniHom**$^{\mathcal{T}}$ Richness:)**.** *For any $n, m, x, y \in \mathbb{I}^+$ such that*[13] *$n|x$ and $m|y$ and any $x \times y$ real matrix M, there are $\hat{s} \in \mathbf{UniHom}^{\mathcal{T}}_{n,m}$ and $\dot{s} \in \underline{h}(\hat{s})$ and enumerations (bijections) $O_i : \{1 \dots x\} \to \mathcal{O}_{\dot{s}}$, $S_i : \{1 \dots y\} \to \mathcal{S}_{\dot{s}}$, $O^*_i : \{1 \dots n\} \to \mathcal{O}_{\hat{s}}$, $S^*_i : \{1 \dots m\} \to \mathcal{S}_{\hat{s}}$, such that*

- $M_{ij} = u_{\dot{s}}(O_i S_j)$, $\forall i \in \{1, \dots, x\}, j \in \{1, \dots, y\}$,
- $O^*_i = \bigvee\{O_j : (x(i-1)/n) + 1 \leq j \leq xi/n\}$, $\forall i \in \{1, \dots, n\}$ *and*
- $S^*_i = \bigvee\{S_j : (y(i-1)/m) + 1 \leq j \leq yi/m\}$, $\forall i \in \{1, \dots, m\}$.

[12] I will make use of the harmless pretense that the utility function in a grand world applies directly to the outcomes therein. This saves having to put everything in terms of utility-level propositions.

[13] Here $\ulcorner a|b \urcorner$ means $\ulcorner b$ is evenly divisible by $a \urcorner$.

Constraint 5 ($\mathcal{T}$-imaging Homogeneity Implies $\mathcal{T}$-probability Homogeneity:)**.** *If $\dot{s}$ is $\mathcal{T}$-imaging homogenous, then $p^{\mathcal{T}}_{\dot{s}}(X) = {}^{1}/_{|\mathcal{S}_{\dot{s}}|}$ for any $X \in \mathcal{X}_{\dot{s}}$.*

4.5 Second Unsolvability Result

With all of these preliminaries out of the way, we proceed now to the second unsolvability result. One last thing is worth noting here at the outset: concepts defined back in §2.3 will be important here, notably $\mathcal{T}$-imaging and utility isomorphisms.

Lemma 8 ($\mathcal{T}$-imaging in $\mathcal{T}$-imaging Homogenous Disjoint Propositions)**.** *For any grand world $\dot{s}$ and set of disjoint propositions Γ, if Γ is $\mathcal{T}$-imaging homogenous in $\dot{s}$, then $\mathcal{I}^{\mathcal{T}}_{\dot{s}}(P \parallel \bigvee\Gamma^*) = {}^{1}/_{|\Gamma^*|}$ for any $P \in \Gamma^* \subseteq \Gamma$, so long as Γ^* is finite.*

Proof. Trivial. □

Corollary 3 ($\mathcal{T}$-imaging Grand Outcome Disjunctions in $\mathcal{T}$-imaging Homogenous Grand Worlds)**.** *If grand world $\dot{s}$ is $\mathcal{T}$-imaging homogenous then for any set of its outcomes $\mathcal{X}^* \subseteq \mathcal{X}_{\dot{s}}$ and subset of that set $\mathcal{X}^{**} \subseteq \mathcal{X}^*$, $\mathcal{I}^{\mathcal{T}}_{\dot{s}}(\bigvee\mathcal{X}^{**} \parallel \bigvee\mathcal{X}^*) = {}^{|\mathcal{X}^{**}|}/_{|\mathcal{X}^*|}$.*

Proof. Trivial. □

Discussion of Lemma 1 and Corollary 3. These first two little results simply establish the way that imaging with finite collections of outcomes in $\mathcal{T}$-imaging homogenous grand worlds looks like applying the principle of indifference.

Lemma 9 ($\mathcal{T}$-imaging Uniform Small World Outcome Disjunctions in $\mathcal{T}$-imaging Homogenous Grand World)**.** *For any $x, y > 0$, any small world $\hat{s} \in \mathbf{UniHom}^{\mathcal{T}}_{x,y}$ and grand world $\dot{s} \in \underline{h}(\hat{s})$, any option O and state S in $\hat{s}$, and for any $X, Y \in \mathcal{A}(O,S)$, $\mathcal{I}^{\mathcal{T}}_{\dot{s}}(X \parallel Y) = \frac{\xi^{x,y}_{O,S}(X)}{\xi^{x,y}_{O,S}(Y)}$ if $X \subseteq Y$, and $\mathcal{I}^{\mathcal{T}}_{\dot{s}}(X \parallel Y) = 0$ otherwise.*

Proof. Given the GWO uniformity of $\hat{s}$ in $\dot{s}$, every $\hat{s}$ outcome contains (supersets) the same number of $\dot{s}$ outcomes; let that number be n. (That is, $|\{X \in \mathcal{X}_{\dot{s}} : X \subseteq Y\}| = n$ for any $Y \in \mathcal{X}_{\hat{s}}$.) Let $\mathrm{o}_{\dot{s}}(P) \stackrel{\mathrm{df}}{=} |\{X \in \mathcal{X}_{\dot{s}} : X \subseteq P\}|$ for any $P \in \mathcal{P}(\Omega)$. Given the $\mathcal{T}$-imaging homogeneity of $\dot{s}$ and Corollary 3, it follows that $\mathcal{I}^{\mathcal{T}}_{\dot{s}}(X \parallel Y) = {}^{n\,\mathrm{o}_{\dot{s}}(X)}/_{n\,\mathrm{o}_{\dot{s}}(Y)} = {}^{\mathrm{o}_{\dot{s}}(X)}/_{\mathrm{o}_{\dot{s}}(Y)}$ if $X \subseteq Y$ and $\mathcal{I}^{\mathcal{T}}_{\dot{s}}(X \parallel Y) = 0$ otherwise for any $X, Y \in \mathcal{A}(O,S)$. Clearly, $\mathrm{o}_{\dot{s}}$ restricted to $\mathcal{A}(O,S)$ is just $\xi^{x,y}_{O,S}$ (cf. §4.4.0.5). □

Corollary 4 (4-partitions from Single Outcomes of Small Worlds in $\mathbf{UniHom}^{\mathcal{T}}_{n,m}$ are $\mathcal{T}$-imaging Isomorphic)**.** *For any $n, m > 0$ and small world(s) $\hat{s}, \hat{x} \in \mathbf{UniHom}^{\mathcal{T}}_{n,m}$, and for any $O, S \in \hat{s}$, and $O', S' \in \hat{x}$, there is a bijection f such that $(\{OS, \overline{O}S, O\overline{S}, \overline{O}\,\overline{S}\}, \underline{h}(\hat{s}))$ is $\mathcal{T}$-imaging isomorphic under f to $(\{O'S', \overline{O'}S', O'\overline{S'}, \overline{O'}\,\overline{S'}\}, \underline{h}(\hat{x}))$.*[14]

Proof. Choose $n, m > 0$ and small worlds $\hat{s}, \hat{x} \in \mathbf{UniHom}^{\mathcal{T}}_{n,m}$. Choose $O, S \in \hat{s}$, and $O', S' \in \hat{x}$. Define f in the obvious way. The $\mathcal{T}$-imaging isomorphism follows given that per Lemma 9 for any $P \in \mathcal{A}(O, S)$, $\xi^{n,m}_{O,S}(P) = \xi^{n,m}_{O',S'}(f(P))$. □

Corollary 5 (PSW1 Implies One Corr. Cred. Est. in Hom. GWs For All Outcomes All SWs in $\mathbf{UniHom}^{\mathcal{T}}_{n,m}$)**.** *For any theory $\mathcal{T}$ and credence correctness function $\mathbf{C}^{\mathcal{T}}$ satisfying PSW1, there is a function $f^{\mathcal{T}} : \mathbb{I}^+ \times \mathbb{I}^+ \to [0, 1]$, such that for any $n, m > 0$ and any $\hat{s} \in \mathbf{UniHom}^{\mathcal{T}}_{n,m}$, $O, S \in \hat{s}$, $\dot{s} \in \underline{h}(\hat{s})$, $\mathbf{C}^{\mathcal{T},O}_{\dot{s}}(S) = f^{\mathcal{T}}(n, m)$.*

Proof. Trivial given Corollary 4 and the definition of PSW1. □

Discussion of Lemma 9 and Corollaries 4 and 5. What Corollary 4 shows is that any two outcomes in any two small worlds in a single $\mathbf{UniHom}^{\mathcal{T}}$ collection are $\mathcal{T}$-imaging isomorphic. Hence, and this is what Corollary 5 shows, a theory that satisfies PSW1 will, per the definition of that constraint, have to assign the same $\mathcal{T}$-correct credence estimate value to every small world outcome in every small world in a given $\mathbf{UniHom}^{\mathcal{T}}$ collection. Loosely put, if a theory $\mathcal{T}$ satisfies PSW1, that guarantees that $\mathcal{T}$-correctness in credence estimation for small world outcomes supervenes on $\mathcal{T}$-imaging isomorphism, and all outcomes in all small worlds in a given $\mathbf{UniHom}^{\mathcal{T}}$ collection are $\mathcal{T}$-imaging isomorphic. Hence, $\mathcal{T}$-correct credence estimation values for outcomes in small worlds in $\mathbf{UniHom}^{\mathcal{T}}$ collections are a function solely of the dimensions (i.e., the associated number of options and states) of the collections.

Lemma 10 (Same UMatrix in Small World Outcomes Entails Utility Isomorphism in $\mathbf{UniHom}^{\mathcal{T}}_{n,m}$)**.** *For any theory $\mathcal{T}$ with utility frame $\mathcal{U}^{\mathcal{T}}$ and any $n, m > 0$, $\hat{s}, \hat{x} \in \mathbf{UniHom}^{\mathcal{T}}_{n,m}$, $\dot{s} \in \underline{h}(\hat{s})$, $\dot{x} \in \underline{h}(\hat{x})$, $O, S \in \hat{s}$, and $O', S' \in \hat{x}$, if $M \in \mathbf{UMat}(\dot{s}, OS) \cap \mathbf{UMat}(\dot{x}, O'S')$, then there are $\mathcal{U}^{*}_{\dot{s}}(S) \sqsubseteq_{\dot{s},O} \mathcal{U}^{\mathcal{T}}_{\dot{s}}(S)$,*

[14] Here we use a generalization presented in §2.3.3.

$\mathcal{U}^*_{\dot{x}}(S') \sqsubseteq_{\dot{x},O'} \mathcal{U}^{\mathcal{T}}_{\dot{x}}(S')$, *and bijection* $f : \mathcal{U}^*_{\dot{s}}(S)///O \cup \mathcal{L}(O,S) \to \mathcal{U}^*_{\dot{x}}(S')///O' \cup \mathcal{L}(O',S')$ *such that (i)* $f(OS) = O'S'$, $f(O) = O'$, $f(S) = S'$, $f(O \vee S) = O' \vee S'$, *(ii)* $(\mathcal{U}^*_{\dot{s}}(S)///O, \{O,S\}, \dot{s})$ *is* $\mathcal{T}$*-imaging isomorphic under* f *to* $(\mathcal{U}^*_{\dot{x}}(S')///O', \{O',S'\}, \dot{x})$ *and (iii)* $(\mathcal{U}^*_{\dot{s}}(S)//O, \dot{s})$ *is utility isomorphic under* f *to* $(\mathcal{U}^*_{\dot{x}}(S')//O', \dot{x})$.

Proof. Choose $n, m > 0$, $\hat{s}, \hat{x} \in \mathbf{UniHom}^{\mathcal{T}}_{n,m}$, $\dot{s} \in \underline{h}(\hat{s})$, $\dot{x} \in \underline{h}(\hat{x})$, $O, S \in \hat{s}$, $O', S' \in \hat{x}$, and $M \in \mathcal{M}$ such that $M \in \mathbf{UMat}(\dot{s}, OS) \cap \mathbf{UMat}(\dot{x}, O'S')$. Choose enumerations $O^{\dot{s}}_i : \{1, \ldots, \mathbf{Rows}(M)\} \to \mathcal{O}_{\dot{s}}$, $O^{\dot{x}}_i : \{1, \ldots, \mathbf{Rows}(M)\} \to \mathcal{O}_{\dot{x}}$, $S^{\dot{s}}_i : \{1, \ldots, \mathbf{Cols}(M)\} \to \mathcal{S}_{\dot{s}}$, and $S^{\dot{x}}_i : \{1, \ldots, \mathbf{Cols}(M)\} \to \mathcal{S}_{\dot{x}}$, such that $M_{ij} = u_{\dot{s}}(O^{\dot{s}}_i S^{\dot{s}}_j) = u_{\dot{x}}(O^{\dot{x}}_i S^{\dot{x}}_j)$, for $i \in \{1, \ldots, \mathbf{Rows}(M)\}$ and $j \in \{1, \ldots, \mathbf{Cols}(M)\}$. Let $\mathcal{U}^*_{\dot{s}}(S) \stackrel{\text{df}}{=} \mathcal{X}_{\dot{s}}/S$ and $\mathcal{U}^*_{\dot{x}}(S') \stackrel{\text{df}}{=} \mathcal{X}_{\dot{x}}/S'$. Observe that $\mathcal{U}^*_{\dot{s}}(S) = \mathcal{U}^*_{\dot{s}}(S)///O$ and $\mathcal{U}^*_{\dot{x}}(S') = \mathcal{U}^*_{\dot{x}}(S')///O'$.

Since $\hat{s}, \hat{x} \in \mathbf{UniHom}^{\mathcal{T}}_{n,m}$ and $|\mathcal{X}_{\dot{s}}/OS| = |\mathcal{X}_{\dot{x}}/O'S'| = |M|$ it follows that $|\mathcal{O}_{\dot{s}}| = |\mathcal{O}_{\dot{x}}|$ and $|\mathcal{S}_{\dot{s}}| = |\mathcal{S}_{\dot{x}}|$. Using the chosen enumerations, choose a bijection $f : (\mathcal{X}_{\dot{s}}/S) \cup \mathcal{L}(O,S) \to (\mathcal{X}_{\dot{x}}/S') \cup \mathcal{L}(O',S')$ that satisfies the following: for $i \in \{1, \ldots, \mathbf{Rows}(M)\}, j \in \{1, \ldots, \mathbf{Cols}(M)\}$, $f(O^{\dot{s}}_i S^{\dot{s}}_j) = O^{\dot{x}}_i S^{\dot{x}}_j$, $f(OS) = O'S'$, $f(O) = O'$, $f(S) = S'$, and $f(O \vee S) = O' \vee S'$. Clearly, $(\mathcal{U}^*_{\dot{s}}(S)//O, \dot{s})$ is utility isomorphic under f to $(\mathcal{U}^*_{\dot{x}}(S')//O', \dot{x})$.

Given (1) Lemma 8, (2) the fact that $\dot{s}$ and $\dot{x}$ are $\mathcal{T}$-imaging homogenous, (3) the fact that for any $Y \in \mathcal{L}(O,S)$, $|\mathcal{X}_{\dot{s}}/Y| = |\mathcal{X}_{\dot{x}}/f(Y)|$, and (4) the fact that for any $X \in \mathcal{X}_{\dot{s}}/S$ and $Y \in \mathcal{L}(O,S)$, $X \subseteq Y$ iff $f(X) \subseteq f(Y)$[15], it follows that for any $X \in \mathcal{U}^*_{\dot{s}}(S)$ and $Y \in \mathcal{L}(O,S)$,

$$\mathcal{I}^{\mathcal{T}}_{\dot{s}}(X \parallel Y) = \frac{1}{|\mathcal{X}_{\dot{s}}/Y|} = \frac{1}{|\mathcal{X}_{\dot{x}}/f(Y)|} = \mathcal{I}^{\mathcal{T}}_{\dot{x}}(f(X) \parallel f(Y)), \quad \textit{if } X \subseteq Y, \tag{4.10}$$

and

$$\mathcal{I}^{\mathcal{T}}_{\dot{s}}(X \parallel Y) = 0 = \mathcal{I}^{\mathcal{T}}_{\dot{x}}(f(X) \parallel f(Y)), \quad \textit{otherwise}. \tag{4.11}$$

Hence, $(\mathcal{U}^*_{\dot{s}}(S), \{O,S\}, \dot{s})$ is $\mathcal{T}$-imaging isomorphic under f to $(\mathcal{U}^*_{\dot{x}}(S'), \{O',S'\}, \dot{x})$. □

Corollary 6 (PSW2 Implies U-Corr. in Hom. GWs a Func. of UMatrices for SWs in $\mathbf{UniHom}^{\mathcal{T}}_{n,m}$)**.** *For any theory* $\mathcal{T}$ *with utility correctness function* $\mathbf{U}^{\mathcal{T}}$ *and utility frame* $\mathcal{U}^{\mathcal{T}}$ *satisfying PSW2, and any* $n, m > 0$ *there is* $g^{\mathcal{T}}_{n,m} : \mathcal{M} \to \mathbb{R}$ *such that for any* $\hat{s} \in \mathbf{UniHom}^{\mathcal{T}}_{n,m}$, $\dot{s} \in \underline{h}(\hat{s})$, $O, S \in \dot{s}$, *if* $M \in \mathbf{UMat}(\dot{s}, OS)$, *then* $\mathbf{U}^{\mathcal{T}}_{\dot{s}}(OS) = g^{\mathcal{T}}_{n,m}(M)$.

[15] Remember, that the matrix-based enumeration is a utility-matrix for the small world outcomes OS and $O'S'$.

Proof. Follows trivially from Lemma 10 and the definition of PSW2. □

Discussion of Lemma 10 and Corollary 6. This is a result analogous to the previous lemma and its corollaries except here we deal with utility. In a nutshell, any two outcomes from any pair of small worlds in a $\mathbf{UniHom}^{\mathcal{T}}$ collection will have the same $\mathcal{T}$-correctness condition for utility estimation in any two grand worlds, provided only that they share the same utility matrix in those grand worlds, respectively.

Lemma 11 (PSW1&2 Imply $\mathcal{T}$-corr. for SWs in $\mathbf{UniHom}^{\mathcal{T}}_{n,m}$ in Hom. GWs a Func. of UMatrices)**.** *For any theory $\mathcal{T}$ with correctness functions $\mathbf{C}^{\mathcal{T}}$ and $\mathbf{U}^{\mathcal{T}}$ and utility frame $\mathcal{U}^{\mathcal{T}}$ satisfying PSW1 and PSW2, and for any $n, m > 0$, there is $\mathscr{C}^{\mathcal{T}}_{n,m} : \mathcal{M} \to \mathbb{R}$ such that for any $\hat{s} \in \mathbf{UniHom}^{\mathcal{T}}_{n,m}$, $\dot{s} \in \underline{h}(\hat{s})$, $O, S \in \hat{s}$, we have $\mathscr{C}^{\mathcal{T}}_{n,m}(M) = \mathbf{C}^{\mathcal{T},O}_{\dot{s}}(S) \times \mathbf{U}^{\mathcal{T},O}_{\dot{s}}(S)$ if $M \in \mathbf{UMat}(\dot{s}, OS)$.*

Proof. Choose $\mathcal{T}$ with $\mathbf{C}^{\mathcal{T}}$, $\mathbf{U}^{\mathcal{T}}$, and $\mathcal{U}^{\mathcal{T}}$ satisfying PSW1 and PSW2. Choose $n, m > 0$.

Given Corollary 5, there is $f^{\mathcal{T}} : \mathbb{I}^+ \times \mathbb{I}^+ \to [0,1]$ such that $\mathbf{C}^{\mathcal{T},O}_{\dot{s}}(S) = f^{\mathcal{T}}(n,m)$ for all $\hat{s} \in \mathbf{UniHom}^{\mathcal{T}}_{n,m}$, $\dot{s} \in \underline{h}(\hat{s})$, $O, S \in \hat{s}$. Given Corollary 6, there is $g^{\mathcal{T}}_{n,m} : \mathcal{M} \to \mathbb{R}$ such that $\mathbf{U}^{\mathcal{T},O}_{\dot{s}}(S) = g^{\mathcal{T}}_{n,m}(M)$ for any $\hat{s} \in \mathbf{UniHom}^{\mathcal{T}}_{n,m}$, $\dot{s} \in \underline{h}(\hat{s})$, $O, S \in \hat{s}$, provided that $M \in \mathbf{UMat}(\dot{s}, OS)$.

For any $n, m > 0$ we define $\mathscr{C}^{\mathcal{T}}_{n,m} : \mathcal{M} \to \mathbb{R}$ as follows: $\mathscr{C}^{\mathcal{T}}_{n,m}(M) \stackrel{\text{df}}{=} f^{\mathcal{T}}(n,m) \times g^{\mathcal{T}}_{n,m}(M)$. □

Discussion of Lemma 11. This key lemma shows that for any $\mathbf{UniHom}^{\mathcal{T}}$ collection C, or rather any pair of natural numbers (which uniquely isolate such a collection), there is a function from matrices to real numbers, such that for any matrix M, there is some x such that any outcome (in any small world in C) having M as its utility matrix in a given grand world $\dot{s}$ must have x as its $\mathcal{T}$-correctness term in $\dot{s}$.

Theorem 7 (Second Unsolvability Result)**.** *No theory meeting Constraints 4 & 5 solves PSW1–3.*

Proof. Choose theory $\mathcal{T}$ with correctness functions $\mathbf{C}^{\mathcal{T}}$ and $\mathbf{U}^{\mathcal{T}}$ and utility frame $\mathcal{U}^{\mathcal{T}}$ that satisfy PSW1 and PSW2.

Given Lemma 11 there is $\mathscr{C}^{\mathcal{T}}_{2,3} : \mathcal{M} \to \mathbb{R}$ such that for any $\hat{s} \in \mathbf{UniHom}^{\mathcal{T}}_{2,3}$, $\dot{s} \in \underline{h}(\hat{s})$, $O, S \in \hat{s}$, we have $\mathscr{C}^{\mathcal{T}}_{2,3}(M) = \mathbf{C}^{\mathcal{T},O}_{\dot{s}}(S) \times \mathbf{U}^{\mathcal{T},O}_{\dot{s}}(S)$ if $M \in \mathbf{UMat}(\dot{s}, OS)$. Given Theorem 5 there is a 2×3 real block matrix

$$\begin{bmatrix} A_1 & A_2 & A_3 \\ B_1 & B_2 & B_3 \end{bmatrix} = \begin{bmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \\ d_1 & d_2 & d_3 \end{bmatrix} = M^*$$

such that both:

1. $\mathscr{C}^{\mathcal{T}}_{2,3}(A_1) + \mathscr{C}^{\mathcal{T}}_{2,3}(A_2) + \mathscr{C}^{\mathcal{T}}_{2,3}(A_3) \geq \mathscr{C}^{\mathcal{T}}_{2,3}(B_1) + \mathscr{C}^{\mathcal{T}}_{2,3}(B_2) + \mathscr{C}^{\mathcal{T}}_{2,3}(B_3)$.
2. $c_1 + c_2 + c_3 > \max\{(a_1 + a_2 + a_3), (b_1 + b_2 + b_3)\}$.[16]

Given that $\mathcal{T}$ satisfies Constraint 4, there are $\hat{s} \in \mathbf{UniHom}^{\mathcal{T}}_{2,3}$ and $\dot{s} \in \underline{h}(\hat{s})$ such that $M^* \in \mathbf{UMat}(\dot{s}, \top)$ and there are enumerations $O_i : \{1,2\} \to \mathcal{O}_{\hat{s}}$ and $S_j : \{1,2,3\} \to \mathcal{S}_{\hat{s}}$ such that $A_k \in \mathbf{UMat}(\dot{s}, O_1 S_k)$ and $B_k \in \mathbf{UMat}(\dot{s}, O_2 S_k)$. This implies

$$\sum_{i=1}^{3} \mathbf{C}^{\mathcal{T},O_1}_{\dot{s}}(S_i) \times \mathbf{U}^{\mathcal{T},O_1}_{\dot{s}}(S_i) \geq \sum_{i=1}^{3} \mathbf{C}^{\mathcal{T},O_2}_{\dot{s}}(S_i) \times \mathbf{U}^{\mathcal{T},O_2}_{\dot{s}}(S_i).$$

But there are also $O^*_i : \{1,2,3,4\} \to \mathcal{O}_{\dot{s}}$ and $S^*_j : \{1,2,3\} \to \mathcal{S}_{\dot{s}}$ such that $O_1 = O^*_1 \vee O^*_2$, $O_2 = O^*_3 \vee O^*_4$, $S_1 = S^*_1$, $S_2 = S^*_2$, $S_3 = S^*_3$, and $M^*_{ij} = u_{\dot{s}}(O^*_i S^*_j) = \upsilon^{\mathcal{T}}_{\dot{s}}(O^*_i S^*_j)$.[17] Hence,

$$\sum_{i=1}^{3} \upsilon^{\mathcal{T}}_{\dot{s}}(O^*_3 S_i) > \max\{\sum_{i=1}^{3} \upsilon^{\mathcal{T}}_{\dot{s}}(O^*_1 S_i), \sum_{i=1}^{3} \upsilon^{\mathcal{T}}_{\dot{s}}(O^*_2 S_i)\},$$

and hence,

$$\sum_{i=1}^{3} \upsilon^{\mathcal{T}}_{\dot{s}}(O^*_3 S_i)/3 > \max\{\sum_{i=1}^{3} \upsilon^{\mathcal{T}}_{\dot{s}}(O^*_1 S_i)/3, \sum_{i=1}^{3} \upsilon^{\mathcal{T}}_{\dot{s}}(O^*_2 S_i)/3\}.$$

This, given that $\mathcal{T}$ satisfies Constraint 5 and $\dot{s}$ is $\mathcal{T}$-imaging homogenous, is equivalent to

$$\sum_{i=1}^{3} p^{\mathcal{T}}_{\dot{s}}(O^*_3 S_i) \times \upsilon^{\mathcal{T}}_{\dot{s}}(O^*_3 S_i) > \max\{\sum_{i=1}^{3} p^{\mathcal{T}}_{\dot{s}}(O^*_1 S_i) \times \upsilon^{\mathcal{T}}_{\dot{s}}(O^*_1 S_i), \sum_{i=1}^{3} p^{\mathcal{T}}_{\dot{s}}(O^*_2 S_i) \times \upsilon^{\mathcal{T}}_{\dot{s}}(O^*_2 S_i)\}$$

[16] To be clear, $a_i = M^*_{1i}$, $b_i = M^*_{2i}$, $c_i = M^*_{3i}$, and $d_i = M^*_{4i}$, for $i \in \{1,2,3\}$.

[17] The last equation holds given Assumption 1 from §1.6.

and hence,

$$\mathrm{EU}^{\mathcal{T}}_{\dot{s}}(O^*_3) > \max\{\mathrm{EU}^{\mathcal{T}}_{\dot{s}}(O^*_1), \mathrm{EU}^{\mathcal{T}}_{\dot{s}}(O^*_2)\}.$$

We observe that $O^*_3 \notin O_1$. By Assumption 4 $\dot{s} \in \mathbf{I}(\hat{s})$, hence PSW3 is not satisfied by $(\mathcal{T}, \mathbf{C}^{\mathcal{T}}, \mathbf{U}^{\mathcal{T}})$.

□

4.6 "Normal" Theories

What the unsolvability results show is that for any "normal" decision theory, regardless of how it defines correctness, there will be at least one small world $\hat{s}$ with an option that maximizes correctly estimated expected utility that is not compatible with any grand world option $O \in \dot{s}$ that maximizes expected utility in $\dot{s}$, where $\dot{s}$ is a possible idealization of $\hat{s}$. The result is thus not that there is some specific class of counterexample cases that prevent any theory from solving the problem. It's that each theory is guaranteed to have its own counterexample cases; what these are will depend only on that theory's correctness functions.

What is a "normal" theory in the relevant sense? Answer: a theory satisfying Constraints 1, 2, and 3, or else satisfying Constraints 4 and 5. Each of these two groups of constraints contains one that's a little more substantive (viz., Constraints 1 and 5). I assume not much needs to be said about the more trivial constraints; but let us briefly consider what it would mean to violate the substantive constraints.

A theory $\mathcal{T}$ that violates Constraint 1 is one where the $\mathcal{T}$-correctness term[18] for a small world outcome X at a grand world is not fixed relative to the $\mathrm{EU}^{\mathcal{T}}$-terms for grand world outcomes compatible with X. In other words, you could have two grand worlds, with the same exact layout of $\mathrm{EU}^{\mathcal{T}}$-terms in outcomes compatible with the given small world outcome, and the product of the two estimates associated with the given small world outcome could be $\mathcal{T}$-correct in one but not the other grand world. Whatever could explain this would have to be something that does not show up at all when it comes time to evaluate grand world options' rationality — otherwise it would affect the $\mathrm{EU}^{\mathcal{T}}$-term of some grand world outcome. One might point out that $\mathrm{EU}^{\mathcal{T}}$-terms can remain the same despite changes in the grand world probabilities and/or utilities, and this is true, but

[18] Which, recall, is a product of a correct credence and a correct utility estimate.

ex hypothesi these changes in probabilities and/or utilities would have no affect at all on expected utilities, or indeed on the determinants of expected utility at the level of grand world outcomes. It would be surprising to say the least to discover a way that such changes could somehow affect the $\mathcal{T}$-correctness term.

So much for Constraint 1. What about a theory that violates Constraint 5? Here we have something closer to a kind of *a priori* structural failure. Such a theory $\mathcal{T}$ will be one that assigns different agent probabilities to grand world outcomes despite not only the fact that that the (ideal) agent's prior credence is entirely evenly distributed over grand world outcomes, but that their credence $\mathcal{T}$-images into collections of outcomes in an entirely homogenous way. This would imply that there are factors relevant to the proper assignment of agent probabilities to possible outcomes that do not affect how credence is rationally distributed on any practical supposition.[19] Even if the probabilities relevant to determining expected utilities are not simply to be identified with imaging values, it seems the two phenomena should be determined by the same underlying doxastic facts about the agent, whatever those might be.

There you have my (admittedly brief) defense of the constraints on decision theories assumed in the proofs above. An adventurous reader might take up the challenge to produce a plausible theory that violates those constraints. My conjecture is that that's a dead end.

4.7 Proliferation

One might hope that small worlds with "problematic" grand world idealizations will be rare or unusual, such that — in some appropriate sense — it will be the case that in "normal" or in "most" small worlds, maximizing correctly estimated expected utility suffices for compatibility in every idealization with a grand world option that maximizes expected utility. This seems dubious. Consider some small world $\hat{s}$ with two states $\hat{S}_1, \hat{S}_2$ and two options $\hat{O}_1, \hat{O}_2$. Consider a grand world idealization $\dot{s}_1$ of $\hat{s}$ that does not refine any of the options or states of $\hat{s}$ — that is, all idealization to $\dot{s}_1$ does is add a credence and utility function. Consider another grand world idealization $\dot{s}_2$ of $\hat{s}$ with three options and two states (viz., $\hat{S}_1, \hat{S}_2$) such that $\hat{O}_1 = \dot{O}_1 \vee \dot{O}_2$ and $\hat{O}_2 = \dot{O}_3$. Regardless

[19] Including the necessary supposition, i.e., $\top$.

of how we define correct estimation, it seems reasonable to expect that the same estimates for the small world outcomes of $\hat{s}$ can be correct in $\dot{s}_1$ and $\dot{s}_2$ regardless of what the expected utility is for $\dot{O}_1$. That is, we can set the utilities for grand world outcomes compatible with $\dot{O}_1$ as high as we like while preserving the same correct estimate values for the two small world outcomes compatible with $\hat{O}_1$ that we find relative to $\dot{s}_1$. All we should need to do is set the utilities and credences for grand world outcomes compatible with $\dot{O}_2$ to compensating values — after all, correct estimation for small world outcomes is surely a kind of averaging.[20] But if we can do this trick, then nothing stops us from doing it when $\hat{O}_2$ is the one that maximizes correctly estimated expected utility. The reasoning generalizes of course to small worlds and grand worlds of all sizes.[21]

Now, even if every small world has these "problematic" idealizations it is not immediately clear how often it will happen that a concrete and realistic possible agent's ideal self can only maximize expected utility by performing a grand world option incompatible with those of the agent's small world options that maximize correctly estimated expected utility.[22] But it seems reasonable to expect the answer will not be never. Further, it is reasonable to conjecture that the more coarse the agent's representation base is relative to their ideal self's base, the likelier it will be that the small world option(s) that maximizes correctly estimated expected utility will fail to be compatible with a grand world option that maximizes — the more grand world options are agglomerated by the agent's small world options, the more room there is for the averaging which constitutes correct estimation to fail to reflect extremal values.

---

[20] It partly because they do not make this (in principle contestable) assumption that the unsolvability proofs are not much simpler.

[21] I have been assuming that the space of worlds is large but finite for simplicity, but if we suppose that it is infinite, and in particular any option can be partitioned into an arbitrary number of sub-options, then it seems reasonable to expect that literally every small world will have many idealizations where the small world option that maximizes correctly estimated expected utility fails to be compatible with a grand world option that maximizes utility. Even failing this, it seems likely to be the exceedingly rare case where a small world is such that the option(s) that maximize correctly estimated expected utility are compatible with maximizing grand world options in every grand world idealization.

[22] A small world, of course, is an abstraction, and the set of its idealizations corresponds to the ideal selves of all possible concrete agents who "instantiate" it (cf. §2.2.3).

CHAPTER 5

DIVERGING PATHS

5.1 Introduction and Guaranteed Prospective Coherence

The First Formulation of the Unrestricted Problem attempts to conjoin the following three plausible claims about small world rationality that together produce the unsolvability results:

Maximising Correct Estimation (MCE) Any small world option that maximizes correctly estimated expected utility (i.e., the sum of products by small world outcomes of correct credence and utility estimates) is rational.

Correctness Is Local (CIL): Correctness in estimates is determined for small world outcomes "locally" in accord with PSW1 and PSW2 (cf. §2.5).

Rationality Entails Compatibility (REC): Any rational small world option is compatible with some rational grand world option.

At least one of these must be rejected — but which one? Before we investigate possible answers to this question, let us take note of an interesting consequence of unsolvability — or rather, of MCE and CIL — that will be of some help in framing subsequent thinking.

It is reasonable to hope that a certain *guaranteed prospective coherence* — a guaranteed harmony between the subjectively optimal long-term plan(s) of one's ideal self on one hand and what it is rational to do here and now in the small world on the other — might be an inherent aspect of practical rationality. To see what I mean, suppose grand world options are not just refinements of small world options, but full life plans that entail ("include") the small world options they refine. Nothing in our framework is inimical to this.[1] Now, the core of the Joycean picture of deliberation

[1] Ralph Wedgwood (2011) prefers this interpretation. He writes, "To solve the problem, it is not enough to ensure that the evaluation of each 'small-world act' will be the same whether one evaluates it from the standpoint of a finely

resides in the conjunction of MCE and CIL. If we interpret grand worlds in the way just described — and we reject "abnormal" decision theories (§4.6) — MCE and CIL imply that making a perfectly rational small world decision is no proof of adherence to any ideally subjectively rational life plan. Note that this is not some feature of the discarded First Formulation, nor does it depend on REC. It is perhaps that part of the unsolvability results that persists even if we deny the First Formulation and REC — indeed even if we reject the Problem of Small Worlds entirely. It must be accepted by anyone who accepts CIL and MCE. Further, if my suggestion in §4.7 is right, and the likelihood that a randomly chosen concrete possible agent will have a "problematic" ideal self increases as we consider classes of such agents with coarser and coarser options relative to their ideal selves, then, insofar as it seems reasonable to expect one's small world options to typically be very coarse by comparison with the life plans available to one's ideal self, it seems right to suppose that it will not be a rare occurrence that our perfectly rational small world choices logically clash with our subjectively rational ideal life plans.

I am unsure what to make of this.[2] Is it benign and unsurprising, or does it threaten to force us to live with an intolerable absurdity endemic to practical rationality itself? What is clear is that it can be averted if we reject CIL or MCE (or both) — but to reject either will do violence to the Joycean conception of small world rationality.

---

individuated specification of the relevant states, or from the standpoint of a more coarsely individuated specification. What matters is not how the agent evaluates each of these small-world acts, taking them one by one. What matters is whether the overall result of the whole series of rational piecemeal decisions that the agent makes is at least roughly the same overall course of action as the result of the grand-world decision."

[2]It is instructive to compare the loss of guaranteed prospective coherence with some basic ideas from the literature on dynamic decision-making (see e.g., Hey and Lotito (2009)). To use a staple example from Hammond (1976), suppose you can take an addictive drug or abstain. If you take it you will face a subsequent choice to keep taking it or never take it again. The optimal outcome is to take it once; the worst outcome is to take it and keep taking it; the middle outcome is to never take it. Suppose you think if you take it now, you'll likely choose to keep taking it at the next choice point. What is it rational to do? What's known in this literature as a *sophisticated chooser* recognizes that their preferences might change, and evaluates their current options based in part on how they expect their future self to choose at later choice points. In the drug choice situation, the sophisticated chooser abstains, even though that is incompatible with the plan with maximal utility, i.e., to take it once and only once. Thus for proponents of sophisticated choice, the loss of guaranteed prospective coherence may come as little surprise. A competing account is known as *resolute choice*. The resolute chooser does whatever option now is consistent with what now looks to them to be the optimal path through future decision points, and then follows that path at subsequent choice points even when and if it clashes with their changed preferences. The resolute chooser takes the drug, gets addicted, and then acts contrary to their addicted desires by never taking it again. For proponents of resolute choice, perfectly rational choice must be compatible with the optimal plan; hence such theorists are liable to look on the loss of guaranteed prospective coherence as a problem.

The remainder of the chapter will go as follows. We will begin by considering two views that retain both CIL and MCE and thus aim to live with the loss of guaranteed prospective coherence. The first is a simple resolute view that for consistency rejects REC (§5.2.1). The second rejects REC but replaces it with a nearby principle (§5.2.2). The first of these responds to the loss of guaranteed prospective coherence with little more than a hard stare, while the second argues that something close enough is achievable. We next will look at two approaches that avoid the loss of guaranteed prospective coherence. The first rejects CIL but keeps MCE (§5.3.2). The second rejects MCE but keeps CIL (§5.3.3). I conclude this part of the dissertation by taking stock and offering some perspective on these divergent paths (§5.4). Some reasons will be given for favoring a pluralistic approach to the norms governing the relationship between small world rationality and grand world expected utility.

5.2 Relinquishing Guaranteed Prospective Coherence

5.2.1 First Path: Just Maximize Correctly Estimated EU

This first view is the simplest of them all. It accepts the loss of prospective coherence and does not try to take any of the sting out of it. Instead, it simply rejects REC and calls it a day. In fact, it rejects this weaker principle:

REC⁻: What it is rational to do in a small world depends on what it is rational to do in its grand world idealization.

More carefully, this first view says what it is rational to do in the small world does not depend on what it is rational to do in the grand world. On this view, the grand world is involved in small world rationality only as a determinant of estimate correctness; the question of which grand world options maximize expected utility is looked on as of no small world rational significance. All that matters in the small world is maximizing correctly estimated expected utility; small world actions that do this are rational.

Does something worthy of being called Problem of Small Worlds still arise? Not really. All that remains of the Problem of Small Worlds on this view is the requirement to define estimate correctness in a way that satisfies what we've been calling PSW1 and PSW2. But beyond meeting

those two "locality" conditions (§2.5) the only measure of whether estimate correctness has been defined well or badly will presumably be intuition. Since grand world options' rationality is of no significance, there is no general desideratum to show that maximizing correctly estimated expected utility properly aligns with grand world expected utility.

The attractiveness of this perspective lies in its honoring the intuitive core of Joyce's conception of small world deliberation and rationality — namely that in the small world we are making localized estimates of what our considered judgments would be as to the likelihoods and desirabilities of the outcomes in our small world representation of our choice, and what it is rational to do is to maximize correctly estimated expected utility. It is unattractive insofar as we think the loss of guaranteed prospective coherence looks more like bad theory than a genuine aspect of practical rationality. One might also object to the total rejection in the small world of the rational significance of grand world options; can the facts as to what it would be rational for our ideal selves to do really be simply irrelevant in the small world? We will return to this question in §5.4.

5.2.2 Second Path: Maximize Average GWEU

How close can we get to REC if we accept both CIL and MCE? Related question: is there something close to prospective coherence that can be guaranteed by rational small world choice, if we accept CIL and MCE? Let us consider the possibility that the expected utilities of grand world options *are* relevant in the small world, but the constitutive rational aim in the small world is *not* compatibility with a maximal grand world option. How might this work? An intriguing idea arises fairly naturally. Any small world option is a disjunction of grand world options.[3] That is, perhaps what we should care about is the *average expected utility* of the grand world options compatible

[3] *So long as we retain Assumption 2 (Recoverability),* one might add. But how could Recoverability fail? Well, perhaps some distinction between behaviors drawn in the small world might be practically otiose in the grand world, once things are precisified. That is, there might be a division between small world behaviors that, once grand world options are in view, amounts to a division between two in principle separable behaviors whose outcomes have exactly the same pattern of utilities and probabilities in the grand world. For instance, in the small world we might think in terms of RUN versus DON'T RUN, whereas idealizing to the grand world might require seeing things in terms of INTENSE AEROBIC EXERCISE, MILD AEROBIC EXERCISE, and DON'T EXERCISE, where the middle option can be done by running or not by running. In this case we can, if we like, suppose that any idealization must retain the small world distinction and have both MILD AEROBIC EXERCISE & RUN and MILD AEROBIC EXERCISE & NOT RUN, say, or we might say that this is unnecessary. If what we're interested in is the average expected utility of grand world options compatible with a small world option, we get the same results whether we insist that small world divisions be retained or not.

with each small world option in a given small world — perhaps rationality requires us to perform a small world option that maximizes that average. Since this does not require compatibility with a grand world option that maximizes expected utility (an average of a subset of values $V \subset S$ can be a maximal average out of a set of such averages despite lacking a maximal value from S) we sidestep the unsolvability results though we still have to accept the loss of guaranteed prospective coherence.

The Unrestricted Problem of Small Worlds can be reformulated on this picture. It becomes a challenge of proving that for any $\hat{s}$ and $\dot{s} \in \mathbf{I}(\hat{s})$, if $O \in \hat{s}$ maximizes correctly estimated expected utility, then it disjoins a set of grand world options $O^* \in \dot{s}$ with maximal average expected utility in $\dot{s}$. One thing that requires clarification is the sense in which we intend the word 'average'. The right answer, I believe, is to use theories' imaging functions.[4] Here then is the Second Formulation of the Unrestricted Problem of Small Worlds:

5.2.2.1 The Unrestricted Problem of Small Worlds for Theory $\mathcal{T}$, Second Formulation

Prove there are *correctness functions* $\mathbf{C}^{\mathcal{T}} : \dot{s} \mapsto \big(\mathbf{SWO}(\dot{s}) \times \mathbf{SWS}(\dot{s}) \to [0,1]\big)$, $\mathbf{U}^{\mathcal{T}} : \dot{s} \mapsto \big(\mathbf{SWO}(\dot{s}) \times \mathbf{SWS}(\dot{s}) \to \mathbb{R}\big)$, a utility frame $\mathcal{U}^{\mathcal{T}} : \dot{s} \mapsto [S \in \mathbf{SWS}(\dot{s}) \mapsto \mathcal{U}^{\mathcal{T}}_{\dot{s}}(S)]$, such that:

(PSW1) Same as before.

(PSW2) Same as before.

(PSW3‡) For any $\hat{s}$ and $\dot{s} \in \mathbf{I}(\hat{s})$, if $\mathbf{c}_{\hat{s}}$ and $\mathbf{u}_{\hat{s}}$ are such that for all $O, S \in \hat{s}$, $\mathbf{c}_{\hat{s}}(OS) = \mathbf{C}^{\mathcal{T},O}_{\dot{s}}(S)$ and $\mathbf{u}_{\hat{s}}(OS) = \mathbf{U}^{\mathcal{T},O}_{\dot{s}}(S)$ then any option $O \in \mathcal{O}_{\hat{s}}$ that maximizes $\mathbf{EstEU}_{\hat{s}}$ must also maximize $\sum_{O' \in \dot{s}} \mathcal{I}^{\mathcal{T}}_{\dot{s}}(O' \parallel O)\,\mathrm{EU}^{\mathcal{T}}_{\dot{s}}(O')$.[5]

5.2.2.2 An EDT Solution

EDT can solve the Second Formulation. It isn't hard to see why. Given EDT's definitions for correct estimation, correctly estimated expected utility for a given small world option is just

[4] Insofar as theories' imaging functions capture their attendant conception of how to adjust credence under practical supposition they seem unavoidable here.

[5] Here "maximize" is meant to be relative to the small world options in $\hat{s}$. I.e., for O to maximize is for there to be no $O^* \in \hat{s}$ such that $\sum_{O' \in \mathcal{O}_{\dot{s}}/O} \mathcal{A}^{\mathcal{T},O}_{\dot{s}}(O')\,\mathrm{EU}^{\mathcal{T}}_{\dot{s}}(O') < \sum_{O' \in \mathcal{O}_{\dot{s}}/O^*} \mathcal{A}^{\mathcal{T},O^*}_{\dot{s}}(O')\,\mathrm{EU}^{\mathcal{T}}_{\dot{s}}(O')$.

the grand world expected value of that option, which is also expressible as a conditional-credence-weighted average of the expected values (viz., the EDT expected utilities) of the grand world options compatible with the small world option.

Theorem 8 (EDT Solves the Strong Unrestricted Problem of Small Worlds, 4th Form)**.** *EDT Solves the Strong Unrestricted Problem of Small Worlds, Fourth Formulation, with:*

- $\mathbf{C}^{\mathrm{EDT}}_{\dot{s}} \stackrel{\mathrm{df}}{=} O \in \mathbf{SWO}(\dot{s}) \mapsto (S \in \mathbf{SWS}(\dot{s}) \mapsto c_{\dot{s}}(S \mid O))$,
- $\mathbf{U}^{\mathrm{EDT}}_{\dot{s}} \stackrel{\mathrm{df}}{=} O \in \mathbf{SWO}(\dot{s}) \mapsto (S \in \mathbf{SWS}(\dot{s}) \mapsto \mathbf{V}_{\dot{s}}(OS))$, *and*
- $\mathcal{U}^{\mathrm{EDT}}_{\dot{s}} \stackrel{\mathrm{df}}{=} S \in \mathbf{SWS}(\dot{s}) \mapsto \mathcal{X}_{\dot{s}}/S$.

Proof. Given Lemmas 2 and 3, it suffices to show that EDT solves PSW3‡. Choose $\hat{s}$ and $\dot{s} \in \mathbf{I}(\hat{s})$ such that $\mathbf{c}_{\hat{s}}(OS) = c_{\dot{s}}(S \mid O)$ and $\mathbf{u}_{\hat{s}}(OS) = \mathbf{V}_{\dot{s}}(OS)$ for all $OS \in \mathcal{X}_{\hat{s}}$. Now, for any $O \in \hat{s}$,

$$\mathbf{EstEU}^{\mathrm{EDT}}_{\hat{s}} = \sum_{S \in \hat{s}} c_{\dot{s}}(S \mid O)\mathbf{V}_{\dot{s}}(OS) = \sum_{O' \in \dot{s}} c_{\dot{s}}(O' \mid O)\mathbf{V}_{\dot{s}}(O') = \sum_{O' \in \dot{s}} \mathcal{I}^{\mathrm{EDT}}_{\dot{s}}(O' \parallel O)\,\mathrm{EU}^{\mathrm{EDT}}_{\dot{s}}(O').$$

□

5.2.2.3 Trouble for Lewisians

Our Lewisian definition of correct utility estimates (§3.7) makes use of dependency hypotheses that are defined for small world options:

$$\forall \dot{s}, O \in \mathrm{SWO}(\dot{s}), S \in \mathrm{SWS}(\dot{s}) \quad \mathbf{U}^{\mathrm{Lew},O}_{\dot{s}}(S) \stackrel{\mathrm{df}}{=} \sum_{K \in \mathcal{K}^{O}_{u_{\dot{s}}}} c_{\dot{s}}(K \mid S)\mathbf{V}_{\dot{s}}(OK). \tag{5.1}$$

Given the definition of correctness for credence estimates (cf. Cor. 2 in §3.4), correctly estimated expected utility simplifies thus to:

$$\forall \hat{s}, \dot{s} \in \mathbf{I}(\dot{s}), O \in \hat{s}, \quad \mathrm{CorEstEU}^{\mathrm{Lew}}_{\hat{s},\dot{s}}(O) = \sum_{K \in \mathcal{K}^{O}_{u_{\dot{s}}}} c_{\dot{s}}(K)\mathbf{V}_{\dot{s}}(OK), \tag{5.2}$$

which is just the canonical definition of Lewisian expected utility except that it's defined for small world options relative to grand worlds. Is there any way to understand this as an average of Lewisian grand world expected utilities?

To answer this, we need to specify the Lewisian imaging function, since averaging grand world expected utilities will go by imaging on small world options. The natural definition for the Lewisian uses dependency hypotheses:

$$\forall P, Q \in \mathcal{P}(\Omega), \quad \mathcal{I}_s^{\text{Lew}}(Q \parallel P) \stackrel{\text{df}(?)}{=} \sum_K c_s(K) c_s(Q \parallel KP). \tag{5.3}$$

But which dependency hypotheses? Noting that any contingent proposition can be treated as an option in a binary partition, the simplest approach for any $P, Q \in \mathcal{P}(\Omega)$ is to just let the set of dependency hypotheses for $\mathcal{I}_s^{\text{Lew}}(Q \parallel P)$ be $\mathcal{K}_s^P$.[6] Thus, the relevant average of (Lewisian) grand world expected utilities will look like this:

$$\sum_{\dot{O} \sqsubseteq O} \sum_{K \in \mathcal{K}_s^O} c_{\dot{s}}(K) c_{\dot{s}}(\dot{O} \mid OK) \sum_{\dot{K} \in \mathcal{K}_{\dot{s}}} c_{\dot{s}}(\dot{K}) \mathbf{V}_{\dot{s}}(\dot{O}\dot{K}). \tag{5.4}$$

This then is what the correctly estimated expected utility of O would need to equal for the Lewisian to have a solution to PSW3‡.

If we expand the expected value term in our Lewisian definition of correctly estimated expected utility (5.2), we get something superficially similar to (5.4):

$$\sum_{\dot{O} \sqsubseteq O} \sum_{K \in \mathcal{K}_s^O} c_{\dot{s}}(K) c_{\dot{s}}(\dot{O} \mid OK) \mathbf{V}_{\dot{s}}(\dot{O}K).[7] \tag{5.5}$$

But (5.4) and (5.5) will not, as a rule, be equal for the same option and grand world state, since it will sometimes be the case that:

$$\left[\sum_{\dot{K} \in \mathcal{K}_{\dot{s}}} c_{\dot{s}}(\dot{K}) \mathbf{V}_{\dot{s}}(\dot{O}\dot{K}) \right] \neq \mathbf{V}_{\dot{s}}(\dot{O}K). \tag{5.6}$$

If this is not immediately obvious, note that $\dot{O}K$ will always, for any $\dot{s}$ we pick, be a utility level proposition in $\dot{s}$, since OK is a utility level proposition in $\dot{s}$ (K being a dependency hypothesis relative to $\dot{s}$ and a partition containing O) and $\dot{O}$ is a subset of O. But the LHS of (5.6) will for

[6] This uses P in the way discussed in the final paragraph of §3.7.

[7] Of course, $O\dot{O}K = \dot{O}K$.

many $\dot{s}$ be a genuine average of different utility levels, i.e., those of the $\dot{O}\dot{K}$ for the various $\dot{K} \in \mathcal{K}_{\dot{s}}$, and these utility levels may be perfectly arbitrary relative to $\dot{O}K$. Note also that trying to fix the RHS by reverting to our original definition of correct Lewisian utility estimation with its use of dependency hypotheses for grand world options rather than small world options (cf. §3.7) will do nothing to avert (5.6), as we would only succeed in replacing the dependency hypothesis on the RHS with a grand world dependency hypothesis.

Lewisian CDT appears to be in trouble. It seems the most natural way to regroup is to rethink the correctness conditions for utility estimates. This means reconsidering the role played by Lewisian centers of mass. This notion amounts to an average of expected values ($\mathbf{V}$'s) of all conjunctions of an option with dependency hypotheses — that is, an average of utility levels — where the weights in the average are credences in the dependency hypotheses conditional on a state. So, using the state as evidence about the causal background, we obtain a measure of the value of the option conditional on that causal background. The trouble is this is not always equal to an average of Lewisian expected utilities of grand world options. Why not then simply define correct Lewisian utility estimation in terms of such an average:

$$\mathbf{U}_{\dot{s}}^{\mathrm{Lew},O}(S) \stackrel{\mathrm{df}(?)}{=} \sum_{K \in \mathcal{K}_{\dot{s}}^{O}} c_{\dot{s}}(K \mid S) \sum_{\dot{O} \subseteq O} c_{\dot{s}}(\dot{O} \mid OK) \sum_{\dot{K} \in \mathcal{K}_{\dot{s}}} c_{\dot{s}}(\dot{K}) \mathbf{V}_{\dot{s}}(\dot{O}\dot{K}). \tag{5.7}$$

An immediate result of this is that correctly estimated Lewisian expected utility reduces to (5.4).

Theorem 9. *Lewisian CDT (Modified with Eq. 5.7) Solves the Weak Unrestricted Problem of Small Worlds, Fourth Formulation, with:*

- $\mathbf{C}_{\dot{s}}^{\mathrm{Lew}} \stackrel{\mathrm{df}}{=} O \in \mathbf{SWO}(\dot{s}) \mapsto (S \in \mathbf{SWS}(\dot{s}) \mapsto c_{\dot{s}}(S))$,
- $\mathbf{U}_{\dot{s}}^{\mathrm{Lew}} \stackrel{\mathrm{df}}{=} O \in \mathbf{SWO}(\dot{s}) \mapsto (S \in \mathbf{SWS}(\dot{s}) \mapsto \sum_{K \in \mathcal{K}_{\dot{s}}^{O}} c_{\dot{s}}(K \mid S) \sum_{\dot{O} \subseteq O} c_{\dot{s}}(\dot{O} \mid OK) \sum_{\dot{K} \in \mathcal{K}_{\dot{s}}} c_{\dot{s}}(\dot{K}) \mathbf{V}_{\dot{s}}(\dot{O}\dot{K}))$.

Proof. Trivial. □

This totally reconceives correct utility estimation from the Lewisian perspective — centers of mass are abandoned. Perhaps a Lewisian can live with that. There is, however, another difficulty that will be harder to swallow. This definition of correct utility estimation fails PSW2. Thus, the Lewisian must give up on CIL.

Hopefully this is plain without a proof. The utilities on which correct utility estimation for an option-state pair depend under this definition include utilities at worlds incompatible with the state. For a deliberator focused on a particular small world option and state to get the utility estimate for that combination right will thus require somehow mentally tracking the expected utilities of all the grand world options compatible with whatever small world option is in focus, and those quantities depend on what's going on at all grand world states, not just ones compatible with the small world state in view. I see no alternative solution for a Lewisian that salvages PSW2. Hence, my reluctant conclusion is that the Lewisian has no solution to the Strong Unrestricted Problem Second Formulation.

5.2.2.4 Imaging Theories

EDT's definition of expected utility and its approach to the Problem of Small Worlds flow from its conception of imaging: conditional credence. Lewisian CDT is not imaging-centric, but other versions of CDT are (see e.g., Sobel (1978, 1986), Joyce (1999)). Is the trick to solving the Second Formulation just basing decision theory on imaging? Let us call a theory $\mathcal{T}$ an *Imaging Theory* iff

- $\forall s; O \in s, \mathrm{EU}_s^{\mathcal{T}}(O) \stackrel{\mathrm{df}}{=} \sum_{S \in \mathcal{S}_s} \mathcal{I}_s^{\mathcal{T}}(S \parallel O) \sum_{w \in S} \frac{\mathcal{I}_s^{\mathcal{T}}(w \| O)}{\mathcal{I}_s^{\mathcal{T}}(S \| O)} u_s(w) = \sum_w \mathcal{I}_s^{\mathcal{T}}(w \parallel O) u_s(w)$.
- $\forall \hat{s}; O, S \in \hat{s}; \dot{s} \in \mathbf{I}(\hat{s}), \mathbf{C}_{\dot{s}}^{\mathcal{T},O}(S) \stackrel{\mathrm{df}}{=} \mathcal{I}_{\dot{s}}^{\mathcal{T}}(S \parallel O)$.
- $\forall \hat{s}; O, S \in \hat{s}; \dot{s} \in \mathbf{I}(\hat{s}), \mathbf{U}_{\dot{s}}^{\mathcal{T},O}(S) \stackrel{\mathrm{df}}{=} \sum_{w \in S} \frac{\mathcal{I}_s^{\mathcal{T}}(w \| O)}{\mathcal{I}_s^{\mathcal{T}}(S \| O)} u_s(w)$.[8]

Adopting an imaging theory is not a surefire solution to the Second Formulation. To solve the Second Formulation, imaging on a grand world option $\dot{O} \subseteq \hat{O}$ cannot result in different credence ratios between worlds compared to imaging on the small world $\hat{O}$ that includes it. That is, if for any $w, w' \in \dot{O}$, $\frac{\mathcal{I}(w \| \dot{O})}{\mathcal{I}(w' \| \dot{O})} = \frac{\mathcal{I}(w \| \hat{O})}{\mathcal{I}(w' \| \hat{O})}$, then the only difference between the credence assigned to a world

[8] A few words on my characterization of "imaging" theories seem in order. One might wonder why the value term in imaging theories is $\sum_{w \in S} \frac{\mathcal{I}_s^{\mathcal{T}}(w \| O)}{\mathcal{I}_s^{\mathcal{T}}(S \| O)} u_s(w)$ rather than $\sum_w \mathcal{I}_s^{\mathcal{T}}(w \parallel OS) u_s(w)$. After all, the latter is what EDT uses. Or is it? Sort of. EDT is special in that the two formulas for it are equivalent, as is easily checked. EDT therefore cannot be used as a standard to decide between them. We choose the formula that only involves imaging on O rather than on OS because the latter approach would trivially result in a loss of partition invariance if a theory's imaging function fails to preserve ratios in prior credences ("priors") between worlds. Why then use the ratio instead of just a "dot product" of utilities with credences, imaged on the given option, in worlds in the given state? Simply put, because normalizing seems intuitively required. The utility value of a state given an option shouldn't be low simply because that state receives low credence imaged on the option.

imaged on one versus the other option will be the normalizing factor $\mathcal{I}(\dot{O} \parallel \hat{O})^{-1}$, and hence for any $w \in \dot{O} \subseteq \hat{O}$, $\mathcal{I}(w \parallel \hat{O}) = \mathcal{I}(\dot{O} \parallel \hat{O})\mathcal{I}(w \parallel \dot{O})$ and the Second Formulation of the problem is solved. If on the other hand imaging on a grand world option can result in different ratios of credence than imaging on the small world that contains it, the Second Formulation cannot be solved.[9] This implies that no imaging theories other than EDT will be able to solve the Second Formulation. Adoption of the Second Formulation of the Problem of Small Worlds and the attendant conception of small world rationality is thus a coup for EDT. Stepping back, we may observe that there will be no difference at the level of case verdicts between this conception of small world rationality and the previous one (§5.2.1), at least insofar as EDT is used, since for EDT there is no difference between maximizing correctly estimated expected utility and maximizing average grand world expected utility; the difference is thus purely conceptual.

5.3 Ensuring Guaranteed Prospective Coherence

Suppose we say, *give us our REC — we must have guaranteed prospective coherence.* The picture of rationality we adopt will depend on how we answer two questions. The first is whether mere compatibility with a rational grand world option is all that small world rationality requires, and nothing stronger. The second is whether we wish to accept CIL or MCE or neither. The first question, I believe, has a definite answer: mere compatibility is not enough. Instead, a proponent of this perspective should say that what rationality requires is maximal subjective likelihood that one will perform a rational grand world option. I will argue for this below. The second question, on the other hand, will be a genuine theoretical choice point and I expect intuitions to differ among fans of REC and guaranteed prospective coherence. There are three alternatives. The most radical of the three rejects both CIL and MCE. This seems to me quite implausible and I will have nothing to say about it.[10] Setting that aside, we are left with a choice between MCE and CIL. If we opt

[9] It is interesting to note that this requirement is stronger than partition invariance, as it is easy to show that imaging theories, as we have defined them, are general partition invariant.

[10] There is perhaps an even more radical view worth pointing out, on which what rationality requires in the small world is performance of a grand world option that maximizes (grand world) expected utility. On this view, a person could be rationally required to perform an option they are not and will not represent as such to themselves prior to the time of decision.

for MCE, we get a view where making estimates and acting in accordance with correctly estimated EU still has a place in our conception of small world rationality, but must be understood in a new way. If instead we choose CIL, then "local" Joycean estimation in small world deliberation is retained and what is lost is the idea that we are rationally bound to maximize correctly estimated EU. In this section we will look at both of these alternatives more closely; the first alternative will be discussed in §5.3.2, the second in §5.3.3. Before we get to either, however, let's first consider what grounds there are for my assertion above that proponents of this overall perspective — which seeks to ensure guaranteed prospective coherence — should think small world rationality requires not just compatibility with some rational grand world option, but maximal subjective likelihood of performing a rational grand world option.

5.3.1 Maximizing the Likelihood of Maximizing

When one acts, a concrete token of behavior occurs. That is, when one chooses a small world option, one must, in carrying it out, "select," as it were, some or other concrete "realization" of that abstract small world option. Hence, one's act either will or will not in fact be an instance of a permissible grand world option. Bearing these platitudes in mind, it seems easy to intuit that performing a small world option that is merely compatible with a rationally permissible grand world option is not in itself a rational good of any sort. Instead, what rationality demands should be performing a small world option with maximal likelihood of being "realized" as a rational grand world option. So goes a fairly intuitive argument. Why might one be led to the contrary view, that compatibility with a rational grand world option is a good of small world rationality? That view seems to depend on the idea that the determinacy of grand world options comes from their being mere sequences of small world acts: precise in the long run, not in the short run. This thought faces pressure from two directions. First, grand world options can be what one might call *act precisifications.* For instance, if my small world option is go for a run, a precisification of that could be to go for a run in my Nikes. If running in my Nikes is what grand world rationality demands and I go for a run in my Asics, it might seem a rather senseless defense of my rationality to insist that I did after all run and running is logically compatible with running in my Nikes. But even if we set this point aside and focus just on cases where the grand world options are *long run extensions*

of small world options, there is pressure to say mere compatibility with a grand-world-rational sequence of small world acts is insufficient for small world rationality.

Consider a small world choice between A and $\neg A$, where each is compatible with a different grand-world-rational time sequence of small world options, say $A, B, \neg C, \dots$ and $\neg A, B, C, \dots$. That is, suppose in the relevant grand world $\dot{s}$, $\mathrm{EU}_{\dot{s}}(A, B, \neg C, \dots) = \mathrm{EU}_{\dot{s}}(\neg A, B, C, \dots) = n$, where n is maximal. Should we — or rather, someone engaged in an effort to preserve guaranteed prospective coherence — conclude straightaway that both small world options are rational? It seems not. It might be more likely that one will enact a grand-world-rational sequence if one does A. That is, it might be that

$$\mathcal{I}_{\dot{s}}(\neg A, B, C, \dots \| \neg A) < \mathcal{I}_{\dot{s}}(A, B, \neg C, \dots \| A). \tag{5.8}$$

If our guiding intuition is that rationality should aim at the performance of rational grand world options, then shouldn't one be required to maximize the likelihood — understood via imaging, as seems appropriate — that one enacts such an option, rather than aim for mere compatibility? That is, given (5.8), isn't A rationally required?[11]

Such are my arguments that anyone seeking to defend guaranteed prospective coherence should maintain that small world rationality requires maximizing subjective likelihood of performing a rational grand world option. A nice way to express this view uses *MaxEU$^{\mathcal{T}}$ propositions*. For a given theory $\mathcal{T}$ and small world $\dot{s}$, we have:

$$\mathrm{MaxEU}^{\mathcal{T}}_{\dot{s}} \stackrel{\mathrm{df}}{=} \bigvee\{O \in \dot{s} : \neg(\exists O' \in \dot{s})(\mathrm{EU}^{\mathcal{T}}_{\dot{s}}(O) < \mathrm{EU}^{\mathcal{T}}_{\dot{s}}(O'))\}.$$

The $\mathrm{MaxEU}^{\mathcal{T}}$ proposition for $\dot{s}$ is thus the disjunction of grand world options of $\dot{s}$ that maximize $\mathcal{T}$-expected-utility in $\dot{s}$. Thus, the central assumption is just this:

[11] One conceivable reply to this argument is that it should not be considered part of one's rational duty to make it likely that one's future behavior will conform to a rational grand world life plan, as that illicitly treats one's future self as a mere part of the world rather than as one's own agential self, or something. This seems dubious. Common sense seems to dictate that it is rational to act now in ways one deems most likely to cause one's future self to make what one now judges to be rational future choices. But regardless of how the dialectic on this point plays out, I can see no good reply to the first argument.

Maxing Likelihood of MaxEU (MLM): According to theory $\mathcal{T}$, small world option $O \in \hat{s}$ is rational relative to grand world $\dot{s}$ if and only if there is no $O' \in \hat{s}$ such that

$$\mathcal{I}_{\dot{s}}(\text{MaxEU}^{\mathcal{T}}_{\hat{s}} \,\|O) < \mathcal{I}_{\dot{s}}(\text{MaxEU}^{\mathcal{T}}_{\hat{s}} \,\|O').$$[12]

5.3.2 Third Path: Radical Estimation

Now that the core assumption of this perspective (i.e., MLM) is clearly before us, we can turn to consider the shape the view takes depending on whether MCE or CIL is retained. In this subsection we will examine the simpler of the two routes, namely the view that MCE is to be retained, CIL rejected.[13] On this view small world rationality still means maximizing correctly estimated expected utility, but estimates are not made individually correct or incorrect by features of the grand world "logically local" to the small world outcomes themselves (§2.5). Instead, correctness is determinate only at the level of estimated expected utility itself. That is to say, the estimates for outcomes of a given small world option O are $\mathcal{T}$-correct in a grand world $\dot{s}$ when and only when the estimated expected utility for that option equals $\mathcal{I}^{\mathcal{T}}_{\dot{s}}(\text{MaxEU}^{\mathcal{T}}_{\hat{s}} \,\|\, O)$.[14,15] Stepping back, we observe that

[12] As explained in §5.2.1 the double relativity is just for presentation. The relativity to theory is meant to be replaced with reference to whatever the correct theory of expected utility is, and the relativity to grand world is meant to be replaced in a given concrete decision-making context by the relevant grand world in that context.

[13] Perhaps it bears reminding that we cannot have MLM, MCE and CIL. The unsolvability results show that if we have PSW1 and PSW2 (i.e., if we have CIL) then for any normal decision theory there will be cases where a small world option maximizes correctly estimated expected utility but is incompatible with any grand world option that maximizes grand world expected utility. The cases are ones with imaging-homogenous grand worlds; hence in these cases the small world option that maximizes correctly estimated expected utility will fail to maximize imaged likelihood of a grand world option with maximal grand world expected utility—its likelihood will be zero, but the likelihood for another option will be non-zero.

[14] One could go further and suppose that the agent only makes one estimate per small world option, i.e., an estimate, roughly speaking, of how likely it is (in their grand world) that they will maximize expected utility given that they do the given small world option. This would call for a revision of our deliberation model (§**??**). Instead of combining estimate functions for small world outcomes with a representation base, we would model a small world decision situation with a single estimate function defined for the options in a representation base.

[15] Does this mean that on this picture the process of rational deliberation cannot be one of breaking down a choice situation into isolated logical pieces, making estimates concerning those pieces, performing a calculation with those estimates, and maximizing the result? No, there is a way to still make sense of all that — but none of it is to be regarded as *rationally essential*. Let's consider how this would go on EDT. Suppose I am deliberating about whether to go for a run or not, and I am considering two possible states of the world, rain and no rain. I don't have to — indeed, shouldn't — try to estimate what credences and utilities "grand-world-me" would assign to my four small world outcomes. Instead I ask myself, if I run and it rains, how likely is it that I will maximize grand world expected utility, and likewise, how likely is it that I will maximize grand world expected utility if I run and it doesn't rain.

as with the first perspective (§5.2.1) — which retained CIL and MCE but rejected REC — this perspective does away with the Problem of Small Worlds. But unlike the first perspective, this one also avoids the loss of guaranteed prospective coherence. Maximize correct estimation on this view and not only are you guaranteed to have chosen a small world option compatible with a rational grand world option, you've maximized your chances at performing such an option. What is the downside? Simply put, this view repudiates a deeply intuitive, commonsense understanding of rational small world deliberation, namely, that it constitutively involves assigning likelihoods and, in particular, desirability values to outcomes in a familiar sense, i.e., to results of determinate choice performed in determinate worldly context, considered in logical isolation from alternative such results in alternative such contexts. It gives up, further, on any point in directly applying the traditional concepts of decision theory (outcomes, utilities, etc.) to everyday choice. Instead of these things, it is advised that we adopt a strikingly "philosophical" conception of small world deliberation, whereby rationality requires directing one's thought towards the hypothetical options one's ideal self would choose between, and aiming to bring about one of those that would maximize expected utility for that ideal self.

5.3.3 Fourth Path: Pluralism

Suppose we want to retain the core of the Joycean picture of deliberation while avoiding the loss of guaranteed prospective coherence and while retaining MLM (§5.3.1). To do this, we must reject MCE, and retain CIL and REC. I can think of three approaches to rendering this combination of principles intelligible:

I can take these likelihoods, which are to be understood as estimates of grand world imaging values — conditional credences, on EDT — weight them by the estimated likelihoods — unconditional credences — I assign to rain and to no rain respectively, sum the products, and get an overall estimated likelihood of maximizing grand world expected utility for running. Notably, if my estimates are correct, then the result will be just the grand world credence that I perform a grand world option that maximizes expected utility conditional on my going for a run. I then would do the same thing for not running and choose the one with the higher estimate.

We can make this work on a causalist account if we adopt a causalist imaging theory and assume that conditional supposition goes by way of causalist imaging rather than conditional credence (cf. §5.2.2.4). Instead of asking how likely it is that I will maximize grand world expected utility if I run and it rains, on the causalist account I should presumably ask myself, on one hand, if I run, how likely is the conjunction of rain and my maximizing grand world expected utility, and on the other, if I run, how likely is the conjunction of no rain and my maximizing grand world expected utility. Then I sum those two estimates.

Distinguish Rationality and Blameworthiness (DRB): We maintain MLM and claim that while maximizing correctly estimated expected utility — which is understood "locally" in accord with PSW1 and PSW2 — may not result in rational small world choice, it at least renders one's choice rationally blameless.

Degrees of Rationality (DOR): The two forms of maximization — i.e., maximizing correctly estimated expected utility and maximizing likelihood of maximizing in the grand world — each suffice for some *degree* of small world rationality.[16]

Deep Pluralism about Rationality (DPR): We maintain that there are two equally legitimate species of small world rationality and neither takes precedence, and that hence our evaluative attitudes must be trained to respect this deep pluralism.[17]

I have grouped these views together under the heading of *pluralism*, since each in some way commits to the idea that small world deliberation can have multiple legitimate aims. In particular, all three in some way suggest that aiming to maximize correctly estimated expected utility in the small world — understood in the "local" Joycean sense via PSW1 and PSW2 — is rationally legitimate. According to DRB, if we succeed in this aim, we are at least rationally blameless. According to DOR, success at least means having achieved some degree of rationality. According to DPR, success means having acted rationally in a sense of 'rationally' that is equally as legitimate as the alternative, prospective coherence guaranteeing sense corresponding to MLM.[18] In view of this commonality, all three approaches face a pair of distinctive objections. The first goes like this: *Shouldn't the aim of small world deliberation just be to act fully rationally? On these views, isn't the centrality of that aim getting overlooked?* Thus the first objection is that according to these views, the aim of rational

[16]This basic idea can be worked out in different ways. To avert the loss of guaranteed prospective coherence, we can maintain that while maximizing correctly estimated expected utility guarantees some non-zero degree of small world rationality, *full* small world rationality always requires maximizing subjective likelihood of maximizing grand world expected utility.

[17]This route claims that guaranteed prospective coherence is one sort of rational value, and is secured by a corresponding concept of small world rationality, but it's not the only rational value.

[18]Notably, DOR and DPR each allow for the revision and reintroduction of MCE in attenuated form. The second requires exchanging 'is rational' in MCE for something like 'is rational to some degree'. (REC would need a similar revision.) The third view requires exchanging 'is rational' for some subscripted notion. REC would also need subscripting, but with a different subscript.

deliberation seems detachable from the pursuit of fully rational action — deliberation instead seems something one does to avoid blameworthiness, or to secure some merely partial degree of rationality, or to act rationally in one but not every sense. The second objection is one that die-hard supporters of MLM and guaranteed prospective coherence — that is, fans of the approach presented in the previous subsection — will want to raise. It goes like this: *What has an agent who maximizes correctly estimated expected utility really accomplished? If maximizing correctly estimated expected utility has no bearing on whether one's choice will lead to rational grand world action, why should we grant that an agent who maximizes correctly estimated utility is immune to rational criticism, or that they have achieved some non-zero degree of rationality, or that they've acted perfectly rationally in any legitimate sense of "rational"?* The second objection is thus a challenge to account for the supposed rational good that results from maximizing correctly estimated expected utility. If this seems strange, remember that it arises in a setting where commitment to MLM as a rational norm is common ground; it therefore seems most forceful against DRB, where, of the three views, that commitment is strongest.

In response to the first objection, which challenges these views to defend their construal of the aim of small world deliberation, I think one can say the following. An agent deliberating in the Joycean small world fashion aims to bring about a desirable outcome. That is, in just the same way as a grand world agent, a small world agent is engaged in the evaluation of options with a view to *outcomes* — the only difference is in the precision or density of detail in their thought. Small world deliberators thus aim at the realization of desirable small world outcomes.[19] Given the impossibility of directly choosing outcomes, the small world deliberator aims at maximizing subjective probability of a desirable outcome, in the sense made precise by the definition of correctly estimated expected utility. The point here is that the aim of maximizing correctly estimated expected utility can be articulated in precisely the way that the aims of expected utility maximization are traditionally articulated, apart from any connection to grand world options.[20] Putting it yet another way, we

[19]Desirable in what way? This is part of what a decision theory should answer with its definition of correct utility estimation. That is, part of the point of specifying the correctness conditions for utility estimates for small world outcomes will be to disclose the relevant concept of desirability, in the sense of some kind of average or agglomeration of the utilities of the worlds in such outcomes.

[20]This point can be made in defense of the view presented in §5.2.1 as well.

might say that the rational aim in the small world is "formally" the same as in grand world. The primary difference lies in the way outcome desirability is understood. In the grand world, it is basic, understood via the idealized agent's utility function. In the small world, it is to be understood in terms of some kind of average of that ideal function over the various specific possibilities embraced by a given outcome.

With a sense, then, of what the aim of small world deliberation and choice comes to, we can turn to the second objection, namely the challenge to say what maximizing correctly expected utility accomplishes. The response should be obvious. The agent in the small world who maximizes correctly estimated expected utility, has, relative to their limited and provisional grasp of their options and the possible states of the world, chosen an act that gives them maximal subjective probability, in a sense suitable to the small world, of a desirable outcome. That is a legitimate practical accomplishment. A defender of one of these three accounts can thus say with some plausibility that achieving the aim of the small world is not a guarantee of perfect subjective rationality, but is not nothing either.

5.4 Final Thoughts

Of the views surveyed in the previous chapter, which is best? I am reluctant to respond to this with any pretense to finality. I much prefer to think there is room for inquiry and dialogue about how to best understand rational deliberation's relation to the grand world. But reflection on the issue does seem to tend hazily in the direction of an answer: pluralism. In this final section I'll present some final thoughts on this.

The argument made in §5.3.3 in support of maximizing correctly estimated expected utility — understood in accord with PSW1 and PSW2 — strikes me as pretty forceful. There seems something of genuine rational value to that sort of maximizing regardless of any inherent connection to rational grand world options. This abstract intuitive sense seems to me strengthened by consideration of cases like the following:

		Utility
Choose Die	Die 1	10
	Die 2	-100
	Die 3	-100
	Die 4	-100
	Die 5	-100
	Die 6	-100
Choose Coin	Coin H	9
	Coin T	8

Figure 5.1: Rejecting a Chance at Grand World Maximizing

Shelly has been presented with a six-sided die and a coin. Shelly knows he can roll the die or toss the coin, but can't do both. Shelly is certain that if the die lands on one he wins $10 but if it lands on any other side he must pay $100. Shelly is certain that if the coin lands heads he wins $9 and if it lands tails he wins $8. What he does not realize — though he would if he were to reflect briefly on the rules of the game as they were (overly hastily and intricately) explained to him — is that he can, if he so chooses, just place the die on whatever side he wants, and the same goes for the coin; in other words, Shelly is not bound to chance — he can dictate the payoff. Alas, Shelly is quite in the dark about this loophole, and believes he must choose to either roll or flip. It seems sufficiently intuitively clear that *in that mental state* flipping the coin is not clearly an irrational act for Shelly.[21] Let's continue the story. Suppose Shelly flips his coin and we immediately come in with his eight or nine dollars and ask him, "Why didn't you just pick up the die and put it down so that the one was facing up?" His reply, it is natural to imagine, might be, "I didn't realize that was an option!" To which we might naturally respond, "Do you recall when they were explaining

[21]This intuition does not seem to change if we add that as a third option Shelly could deliberate further, as by the lights of Shellyr current state, that might seem clearly (if misleadingly) otiose.

the rules and they said . . . ? Well if you just think about that can see that it does not actually rule out just picking up the die and placing it on whatever side you want." Shelly might reply, "Oh, I see! Gosh, if I'd just thought about it a bit I'd have realized that. I feel silly! But you asked why I flipped the coin, and now you know."

A few intuitive details stand out to me as worthy of note. First, we are intuitively reluctant to say that Shelly's action was *perfectly* rational. The reason seems to be that he had the information — and, we may assume, the time — to make what by his lights would have clearly been a better choice, one that in fact never occurred to him as being an option, though it could have, if he'd reflected. Second, if we consider the question of what would've been the most rational choice for Shelly — what he rationally should have done — the answer does not appear to be that he should have just chosen (rolled) the die, even though that small world option would have maximized likelihood of maximizing grand world expected utility — since of course choosing the die is at least compatible with choosing a maximal grand world option — placing the die with the one facing up — unlike choosing the coin. Rather, the answer seems to be that he should've picked up the die and put it down with the one facing up. But third, there is a clear intuitive pull to say that relative to his understanding — his small world partitioning of his options — his choice was not clearly completely subjectively irrational. So this case seems to tell against the first three paths (§5.2.1, §5.2.2, §5.3.2) and in favor of some kind of pluralism.

A slightly different case will throw more light on things. Suppose today is Monday and Sam is deciding whether to run, lift weights, or not exercise. Let us assume that in the grand world, Sam would not just be considering today, but each day of the upcoming week. Let us suppose that a grand world option is a seven-member sequence of such choices — e.g., (R, N, N, L, N, R, N) is the grand world option which consists of running only on Monday and Saturday and lifting weights only on Thursday.[22] Let's suppose that grand world options have expected utilities as follows. Any option with exercise on four total days (Mon, Wed, Fri, Sun) separated from one another by rest days (Tue, Thu, Sat) has expected utility of 12. Any option with exercise on three total days separated from one another by rest periods of at least one day has expected utility of 10. Options

[22]Sam has thus $3^7 = 2187$ grand world options.

like this but with two total days of exercise have expected utility of 7. Those with only one day of exercise have expected utility of 5. Lastly, the unique option with zero exercise days has expected utility of 0. The (grand world imaged) probabilities of these five propositions conditional on what Sam does today are as follows:

	Utility	**Run**	**Lift**	**Nothing**
4 Days	12	0.1	0.3	0
3 Days	10	0.12	0.1	0.7
2 Days	7	0.7	0.1	0.1
1 Days	5	0.1	0.5	0.1
0 Days	0	0	0	0.1
$\mathbf{CorEstEU^{EDT}}$	\	7.8	7.8	8.2

Figure 5.2: Sam's Exercise

Assume for simplicity that the correctly estimated expected utility of not exercising today is the highest out of his three small world options and the correctly estimated expected utilities of the exercise options are equal — this is what we get from EDT (cf. above and §5.2.2.2). Note that Sam maximizes the likelihood that he maximizes grand world expected utility iff he lifts weights today, but that option also makes it 50% likely that he will only exercise once this week. Further, notice that while running fails to maximize the likelihood of maximizing grand world expected utility, it exhibits lower variance with respect to correctly estimated expected utility compared to lifting: running has variance $3.5768 = .1(12-7.8)^2 + .12(10-7.8)^2 + .7(7-7.8)^2 + .1(5-7.8)^2$ compared to $9.76 = .3(12-7.8)^2 + .1(10-7.8)^2 + .1(7-7.8)^2 + .5(5-7.8)^2$ for lifting. For comparison, doing nothing has highest variance: 10.16.

What is it rational for Sam to do? Do you have a clear intuition? The only intuitions I have are that none of the options are clearly required and none are clearly irrational. If that is right, though, then doubt is cast on the first path (§5.2.1) and on the second path (§5.2.2) insofar as these entail

that doing nothing is rationally required, and doubt is cast on the third path (§5.3.2) insofar as it entails that lifting is required. The right approach again seems to be some form of pluralism.[23]

Here is one final question that I cannot address but that I wish to at least flag: what do we say about decisions that maximize all relevant factors — correctly estimated expected utility, likelihood of maximizing grand world expected utility, lowest variance, etc. — but that are made relative to crude binary option and state partitions with no real time spent deliberating? It seems decisions of this sort can can exhibit a sort of rational faultiness that the Joycean framework as we have understood it is not built to capture.[24]

[23]Pluralism of what sort? I am unsure.

[24]We might look to Skyrms (1990) for help with this, with the guiding thought being that such decisions might lack a kind of dynamic deliberative equilibrium of the sort Skyrms discusses. But making sense of such an approach will be no simple matter, particularly insofar as Skyrms's approach takes for granted that the set of options stays fixed throughout deliberation, whereas on the Joycean conception, divvying up one's options is itself in part the very subject of one's deliberation.

PART II: ON GIVING YOURSELF A SIGN

CHAPTER 6

INTRODUCTION AND BACKGROUND

6.1 Introduction

Sometimes people want a sign. They gaze heavenward hoping for a sign from a higher power. They scrutinize text messages for signs of affection. They peek into hot ovens hoping to see signs that their attempt at baking is going well. And so on. Suppose you want a sign and you believe you can *manufacture it* — what's more, if you do, it will *retain the significance you desire*. This isn't normally how it works. We don't usually go around trying to make the signs we desire, since we know in advance that our best effort would only produce a dud lacking the desired significance.[1] But in the weird case where you can create the sign you want and still have it mean what you want it to mean, do you necessarily have a practical reason to do so? In this part of the dissertation I argue that a straightforward means-end principle implies the answer is Yes — so long as we understand 'reason' subjectively (§7).

That result in itself is interesting, but what is also interesting is that such reasons are grounded in a species of extrinsic desire — desire for signs — that you can possess even when you're sure you lack the means to satisfy any of your intrinsic desires. In other words, desires for signs can independently ground practical reasons not aimed, so to speak, at satisfying intrinsic desire. After responding to objections (§8), I investigate the relationship between such reasons and reasons grounded in intrinsic desire; I argue the latter sort of reasons can *trump* the latter. I sketch a *Two-level Means-end Account* of subjective desire-based reasons to regiment everything (§9).

---

[1] Hacking your text message logs won't give you evidence that you'll be be hired at the job you applied for, alas.

6.2 Objective and Subjective Practical Reasons

Assertions to the effect that so-and-so has a reason to do such-and-such can be heard objectively or subjectively. Bernie thinks it's gin and tonic in his glass, but it's gasoline (cf. Williams (1981)). Bernie wants to drink gin and tonic, not gasoline — does he have any reason to take a sip? On one way of hearing the question (the objective way), the answer is No. On another way of hearing the question (the subjective way), the answer is Yes. It seems thus our ordinary talk about having reasons to do things is ambiguous. There are two different sorts of practical reason our talk can be about, depending on how it is disambiguated: objective practical reasons or subjective practical reasons. Bernie has no objective practical reason to take a sip, but he does have a subjective practical reason to take a sip.

Some philosophers deny both the ambiguity and the existence of subjective practical reasons (Dancy (2012); see also Thomson (1986)). Others say that while Bernie has an "apparent" practical reason to take a sip, he in fact has no practical reason to take a sip Parfit (2001); Parfit (2011); Sylvan (2015, 2020). Still others ((Mackie, 1977, 77); Joyce (2001); Schroeder (2007, 2008, 2018)), myself included, grant the ambiguity and the existence of subjective practical reasons. This paper is primarily addressed to those in the second and third camps, i.e., those who in some sense countenance subjective practical reasons.[2] Naturally, members of the first camp are invited to read on in the spirit of understanding the rest of us.

6.3 The Means-End Principle

I will be assuming the following:

Means-end Principle (MEP): An agent S has a subjective practical reason, grounded in a belief B and desire D, to do option O iff D is a desire of S's with content p and B is a belief of S's that doing O is (or might be) a means to the truth of p.[3]

[2] I will often leave 'practical' in 'subjective practical reason' tacit.

[3] I intend 'grounded in' as a placeholder for the distinctive relation between subjective reasons and belief-desire complexes. This relation could be like the 'in virtue of' relation familiar from the literature on metaphysical grounding, or it could be like constitution or identity.

MEP captures a means-end view of subjective reasons that is relatively modest inasmuch as it is compatible with such reasons having sources other than desire. What it denies is the possibility of a subjective reason grounded in a belief-desire complex where the belief isn't means-end.[4] On this view, if an agent wants something, and believes one of their options is a means to that thing, then those facts give rise to a subjective reason for them to do that option, and that is the only way a belief and desire can work together to ground a subjective reason.

MEP offers a plausible explanation of Bernie's having a subjective reason to take a sip. Bernie wants to drink gin and tonic and believes taking a sip is a means to that. Thus, he has a subjective reason to take a sip grounded in his means-end belief and desire.

I intend 'is a means' in a broadly causal sense. In particular, the fact that p is/will be true if O is done is not sufficient for doing O to be a means to the truth of p. It is sufficient (but not necessary) for doing O to be a means to the truth of p in the intended sense that doing O will/would cause p to be true. It is natural, for instance, to say that my drinking coffee is a means to staying alert. This ordinary, broadly causal sense of 'is a means' anchors my usage. Two notable ways for S doing O to be a means to the truth of p in the intended sense: (1) S doing O would increase the objective chance of p, (2) p is the proposition that S does O.

Readers may find MEP more plausible if, on the right-hand-side of the biconditional, instead of 'belief', we write 'justified belief' Joyce (2001); Gerken (2011), or 'justified true belief' Littlejohn (2009), or even 'piece of knowledge' Hawthorne and Stanley (2008). If so, they should read as if this substitution has been made throughout, and as if the same modification has been made in the right hand side of The Rational Movement Condition below. Notably, readers who opt for either of the latter two substitutions will have to say Bernie does not have a reason to take a sip grounded in his means-end belief and desire for gin and tonic; they will require a different sort of case to illustrate the subjective/objective reasons distinction.

[4]MEP can be fairly considered an internalist-adjacent principle cf. Williams (1981); Smith (1994). But care should be taken in drawing this connection. Internalism about practical reason is sometimes understood as the view, roughly, that an agent S has an objective practical reason to do ϕ iff S would be motivated to do ϕ if S were fully informed and deliberating perfectly. Since MEP concerns subjective rather than objective practical reasons, the falsity of Internalism as just stated does not imply (or at least, does not trivially imply) the falsity of MEP.

MEP can be fairly considered a Humean-adjacent principle as well cf. Joyce (2001), 52-53. But similar care should be taken here. Humeanism about practical reason is sometimes understood as the view that, roughly, an agent's objective practical reasons are as a rule grounded in their desires Schroeder (2007). The falsity of Humeanism in this sense does not imply (at least, not trivially) the falsity of MEP.

Here's the reason for the "at least not trivially" parentheticals. Some theorists characterize subjective reasons in terms of objective reasons. For instance, Lord (2010) defends the Factoring Account (cf. Schroeder (2008)), on which subjective reasons are just objective reasons that one in some sense "has." A less reductive view, favored by Schroeder (2007) and Parfit (2011), says roughly that one has a subjective reason to do ϕ just in case one has a belief that if true would give one an objective reason to do ϕ. Other such "objectivist" theories have been proposed Vogelstein (2012); Whiting (2014); Sylvan (2015); Wodak (2017, 2019). If such a theory is true — a question on which I take no stand — then Internalist and Humean theories of objective reasons may turn out to entail MEP.

6.4 The Constitutive Roles of Subjective Practical Reasons

Subjective reasons play constitutive roles in our folk theory of rationality (Schroeder (2007, 2008); Wodak (2017)). Three such roles will be relevant.

The core role has to do with rational action. It can be stated simply with the following principle:

Act Rationality: An action is rational iff (and because) the subjective practical reasons to do it are not outweighed by the subjective practical reasons to do otherwise.

A second role played by subjective reasons has to do with rational criticism (Schroeder (2007); Wodak (2017)). Our practices of rationally criticizing agents for their actions or deliberations are constitutively bound up with perceived failures to suitably respond to subjective reasons. The following two principles capture this:

Act Criticizability: An agent is rationally criticizable for doing an act A iff (and because) whatever subjective practical reason(s) they have to do A are outweighed by the subjective practical reason(s) they have to do one of their alternative options.

Deliberation Criticizability: An agent's practical deliberation is rationally criticizable iff (and because) they failed to suitably appreciate, individually or collectively, some of the subjective practical reasons they have for or against one of their options.[5]

A third role played by subjective reasons has to do with rational motivation (Schroeder (2007, 2008)). I'll rely on the following plausible necessary condition:

Rational Movement Condition: An agent is in a position to be rationally moved by a belief B and desire D to do one of their options O only if they have a subjective practical reason R, grounded in B and D, to do O.[6]

[5] There might be some subjective reasons that it is suitable to ignore when deliberating (cf. Wedgwood (2013)).

[6] The Rational Movement Condition coheres with a plausible account of motivating reasons provided by Schroeder (2007), which states that for R to be an agent's motivating reason for doing some act A is just for R to be both a subjective reason for her to do A and an explanatory reason for why she did A.

Being rationally moved by a belief B and desire D to do O is at least materially equivalent to doing O and having a *motivating reason* grounded in B and D. Motivating reasons are reasons that are *acted on* — reasons *for which* an agent does an action. They are not mere causes of action. They *rationalize* action (cf. Davidson (1963)). Subjective reasons are what rationalize actions; to act and have a motivating reason is to act on a subjective reason one has.

6.5 Desire

Subjective reasons are one main element in the story I want to tell; the other is desire. I will assume that desires are propositional attitudes in the sense that a token desire with content p is essentially such that it is *satisfied* iff p is true. So for example, where we might ordinarily speak of Scooby wanting a club sandwich, I will understand Scooby as desiring that Scooby has a club sandwich. Having made this clear, I'll sometimes state things in a way superficially inconsistent with it when doing so is convenient for presentation.

6.5.1 Intrinsic and Extrinsic Desires

To intrinsically desire something is to desire it in and of itself, for its own sake. To extrinsically desire something is to desire it for the sake of some further, logically distinct thing. Intrinsic desire is logically independent of extrinsic desire; you can have or not have an intrinsic desire for something you extrinsically desire.

6.5.2 Extrinsic Desires are Derivative

Extrinsic desires are *derivative*. They typically derive from other desires in virtue of "connecting beliefs," e.g., a belief that if an extrinsic desire is satisfied some further desire will be satisfied ((Smith, 1994, 157); Smith (2004); Arpaly and Schroeder (2014)).

The claim that extrinsic desires are derivative can be unpacked in two ways. First, extrinsic desires tend to vanish if and when the beliefs in virtue of which they derive are lost — e.g., if Joe the health nut stops believing that exercise causes health his desire to exercise vanishes. Second, extrinsic desires are — at least stereotypically — *rationally explainable* by appeal to further desires (i.e., those from which they derive) together with beliefs (i.e., those in virtue of which they derive).

Rational explanation in this context is not merely causal. We ask questions like "Why do you want to go to med school?" and expect answers like "I want to become a doctor and (I believe) going to med school is a means to that" which intuitively differ substantially from answers like "I was brainwashed."[7]

6.5.3 Signatory Desire

The key type of desire in this paper is a species of non-instrumental extrinsic desire that McDaniel and Bradley (2008) have dubbed *signatory* desire.[8] Consider Tom, who just texted his crush Judy and asked if she's into him. Tom very much wants to receive a text that reads "Yes." Tom's desire for a "Yes" is not intrinsic, but neither does he believe a "Yes" would be a means to anything else he wants. He only wants it because it would be a sign that Judy reciprocates his feelings. His desire for the text is signatory.

The following definition will suffice for our purposes:

Signatory Desire: A desire D for p is signatory iff there is some q such that D derives from a desire for q in virtue of a belief that p's truth is a sign (but not a cause) of q's truth.[9]

Being a sign but not a cause is a familiar notion. The failure of the bathtub to drain is a sign but not a cause of a clog in the pipe. A persistent cough is a sign but not a cause of an infection.

6.5.4 Desire's Causal Roles

Desire's roles in folk psychology include but aren't limited to motivating and rationalizing action in combination with belief (cf. Sinhababu (2009); for methodological background, see Braddon-Mitchell and Nola (2009)). Here are five more causal roles. First, the believed satisfaction or frustration of desire is causally tied to pleasure, displeasure, and related feelings and attitudes: coming to believe a desire is or will be satisfied causes pleasure (or excitement, relief, contentment,

[7] Exceptions to the letter (but not the spirit) of this characterization of extrinsic desires come from cases of desires for knowledge. I'll discuss these cases in §6.5.6.

[8] Intrinsic desire is often contrasted with instrumental desire, i.e., desire for something as a means to some end. But as the category of signatory desire illustrates, not all extrinsic desires are instrumental Harman (2000), 128-9; McDaniel and Bradley (2008); Arpaly and Schroeder (2014).

[9] McDaniel and Bradley (2008) leave out the parenthetical bit.

etc.), whereas coming to believe a desire is or will be frustrated causes displeasure (or disappointment, sadness, annoyance, etc.). Second, desires cause one to fantasize about their objects and dwell on what it is about their objects that is desired (cf. (Scanlon, 1998, 38)).[10] Third, desires direct our attention to things in our environment associated with their objects. Fourth, desires can be intensified by dwelling on vivid representations of their objects. Fifth, desires cause a distinctive phenomenological feel (a "pull") vis-à-vis their objects.

6.5.5 Idle Desire and Working Desire

Both of the following sentences are true:

1) I want a new car.

2) I want it to be the case that both (i) I have a new car and (ii) there are an even number of hairs on my head.

But only the content attributed by the first of these sentences is such that a mental state with that content plays the causal roles of desire in my psychology (viz., I will not be the least bit let down if I come to believe I'm getting a new car but have an odd number of head hairs; I fantasize occasionally about having a new car, but never about the conjunction of that and my having an even number of hairs; etc.). So, we mustn't assume a one-one correspondence between an agent's desires and true attributions of desire to said agent.

Why is 2) true? Is it because I have a desire whose content is entailed by the conjunctive content 2) attributes? No. Ralph is an American seven-year-old. It is natural to think the first but not the second of the following is true:

3) Ralph wants a new toy.

4) Ralph wants it to be the case that both (i) he has a new toy and (ii) vaquitas love knafeh.

[10] Note that it only makes sense to reflect on "what it is about" the object of a desire that is desired if the desire is extrinsic. An intrinsic desire's object is desired in itself!

It's natural to think 4) can be false despite 3) being true because Ralph might have never heard of vaquitas or knafeh — he might not possess the required concepts.[11]

Why then is 2) true? Here is why: the subject of 2) — i.e., me — happens to believe the truth of the content attributed in 2) is a means to (and/or a sign of) the truth of the content attributed in 1), which in turn is the content of one of my desires.

What is it that prevents us from saying that I have a desire with satisfaction condition attributed by 2)? Just some causal constraint on positing psychological states. If we don't mind violating such a constraint, we can *stipulate* a one-one correspondence between true desire attributions and desires. As it turns out, doing this will be useful for us. We will simply keep in mind the distinction between the stipulated states and the states that actually play the causal roles of desire.[12] Call the stipulated states *idle desires* and the states that play the causal roles of desire *working desires*. Only working desires can be intrinsic; idle desires are by nature extrinsic.

Idle desires cannot ground subjective practical reasons. In particular, MEP and other principles quantifying over desires are meant to quantify only over working desires.[13] Henceforth, all my reference to desire will be confined to working desire, unless stated otherwise.

6.5.6 Are Signatory Desires Just Desires for Knowledge?

Perhaps signatory desires are nothing more than desires for knowledge.[14] That is, perhaps

Signatory Desires are Desires for Knowledge (SDK): To have a signatory desire for the truth of p is just to have a desire to know whether (or that) p.

For instance, perhaps Tom doesn't want a "Yes" text to appear on his phone, strictly speaking. Rather, perhaps what he really wants is to *know* whether (or that) a "Yes" has appeared, or to

---

[11] Vaquitas are a rare and endangered species of porpoise. Knafeh is a type of cheesy Arabian cake.

[12] Or rather, play "enough" of them cf. (Braddon-Mitchell and Nola, 2009, §1)

[13] The distinction between idle and working extrinsic desires is what I appeal to in reply to a worry with MEP based on a proliferation of subjective practical reasons grounded in extrinsic desires with gerrymandered contents (e.g., my "desire" for a million dollars cash buried under a foot of cat litter). Such desires are idle.

[14] . . . or evidence or justified belief (etc.). I focus on knowledge but the points generalize as far as I can tell. Thanks to an anonymous referee for raising this issue.

know whether (or that) Judy likes him. SDK says this is what signatory desire just is. I think SDK is false for two reasons.

The first reason is just that we can imagine cases where someone wants a sign but clearly doesn't want the associated knowledge. Examples that come to mind are weird but intelligible. Patrick the alligator-lover believes an Amazing Predictor spared an alligator's life iff she predicted rain in Bangkok tomorrow and that Patrick will never know about it.[15] Patrick hopes for rain in Bangkok tomorrow, but he does not want to know about it.

A second reason to think SDK is false is that agents can give up hope of knowing whether a sign obtains without giving up hope that it does. Damon is a dying man who wants it to be the case that Atlantis existed. He therefore wants it to be the case that Atlantean artifacts exist somewhere at the bottom of the Atlantic. Damon has given up hope of ever knowing whether there are Atlantean artifacts, but he has not given up hope that there are such artifacts. If one can give up hope of Desire A being satisfied while not giving up hope of Desire B being satisfied, then it seems Desire A and Desire B must be distinct.

I'm persuaded by cases like Patrick's and Damon's, but it's not essential to my purposes that SDK be false. As far as the arguments I'm going to give go, what's necessary is just that signatory desires are not as a rule believed to be means to satisfying intrinsic desire. So, would SDK being true throw a wrench into my arguments? There are two ways it could. First, it could be that any time one believes an option is a means to desired knowledge, one also believes it is a means to some intrinsic desire. Second, it could be that desires for knowledge of signs are intrinsic any time the desired knowledge is believed not to be a means to further desire satisfaction. I'll consider these possibilities in order.

Suppose Terry's only intrinsic desire is to love and to be (and to have been) loved. Terry desires to know whether his deceased relative really loved him. He believes he can gain this knowledge if he reads a special letter they wrote to him. It seems coherent to suppose that Terry does not

[15] All the "Amazing Predictors" in this paper are, unless stated otherwise, known by the agents in the cases to be almost but not quite infallible predictors of future events.

believe reading the letter would be a means to loving or to being loved.[16] Cases like this show that believing x is a means to desired knowledge doesn't imply believing x is a means to intrinsic desire.

Is all desire for knowledge intrinsic? Clearly not.[17] But what about desires to know whether mere signs obtain, where one is believes that that knowledge is not a means to any further desire — are all *those* desires intrinsic? Consider a concrete example. We ask Yancy why he wants to know whether canaries have been dying in the coal mine. He answers, "Because that would be a sign of dangerous carbon monoxide levels, and I don't want the miners getting poisoned." Note that Yancy can desire the canary knowledge even if he knows he can do nothing to stop miners getting carbon monoxide poisoning. Now, assuming Yancy has no other reasons for caring about canaries, the intensity of his desire to know whether canaries are dying will vary with his concern for carbon monoxide levels in the mine and his confidence that canary deaths are a sign. Further, Yancy's desire for the knowledge will be rationally explainable via his concern over carbon monoxide levels and his belief that canary deaths are a sign about that. So, desires to know whether signs obtain can be extrinsic, even when such knowledge is believed not to be a means to further desire satisfaction.[18]

[16]Maybe Terry believes himself to not be the sort of person to behave any differently given such knowledge, or perhaps Terry is a very old recluse who does not foresee any future loving relationships.

[17]Suppose you want to know whether Stop and Shop is open midnight on Saturdays. It is natural to think this desire derives from further desire(s) by virtue of belief that the knowledge is a means to getting groceries.

[18]Two tangential observations seem in order here. First, it's interesting to note that the way desires to know whether signs obtain derive seems structurally different from the way desires for means derive. In the latter case one's belief is that the thing desired stands in a means-end relation to some further thing one desires; but in the former case one's belief is not that the knowledge one desires stands in the is-a-sign-of relation to some further thing one desires, but rather that the content of the knowledge desired stands in that relation. Second, it seems a desire for knowledge can be extrinsic without deriving in virtue of *any* belief. If one wants a particular fact to obtain one can, in virtue of that alone, desire to know whether it obtains; desire for such knowledge seems extrinsic insofar as it waxes and wanes with, and is explained by, one's concern for the particular state of affairs.

CHAPTER 7

SIGNATORY DESIRES GROUND SUBJECTIVE PRACTICAL REASONS

The preceding chapter familiarized us with signatory desires, subjective practical reasons, and the Means-end Principle. In this chapter I connect the dots. Signatory desires together with means-end beliefs can ground subjective practical reasons, or so I will argue.

7.1 The Case of Donny

Donny wants to bowl a strike. Naturally, he thinks bowling is a means to that end. But Donny's case is unusual. His reason for wanting to bowl a strike isn't that he thinks it would be pleasant. In fact, he fully expects bowling a strike would overwhelm him with restless doubts. He expects he would doubt whether the strike really happened or was some cruel hoax; he's sure his resulting anxiety would keep him awake all night. Donny is a nervous, fitful man; he trembles at great turns of fate — especially when they involve money. What is Donny's deal? It's this: Donny's reason for wanting to bowl a strike is that he believes an Amazing Predictor mailed him one million dollars iff she predicted he'd bowl a strike tonight. The money, he thinks, will arrive tomorrow if ever. So, on this night, Donny is cold to the usual appeal of bowling, and doesn't think it'll be a means to anything he desires besides the strike. He believes bowling is a means to a strike, and his desire for a strike derives from his desire for wealth in virtue of his belief that a strike would be a sign (not a cause) that he'll soon be a millionaire. In particular, Donny believes bowling is a means to a sign *that will retain its significance as a sign* of the million. In other words, Donny believes that even if he "causally intervenes" in the production of the sign, it will still be a sign of the million. His desire to bowl a strike is signatory.

7.2 My Two Main Arguments

In this section I give two arguments for the claim that Donny has a subjective reason to bowl grounded in his means-end belief and his desire to bowl a strike. Some premises have been given specific titles in parentheses.

Here is my first argument:

Argument 1

Premise 1. MEP is true.

Premise 2. Donny believes that bowling is a means to bowling a strike.

Premise 3. Donny desires to bowl a strike. ("Desire to Bowl")

Conclusion 1. Donny has a subjective reason to bowl grounded in his means-end belief and his desire to bowl a strike. From P1, P2, and P3

How can this argument be resisted? An outright rejection of MEP would be implausibly ad hoc. A more serious objection would seek to limit the application of MEP to desires that aren't signatory. We'll consider such objections in §8.

Here's my second argument:

Argument 2

Premise 4. Prior to making his decision Donny is in a position be rationally moved to bowl by his means-end belief together with his desire to bowl a strike. ("In Position to be Rationally Moved")

Premise 5. The Rational Movement Condition is true.

Conclusion 1. Donny has a subjective reason to bowl grounded in his means-end belief and his desire to bowl a strike. From P4 and P5

In support of In Position to be Rationally Moved, I would say that Donny is intuitively in a position to bowl and rationalize his choice by saying, "I want to bowl a strike tonight, and bowling is a way to make that happen!" Denying The Rational Movement Condition outright just to avoid the conclusion would, like a blanket rejection of MEP above, seem implausible and ad hoc. It seems therefore that the best way to resist this argument is to give some reason for denying In Position to be Rationally Moved. In §8 we'll look at attempts to motivate such a move.

Stepping back, what exactly does Conclusion 1 amount to? Well, if Conclusion 1 holds, Donny's signatory desire to bowl a strike participates in two distinct dependence relations. One is the relation between Donny's subjective practical reason to bowl and the pair consisting of his desire to bowl a strike and his means-end belief. The other is the relation between Donny's desire to bowl a strike and his desire for future wealth (a relation which holds in virtue of his belief that bowling a strike tonight is a sign — but not a cause — of an impending million). Conclusion 1 is about the first of these two relations. The second relation is one that holds between desires in virtue of connecting beliefs. It does not take a subjective practical reason as a relatum.

7.3 Supporting Arguments

In this section I defend Argument 1 with the following arguments, which support the claim that MEP applies to Donny's signatory desire to bowl a strike. Once again, some premises have been given specific titles in parentheses.

Argument 3

Premise 6. Unless Donny has some subjective practical reason to stay home and not bowl, Donny is rationally criticizable if he chooses not to bowl. ("Criticizable Choice")

Premise 7. Act Criticizability is true.

Conclusion 2. Donny has a subjective reason to bowl. From P6 and P7

Argument 4

Premise 8. Donny is rationally criticizable if when deliberating over whether to bowl, he disregards his belief that bowling is a means to bowling a strike. ("Criticizable Deliberation")

Premise 9. Deliberation Criticizability is true.

Conclusion 2. Donny has a subjective reason to bowl. From P8 and P9

Argument 5

Premise 10. The best explanation for Donny's having a reason to bowl is that MEP applies to his signatory desire to bowl a strike. ("Best Explanation For Having a Reason to Bowl")

Conclusion 3. MEP applies to Donny's signatory desire to bowl a strike. From P10 and Conclusion 2

How can these arguments be resisted? There are two main ways. The first way to respond is to motivate a denial of Criticizable Choice and Criticizable Deliberation, and then reject Best Explanation For Having a Reason to Bowl as falsely presupposing the existence of a subjective practical reason to bowl. The second way to respond is to accept Criticizable Choice and Criticizable Deliberation and motivate a denial of Best Explanation For Having a Reason to Bowl by supplying a competing explanation for Donny's having a subjective reason to bowl. We'll look at both of these routes in §8.

7.4 Summing Up

Those are my arguments about Donny. I take them to generalize. That is, I assume that Donny's case points the way to all sorts of other possible cases with the same basic structure. Hence, I take these arguments to show that signatory desires, together with associated means-end beliefs, can ground subjective practical reasons.

CHAPTER 8

OBJECTIONS AND REPLIES

8.1 Objection One: Refusing to Satisfy Signatory Desire is Uncriticizable

This objection asserts that signatory desires are such that refusing to try to satisfy them is rationally uncriticizable. It directly targets Criticizable Choice and Criticizable Deliberation. A proponent of this would reject Best Explanation for Having a Reason to Bowl as falsely presupposing that Donny has a reason to bowl; further, they would consider MEP implausible if it isn't restricted to non-signatory desires. This objection challenges Argument 1 by attacking my unrestricted use of MEP and the support I offer for that in Arguments 3, 4, and 5. Notably, it does not challenge Argument 2. So even if this objection were successful, its proponents would have more work to do to rebut Conclusion 1.

8.1.1 Reply

If you want p to be true, and believe that the only way you can cause p to be true is by doing O, and you have no subjective practical reason to do anything else, then if you choose not to do O, it's just obvious that you can be rationally criticized. "Don't you want this? Don't you believe that the only way you can make it happen is to do O? Instead, you've done something you have no reason to do! What gives?"[1] It is inadequate to reply: "I do want p to be true, but the only reason I want it to be true is that I want q to be true and I believe that the truth of p is a sign — but not a cause — of the truth of q." This is inadequate because the reason for wanting p to be true is irrelevant. We can reply, "Yes, that is a perfectly reasonable basis for wanting p to be true. What's

[1] Is the sense of irrationality to be explained by the fact that the person has done something they've no reason to do? No, since this is not inherently irrational; it is rationally permissible to do what one has no reason to do if all of one's alternatives are such that one has no reason to do them either.

your point? Why aren't you doing the thing you believe can cause p to be true, given that you have no reason to do anything else?" Perhaps they might protest, "It's the *type* of basis the desire has. It does not derive from another desire in virtue of a means-end belief, but in virtue of a belief about being a sign. *That's* the excuse." I reject this invidious distinction. What is disqualifying about the distinctive basis of a signatory desire for p? Does the agent have a reason for desiring p? Yes. Is the reason a perfectly acceptable one? Yes. Is it granted, therefore that the desire is not an irrational one? Yes.[2] I fail to see any excuse for not doing the thing they think will cause p's truth, given that they have no reason to do anything else.

8.2 Objection Two: Signatory Desires are Inherently Irrational

According to this objection, signatory desires are inherently irrational qua desires. Hence, it's irrational to act with the aim of satisfying them. In other words, according to this objection, mere signs are always, as a rule, irrational *to want*, and therefore it is always, as a rule, irrational to act with the aim of satisfying desires for mere signs. Like the previous objection, this objection challenges MEP and a proponent of it would likely want to restrict MEP to non-signatory desires. But unlike the last objection, this objection also challenges Argument 2 by targeting In Position to be Rationally Moved: it is always irrational, according to this objection, to be moved by a signatory desire (together with a means-end belief) since signatory desires are themselves always irrational. Hence, contra In Position to be Rationally Moved, Donny was never in a position to be rationally moved by his signatory desire.[3]

8.2.1 Reply

Recall Tom from §6.5.3. Is there really something inherently irrational about him wanting to receive a "Yes" text? He seems in a position to give a quite ordinary and reasonable account of his desire. Receiving a "Yes" text would be a sign that something he hopes to be the case is the case: Judy is attracted to him. What's irrational about this? Consider someone who doesn't want to get

[2] Well, maybe not so fast. See Objection Two.

[3] That is, he may have been in a position to be moved, but not *rationally* moved.

sick, and who therefore desires to not feel a tickle in their throat. Is it really inherently irrational for them to want there to be no tickle? Or consider wanting to see an "OPEN" sign in the window of a restaurant you hope is still open, or wanting to hear your doctor say your newborn is healthy. Examples of what seem plainly rational signatory desires can be multiplied ad nauseam. The claim that such desires are all inherently irrational must surely strike any unbiased judge as being at odds with common sense. At the very least, it is not the defender of the rationality of signatory desires who bears the burden of proof, since having and explaining signatory desires is so commonplace. Let's consider, then, how my opponent might defend their doctrine.

Prevailing accounts of rational desire come in four flavors. On *content-based* accounts, a desire is rational iff it has the right sort of content — e.g., it is a desire for the Good, or for something believed good, or for something there is objective reason to want (Anscombe (1963); Audi (2001); Parfit (2011)). On *deliberative* accounts, desires are rational iff they are produced, controlled, or sanctioned by an actual or idealized process of rational deliberation ((Rawls, 1971, 407-424); Brandt (1979); (Smith, 1994, 157-161)). On *information-based* accounts, a desire is rational iff it is not based on false beliefs and/or it would be maintained even if the desirer were suitably well-informed (Hume (1739); Brandt (1979); Savulescu (1994)). On *coherence* accounts, a desire is rational iff it suitably coheres with the desirer's beliefs, desires, and/or self-conception (Velleman (1989); Smith (1997, 2004); Verdejo (2020)). Let's look at each of these in turn.

The sort of account that seems most apt to deliver the result that signatory desires are all as a rule irrational is a content-based account that classes a desire as irrational if its satisfaction is not, as a matter of fact, objectively good for the agent. One might assert

The Means to My Intrinsic Desire is My Good (MIDIG): The objective good for an agent consists in (α) the satisfaction of their intrinsic desires and (β) any proposition whose truth would cause the satisfaction of (one or more of) their intrinsic desires.

Next one might claim that since satisfaction of signatory desires belongs to neither to (α) or (β), it is never objectively good for an agent. Hence, on this account of rational desire, such desires are as a rule irrational. There are two problems with this.

The first problem is that MIDIG begs the question. It is not obvious that it isn't objectively good for an agent to get a sign they want, and more to the point it's hard to see why mere causes of intrinsic desire satisfaction should be part of an agent's good but mere signs of intrinsic desire satisfaction should not be. We could toss out condition (β), but then by parallel reasoning we wind up saying all extrinsic means-end desires are inherently irrational.

The second problem is that the sort of content-based account just sketched seems promising as an account of when there is objective reason for an agent to desire something but doesn't seem promising as an account of when there is subjective reason for an agent to desire something. Recall Bernie, who thinks his glass is full of gin and tonic when in fact it's full of gasoline. Is Bernie's desire to take a sip from his glass rational? Here we seem to see the familiar ambiguity. Bernie's desire seems unsupported by objective reason but nevertheless subjectively rational. Since it is desires' subjective rationality that would be relevant to whether they can ground subjective reasons to act so as to satisfy them, a content-based account seems like the wrong sort of account to oppose the rationality of signatory desires.

If content-based accounts don't support the claim that signatory desires are as a rule irrational, might some combination of the other three types of account do so? On the contrary, these accounts seem quite friendly to the rationality of such desires. It would be odd to deny that an informed and ideally reflective agent might desire a sign of something else they want, and that that desire might cohere with the agent's beliefs and other desires.[4]

Finally, it bears mentioning that even if signatory desires were irrational, that on its own would not trivially imply that seeking to bring about their satisfaction would be irrational. That inference is mediated by the non-trivial assumption that it is as a rule *practically* irrational to seek to satisfy irrational *desires*. But it is one thing for a desire to be irrational, it is another for behavior aimed at satisfying that desire to be irrational. Personally, I'm for a sharp separation of these two domains

[4] *Interlocutor*: Wouldn't being *sufficiently* informed mean not needing or wanting the sign? *Me*: If we understand "sufficient information" in that strong a way, it'll tend to make instrumental desires irrational. Suppose I think Pete's attendance will cause the party to be fun, and hence I want Pete to attend. Would my desire for Pete's attendance survive my being sufficiently informed if that means knowing whether Pete will attend? *Interlocutor*: Sure. If you knew he was not going to be there you would wish he were, and if you knew he was going to be there you would be glad. *Me*: The same can be said about signs. Suppose I want my X-ray to be clear because I believe that would be a sign I'm cancer free. If I knew that the X-ray was clear, I'd be glad, and if it knew wasn't, I'd wish it were.

of rationality. An agent whose actions are sanctioned by the beliefs they have about how to best effect their desires' satisfaction seems to me one whose behavior makes rational sense, even if the desires they aim to satisfy are irrational.[5] Such a position seems in keeping with a broadly Humean perspective on rationality.[6]

8.3 Objection Three: Do You *Really* Want What You Extrinsically Want?

This objection claims that when pressed we admit that we don't *really* want the things that we only extrinsically want. Therefore, extrinsic desires don't exist (Marks (1986); Finlay (2007)). For instance, if we press Donny by asking, "But is bowling a strike *really* what you want?" it would be natural for him to respond, "Well, no. All I *really* want is wealth." So perhaps Donny's desire for a sign isn't a real desire. If that's right, I was confused when I stipulated Desire to Bowl (i.e., P3), and Argument 1 fails for the simple reason that Donny doesn't want to bowl a strike — indeed he could not desire to bowl a strike unless he intrinsically desired to bowl a strike. Argument 2 fails because In Position to be Rationally Moved is false — Donny can't be in a position to be moved by a desire that isn't even real.

8.3.1 Reply

It is undeniable that sometimes people say they want things only to admit under pressure they don't *really* want them. But the fact that someone admits (or is disposed to admit) under pressure that they don't *really* want something does not immediately imply that the desire doesn't exist, for as I will now explain, the meaning of questions like 'Is that *really* something you want?' and 'What do you *really* want?' is context-sensitive.

---

[5] One might protest that for a desire to be irrational just is for it to be inherently irrational to seek to satisfy. This reduces Objection Two to the nakedly question-begging claim that signatory desires are irrational to seek to satisfy.

[6] Hume writes, "Where a passion is neither founded on false suppositions, nor chuses means insufficient for the end, the understanding can neither justify nor condemn it." I suggest that even if the understanding condemns one's desire, it won't condemn taking means sufficient for satisfying it.

Asking what someone *really* wants, or whether something they profess to want is something they *really* want, can be interpreted in at least three ways depending on conversational context.[7] First, it can be interpreted as asking whether a salient professed desire is a working desire. Second, it can be interpreted as asking what desire the salient professed desire proximally derives from. Third, it can be interpreted as asking whether the salient professed desire is intrinsic and/or what intrinsic desire the salient professed desire distally derives from. Thus, in some contexts you *can* truthfully say you *really* want something that you merely extrinsically desire, and in contexts where you can't, the reason is that the context is such that 'what you *really* want' just denotes what you intrinsically desire. Let me elaborate.

Sometimes in the face of pressure we steadfastly affirm that we *really* want things that we only extrinsically want. If we ask Donny whether he *really* wants to bowl a strike tonight, he might reply, "Absolutely!" We might continue, "But isn't there some further thing you want that explains *why* you want to bowl a strike tonight?" He might reply, "Of course, but that doesn't mean I don't *really* want to!" There's nothing unnatural about this. A competent speaker wouldn't judge his assertions to be baffling or contradictory. It would seem that Donny is taking '*really* want' to denote working desire as opposed to idle desire (§6.5.5), which is a perfectly ordinary and acceptable interpretation. A real-world example might be helpful. It's the day after the 2020 US Presidential Election. Votes are still being counted in swing states like Michigan. I want Joe Biden to win the presidency. If he's not going to win the presidency, I don't care if he wins Michigan. His winning Michigan is desirable to me solely as a potential means to his winning the election — it is an extrinsic desire. I am being sincere when I say I really, genuinely want Biden to win Michigan. What do I mean? I take myself to mean that my desire is a working desire. I fantasize about Biden winning Michigan. My attention is drawn to news alerts about Biden and Michigan. If I learn that Biden wins Michigan I will be excited and pleased; if I learn that he lost Michigan I will be displeased and troubled. And so on.

[7] My discussion will be non-technical, but it fits well with the contextualist Karttunen-style semantics -Karttunen (1977) for questions and embedded wh-complements that Stanley and Williamson (2001) draw on in their account of know-how.

So we've just seen that there is one natural interpretation of '*really* want' on which one can truthfully assert that one *really* wants something one only extrinsically wants. For purposes of blocking the objection, this is all that's needed. But for the sake of completeness, let's consider contexts where a speaker accedes to not *really* wanting something they formerly professed to want. Are all such contexts ones in which you cannot truthfully assert that you *really* want what you only extrinsically want? I will argue that the answer is No. Such contexts come in three varieties; let's look at them one-by-one.

In the first kind of context, the point of asking whether someone *really* wants something is to get them to confirm whether the desire is a working desire. Suppose you ask, "Do you want to hear Wolf Blitzer announce that Biden won Michigan?" and I answer, "Yes." Then you ask, "But is hearing *Wolf* make the call *really* what you want?" I reply, "Well, no, I guess not. What I *really* want is for Biden to win Michigan. I don't really care who announces it." In this example, my professed desire to hear Wolf Blitzer announce a Biden win in Michigan is idle. I believe that its satisfaction would be a sign that Biden won in Michigan, but I have no working desire to hear Wolf Blitzer make the announcement — I do not fantasize about Wolf in particular making the announcement, etc.[8] Note however that in this case I've truthfully said what I *really* want is for Biden to win Michigan — but this, of course, is an extrinsic desire.

In the second kind of context, one asks whether someone *really* wants something they profess to want and what one is after is what desire their professed desire proximally derives from. Suppose you ask, "But is Biden winning Michigan *really* what you want? What are you *ultimately* hoping for?" In this updated context it seems I can't truthfully reply by saying that what I *really* want is for Biden to win Michigan. Instead, I can say, "Well no, what I *really* want is a Biden presidency." If this satisfies you then it seems what you were after is what desire my desire for a Biden win in Michigan proximally derives from. Note that I truthfully said that what I *really* want is a Biden presidency, which is still an extrinsic desire.

[8]Note the difference between an idle desire for a sign, as in this example, and a working desire for a sign, as in Donny and Tom's cases.

In the third kind of context, the point of asking whether someone *really* desires something is to get them to confirm whether their desire is intrinsic and/or to say what intrinsic desire stands at the end of the chain of desires from which the professed desire ultimately derives. Suppose you keep up your pressure. Soon enough I'd catch on and might say, "Well, I suppose what I *really* want is for there to be happiness, health, peace, and justice in human society." In the context created by the repeat questioning it seems I can't truthfully say I *really* want what I only extrinsically want. But note that this isn't the most natural interpretation of '*really* want'; it can take a bit of nudging to fix on it. It can be hard to pin down one's intrinsic desires. Pressing someone to do so is liable to seem weird and demanding. Further, note that in this sort of context 'what S *really* wants' doesn't denote the set of all S's desires, or even the set of all S's working desires — it just denotes a set of S's intrinsic desires.[9] Saying you don't *really* want something in such a context thus doesn't entail lacking a working desire for it.

8.4 Objection Four: Only Intrinsic Desires are Rationally Relevant

The rough idea of this objection is that intrinsic desires can do all the work we need from desires in a theory of practical reason, and therefore, extrinsic desires, including signatory desires, do not ground subjective reasons. This idea can take four forms, each targeting a different subset of my premises. The four forms come in two main variants each with two subvariants. One main variant stays neutral on whether extrinsic desires are real psychological states and asserts that when it comes to practical reason and rational motivation such desires are explanatory third wheels (cf. Marks (1986); Smith (2004)). This variant attacks MEP. On its first subvariant, one claims that MEP needs to be restricted to intrinsic desires — you'd go this route only if you were prepared to deny Criticizable Choice and Criticizable Deliberation. On its second subvariant, one says that a restricted version of MEP is true as a sufficient but not a necessary condition, and that a subjective reason can also be grounded in a pair consisting in an intrinsic desire and a belief that an action would itself be a sign that that intrinsic desire is satisfied — you'd go this route if you wished to

[9] Maybe it denotes the set of all S's intrinsic desires, or maybe the set containing the intrinsic desire(s) from which the desire S originally professed ultimately derives.

account for Criticizable Choice and Criticizable Deliberation without appealing to signatory desire. The other main variant of the objection primarily targets Desire to Bowl and asserts that extrinsic desires simply don't exist, i.e., Quines them[10] (Chan (2004); Finlay (2007)). This variant faces an analogous choice-point dictated by whether Donny is susceptible to rational criticism. On the one hand, the Quiner of extrinsic desires might reject Donny's susceptibility to rational criticism. If they go this route, there is no need for them to also restrict MEP to intrinsic desires — that sort of move is called for only if there might be an extrinsic desire that Donny takes bowling to be a means to, and the Quiner says such desires don't exist. On the other hand, if the Quiner agrees Donny is susceptible to criticism, then there has to be a subjective reason to bowl and MEP must be denied as a necessary condition; it is again natural to deny Best Explanation For Having a Reason to Bowl and offer in its place the sort of explanation mentioned above. Regardless of which form this objection takes, it will target Argument 2 via In Position to be Rationally Moved by claiming that only intrinsic desires can rationally move people.

Table 8.1: Summary of the Four Variants of Objection Four

	Neutral on Extrinsic Desires	Extrinsic Desires Don't Exist
Donny IS Susceptible to	Denies Premises 1, 4, & 10.	Denies Premises 1, 3, 4, & 10.
Rational Criticism	(Variant 1A)	(Variant 2A)
Donny is NOT Susceptible to	Denies Premises 1, 4, 6, & 8.	Denies Premises 3, 4, 6, & 8.
Rational Criticism	(Variant 1B)	(Variant 2B)

8.4.1 Reply

Let's start with those who accept my claims about Donny's susceptibility to rational criticism but reject MEP as a necessary condition and deny Best Explanation For Having a Reason to Bowl.[11] Someone going this route will need to explain Donny's susceptibility to rational criticism, and the most natural way to do so is to appeal to Donny's intrinsic desire for wealth together with the fact

[10] Here I'm using 'Quine' as a verb in the way popularized by Daniel Dennett.

[11] This covers Variants 1A and 2A in Tbl. 8.1.

that he believes the act of bowling would itself be a sign of impending wealth. Thus, it seems, someone going this route will accept

Evidential Sufficient Condition (ESC): If an agent S has a desire D for p and a belief B that one of their options O is a sign (but not a cause) of the truth of p, then the agent has a subjective practical reason, grounded in B and D, to do O.

Here I think we're encountering an intuitive fault line. All I can say is that to my mind it is primitive that insofar as rational action relates to desire, its aim is to do the most one thinks one can to *make* the world *into* what one wants it to be. If you believe an action would be a mere sign of that some desire of yours is satisfied, but would do nothing to *affect* whether that desire is satisfied, then that desire doesn't ground any subjective reason to do that action. James Joyce puts the idea nicely when he writes,

> Rational decision makers should choose actions on the basis of their *efficacy in bringing about desirable results* rather than their auspiciousness as harbingers of these results. Efficacy and auspiciousness often go together, of course, since most actions get to be good or bad news only by causally promoting good or bad things. In cases where causing and indicating come apart, however, ... it is the causal properties of the act, rather than its purely evidential features, that should serve as the guide to rational conduct. (Joyce, 1999, 150)

Is this a distinction without a difference? Is denying ESC while maintaining that MEP applies to signatory desires in some sense not meaningfully different from denying MEP as a necessary condition, accepting MEP as a sufficient condition, and maintaining ESC? I will argue No.

The claim that my rejection of ESC amounts to a superficial distinction has force only if the following principle is a necessary truth:

Signatory Belief to Desire (SBD): An agent with option O desires to do O if their beliefs entail that doing O would be a sign (but not a cause) of the satisfaction of one of their desires.

If SBD is a necessary truth, then necessarily, if an agent S is minimally rational, ESC implies S has a subjective reason to do ϕ only if MEP also implies S has a subjective reason to do ϕ.[12] But

[12] If ESC implies that an agent S has a subjective practical reason to do an option O, then S believes doing O would be a sign (but not a cause) of something they desire. Thus, given that SBD is necessary, if ESC implies that

SBD is not a necessary truth. There could be a hard-nosed sort of person, a "stoic causalist," who one day lacks any desire to do a thing they're certain isn't a means to anything they want, despite believing that to do it would be a sign of something they want.

Should we take SBD to be a norm, and hence conclude that the only time ESC and MEP clash is in cases where the agent is "already" irrational? I don't think so. For one thing, SBD only supports the connection between ESC and MEP if it is restricted to working desire. But given that restriction, SBD should strike us as a pretty unintuitive norm. (Would the stoic causalist be criticizable for, e.g., feeling no "pull" to doing the thing that would have no causal impact on what they care about?) For another, even if SBD were a norm, violating it would make one guilty of *conative*, not practical, irrationality. The violation would entirely consist in lacking a rationally required desire cf. Audi (2001), p. 69. It is not obvious that an agent couldn't be conatively irrational in such a way while still being practically rational and vice versa; these seem to be separate normative domains (cf. §8.2). While the broader issue deserves more examination, I am confident that the disagreement between ESC and MEP over what subjective *practical* reasons are had by the stoic causalist (who, ex hypothesi, would be conatively irrational) is substantive regardless of whether SBD is a norm.

So much for the rejection of Best Explanation For Having a Reason to Bowl by way of ESC and SBD. If that route is a dead end to any believer in MEP as a necessary condition, where does that leave this objection? It seems the only available alternative is to deny Criticizable Choice and Criticizable Deliberation and claim Donny is not susceptible to rational criticism in the ways I suggest in §7.3. There are two possible routes. One route is to remain neutral on the existence of extrinsic desires and claim that it is rationally uncriticizable to not do what you think would satisfy such desires, provided you don't think the rejected action might also lead to the satisfaction of any intrinsic desire.[13] I've already presented my case for the untenability of this route in §8.1

S has a subjective practical reason to do O, then S desires to do O. But if S desires to do O, then provided that S is minimally rational, S believes that doing O is a means to satisfy one of their desires, since the fact that doing O is a means to its being that case that one does O is analytic.

[13]This is Variant 1B in Tbl. 8.1.

and §8.2.[14] Setting that route aside leaves the final route that this objection might take. One might argue that extrinsic desires simply don't exist; and no one can be rationally criticized for failing to aim to satisfy desires that just aren't there.[15] This route denies Desire to Bowl, In Position to be Rationally Moved, Criticizable Choice, and Criticizable Deliberation; it dismisses Best Explanation For Having a Reason to Bowl as having a false presupposition (i.e., the presupposition that Donny indeed has a subjective practival reason to bowl). I assume it accepts MEP and does not restrict it (since it denies any need for a restriction — extrinsic desires don't exist). I've already considered and rejected an argument that could be used to support this position in §8.3. Let's consider a different sort of argument.

The sort of argument we turn to now uses inference to the best explanation (IBE). Examples in print focus on motivation. They claim that since rational motivation can be accounted for by intention and intrinsic desire (Chan (2004)) or by belief and intrinsic desire (Finlay (2007, 2008)) there is no reason to posit extrinsic desires. Taking this IBE route to denying the existence of extrinsic desires puts my opponent in a somewhat awkward position, as there appear to be in Donny's case two things intrinsic desire is ill-suited to explain: (1) Donny's meriting blame if he doesn't bowl (and praise if he bowls), and (2) Donny's being in a position to be motivated to bowl. Additionally, aside from cases like Donny's, I think extrinsic desires play the causal roles characteristic of desire (§6.5.4) in many cases better than intrinsic desires — this too presents a serious difficulty for an IBE-based rejection of extrinsic desires.

Let's start with Donny's susceptibility to criticism. The opponent we're now considering thinks there is no such susceptibility. But why not? Well, one kind of inference to the best explanation takes a liberal stance with regard to explananda. It is happy to deny certain would-be explananda if doing so results in a overall pattern of data with a parsimonious and unifying explanation. So the Quiner of extrinsic desire might argue that we should just reject Criticizable Choice and Criticizable

[14]The former section responds directly to the idea that Donny is rationally uncriticizable if he chooses not to bowl. The latter section bears indirectly on the same issue, insofar as it argues against one possible basis for claiming a refusal to bowl would be uncriticizable, namely that signatory desires are inherently irrational. As indicated in that section I personally think that such an argument would be moot, since I think we must strongly separate conative from practical irrationality, but I realize not everyone will agree on that score.

[15]This is Variant 2B in Tbl. 8.1.

Deliberation since in so doing we get rid of a few stray explananda that aren't readily captured via intrinsic desire.[16] The result of rejecting the intuition that Donny merits criticism if he refuses to bowl, goes the thought, is a simpler, more elegant, less gerrymandered combo of explananda and explanans. This take seems to me born out of too severe a lust for parsimony. The existence of extrinsic desires does not seem inelegant or gerrymandered. To be sure, a theory that posits only intrinsic desires is just plain *simpler*, but brute simplicity is a slim basis for rejecting the intuitions regarding Donny's susceptibility to criticism. The tradeoff appears less like good abduction and more like fetishism. But aside from this, there is a firmer reason to resist the Quiner: there are additional explananda, apart from Donny's susceptibility to criticism, that intrinsic desires alone are ill-suited to handle.

Consider motivation. Even if, to avoid begging the question against this objection, we do not assume Donny is in a position to be moved to go bowling by a signatory desire to bowl a strike, surely it must be granted that he is *somehow* in a position to be moved to go bowling. The battle should be over what the best explanation for this potential motivation is. My opponent will claim the best explanation involves Donny's intrinsic desire for wealth, rather than any signatory desire to bowl a strike. Given that Donny does not believe bowling is a means to satisfy any intrinsic desire, how can his intrinsic desires get involved in his motivation to bowl? The reply is that Donny can be motivated by the complex of his intrinsic desire for wealth and his belief that bowling, specifically bowling a strike, would be a sign of the satisfaction of that desire.[17] On this picture, Donny can be moved to bowl despite not thereby being moved to acquire, make manifest, realize, or in short, to cause, something he wants, to take place.[18] This does not square with the sort of explanation we would most naturally expect Donny to give for why he chose to bowl, should he so choose, or why he was tempted to bowl, should he not. Our natural expectation would be that he would explain his motivation in terms of a desire to bowl a strike: he is thinking about that strike and what it would

[16] Alternatively, the Quiner could endorse ESC and say that intrinsic desire captures these explananda via that principle. I've already given my response to that route. The Quiner we focus on now agrees that ESC is a no-go.

[17] An alternative, dialectically equivalent, says his motivation would come via an intention suitably formed on the basis of the belief(s) just mentioned.

[18] Notably, the opponent we're now imagining would hold such motivation to be *irrational*, since deeming it rational requires ESC, which we're assuming has been rejected.

mean, and he wants it to happen. It is less natural to suggest that on the contrary, he doesn't want the strike to happen, and indeed doesn't think bowling would yield anything he wants, but that he believes it would be a sign of something he wants, and can be moved by *that alone*.[19] Indeed I would go as far as to say that the very idea of such motivation, in which there is no prospect of getting anything one wants,[20] is incoherent.[21] Stepping back, though, it is unnecessary to press this incoherence claim; to rebut the IBE-based argument it suffices to point out that the explanatory picture the Quiner offers is less natural than that Donny's motivation would come from a common means-end belief and his desire to bowl a strike.

There is one final set of data that intrinsic desires are ill-suited to help explain, and for which it seems extrinsic desires ought to be posited. Namely, it seems clear that extrinsic desires, including signatory desires like Donny's, are in some cases the most natural states to play the causal roles of desire (cf. §6.5.4). That is, extrinsic desires, it seems, can be working desires (cf. §6.5.5). I've already discussed this in §8.3.

Not everyone agrees. Notably, Chan (2004) presents several cases meant to show that only intrinsic desires can play the causal roles stereotypical of desire. Here is one of his cases:

> John may be afraid of dogs and dislike the sight of dogs. But John may be in love with a woman whom he regularly sees walking her dog around the block. John may therefore welcome, and even look forward to, seeing her dog coming around the corner on its leash, as he knows that it will be followed by his beloved coming into view. Since John does not have an intrinsic desire to see the dog, it is suggested that his positive feelings in favor of seeing the dog can be attributed to an extrinsic desire.

Speaking of this case, Chan says the following:

> It is not the sight of the dog but the thought of seeing his beloved that gives John pleasure, and he would be not in the least disappointed if his beloved appeared around the corner without her dog. The daydreaming that is motivated by John's desires will be about meeting his beloved without the dog appearing, even if this cannot happen in the real world.

---

[19] *Interlocutor*: The claim is not that he can be moved by belief alone — but by belief together with his intrinsic desire. *Me*: How can an intrinsic desire help move him to do something he's sure won't lead to its satisfaction?

[20] And further, no belief that what one is doing is morally required or good. I am neutral here on whether moral judgments can motivate apart from desire.

[21] *Interlocutor:* Donny believes he will get evidence about something he wants. Isn't it coherent that that could be enough to move him? *Reply:* My claim is precisely that he wants that evidence, and that that desire can move him; by contrast, my opponent says he does not want the evidence and yet is somehow moved by it.

Chan's claims strike me as unappealing. I see no reason why imagining or experiencing the sight of the dog rounding the corner cannot be a source of pleasure for John. ("Oh boy, there's her dog! In just a moment, she'll be rounding the corner as well!") To be sure, John's belief that the dog is a sign of the woman's impending presence is (ceteris paribus) a necessary background condition for the causation of pleasure by dog sighting (imagined or real). But nevertheless, the sight itself still intuitively can cause pleasure. By the same token, it is implausible to insist that John cannot daydream about the dog's appearing (only to be followed closely by his crush) absent an intrinsic desire to see the dog. Here my opponent might argue that all that's needed to explain John's pleasure at the dog sighting (imagined or real) is an intrinsic desire to see his crush together with a belief that the dog is a sign that he'll soon see his crush. They might say the same thing about the other causal effects of the supposed extrinsic desire, e.g., John's attention being directed toward the sound of a dog approaching, and the "pull" he feels toward the thought of seeing the dog. This seems to me best taken as an argument for the view, defended by Smith (2004), that extrinsic desires can be reduced to "suitably related" complexes of intrinsic desire and belief.[22] It seems to give us no reason to eliminate John's extrinsic desire altogether. Whether Smith's reduction succeeds or not, extrinsic desire is in no danger of elimination.

The reader might be unsure of my take on Chan's case. Reasonably so. The sort of intuitions I'm aiming to prompt are liable to be muddied due to the "closeness," causally and spatiotemporally, of the objects of John's extrinsic and intrinsic desire (i.e., the dog and John's crush). I suggest we consider a case that lacks this closeness.

Recall my Biden example from §8.3. Consider me in the days after the 2020 election, before Biden's victory was announced. For me in that state, to find out that Biden has won Michigan

[22]Notably, Smith's "suitably related" qualification, which looks like a problem for a pure reduction, is indispensable. This is because it seems possible (cf. the "stoic causalist" described earlier) to have an intrinsic desire and believe that the truth of some proposition would be a sign or means to the satisfaction of that desire, and yet just not want the truth of the proposition (cf. Smith pp. 97–98). But further, leaving out 'suitably related' seems to commit one to something stronger than (the already implausible) SBD. Whereas SBD is limited to desires to do things, the view in question minus the "suitably related" qualification would entail that in general, whenever one has an intrinsic desire and a belief that the truth of a proposition would be a sign or cause of that desire's satisfaction, one thereby has an extrinsic desire for the truth of the proposition. The resulting proliferation of (mostly gerrymandered) extrinsic desires is unattractive, as they would all supposedly count as working rather than idle desires (cf. §6.5.5). One avoids this by saying that only the "suitably related" belief-plus-intrinsic-desire complexes count as (working) extrinsic desires.

would be a source of great pleasure; I fantasize about a Biden win in Michigan; my attention is drawn to news alerts about Michigan; etc. Some desire ought to be playing these roles. My claim is that it is an extrinsic desire for a Biden win in Michigan. But notice: all the ready-to-hand alternative desires that might play those roles are *also* extrinsic. What are those desires? Well, a desire for a Biden presidency, for a reversal of Trump's policies on climate change, for more humane policies on immigration, etc. Assuming these desires exist, none of them are for things I want *in themselves*, and in fact I'm not even sure exactly what intrinsic desires of mine would ultimately support them. If forced to specify the intrinsic desires, they'd be something like desires for general human happiness, equality, health, flourishing — but those too might ultimately turn out to be extrinsic desires. My immediate answer would be nebulous and tentative and I would need to reflect on it to sort it out.[23] In this case, it is, I expect, clearer than in Chan's case both that an extrinsic desire — my desire for a Biden win in Michigan — is playing the desire roles, and that there are no clear intrinsic desires ready-to-hand to usurp those roles.

[23]Facts about what people intrinsically desire seem like they'll ordinarily be subject to a good deal more indeterminacy than facts about what they extrinsically desire. Cf. Rawls (1971) "whether an aim is final or derivative is not always easy to ascertain. The distinction is made on the basis of a person's rational plan of life and the structure of this plan is not generally obvious, even to him" (494).

CHAPTER 9

THE TWO-LEVEL ACCOUNT

I have argued that an extrinsic (i.e., signatory) desire together with a means-end belief can ground a subjective practical reason. The cases I used demonstrate further that such reasons can occur even if the agent is certain none of their options are means to any intrinsic desire. So, there can be a subjective reason to do ϕ grounded solely in extrinsic desire, unaccompanied by any concurrent reason to do ϕ grounded in intrinsic desire. What is the status of such reasons? In particular, what happens when they compete with reasons grounded in intrinsic desire? In this section I defend the view that subjective reasons grounded in signatory desires can be *trumped* by reasons grounded in intrinsic desire, and that this can happen even when the desires and beliefs grounding the respective reasons are of equal respective intensities.

Suppose you only have two options, A and B. You believe A is a means to satisfy an extrinsic desire but not any intrinsic desire, and B is a means to satisfy an intrinsic desire. Suppose the desires are of equal intensity and the beliefs carry equal confidence. You rationally must do B. Why? Not because of any *quantitative* difference between your desires or reasons, rather because of the *qualitative* difference between intrinsic and extrinsic desires. If you intrinsically desire something, you want it for its own sake — it is one of your basic ends. If you only extrinsically desire something, you don't want it for its own sake — it isn't one of your basic ends. It is irrational to pursue something that isn't one of your basic ends at the expense of something that is — or at least this is so when the desires for the two things are of close to equal intensity and your levels of confidence in your ability to obtain each are close to equal.

Let's check this with a case. Walter believes a Predictor whose predictions he's sure are roughly 90% reliable sent him a coupon for a club sandwich, a beer, and a dessert iff she predicted he wouldn't bowl today. He believes the coupon will arrive tonight, at his house, if ever. What Walter intrinsically desires are pleasurable taste sensations (measured in gustatory hedons). He doesn't

believe staying home is a means to such sensations, since the coupon is either already in the mail or it isn't coming regardless. His other notable option is to go bowl with Donny and The Dude. He's confident that if (and only if) he hangs out with them, they'll buy him a club sandwich and a beer. A club sandwich and a beer would yield 9 gustatory hedons; an added dessert would make it 10. So, Walter does not believe staying home will get him any gustatory hedons, but it will get him a *sign* (with roughly 90% reliability) that he is going to get 10 hedons. By contrast, he's more or less certain that hanging out with his friends will procure him 9 gustatory hedons. His desire for the sign of 10 hedons and his desire for 9 hedons are of close to equal intensity, but they are of different quality: one is extrinsic, the other intrinsic. His confidence in the claim that the Predictor is 90% reliable is roughly equal to his confidence in his friends' willingness to buy him dinner. There are no other desires Walter believes might be satisfied as a result of anything he can do tonight. He can't both stay home and go bowl; he must choose one.

By MEP, Walter has a subjective practical reason to stay home, grounded in his signatory desire together with the associated means-end belief, and a subjective practical reason to go hang out with his friends, grounded in his intrinsic desire for nine gustatory hedons together with the associated means-end belief. What I claim is that the reason grounded in the signatory desire is trumped by the reason grounded in the intrinsic desire, even if the two desires have equal intensity. In other words, Walter isn't faced with a toss-up; the only rational option is to see his friends. To reiterate the core intuition, there is an inherent irrationality in choosing to make desirable signs of your ends happen rather than make your ends happen — or this is so at least in a case like Walter's, where the desire for the sign and the intrinsic desire are close to equal in intensity and the respective means-end beliefs carry close to equal confidence. It bears emphasizing that trumping in cases like Walter's isn't due to one reason's arising from a means-end belief and the other's arising from a "signatory belief" — both arise from means-end beliefs.

Are there any cases where trumping does *not* occur? If the extrinsic desire is far more intense than the intrinsic desire or the agent is far more confident in their ability to satisfy the extrinsic desire should we say that the reason grounded in the intrinsic desire still trumps the reason grounded in the extrinsic desire? In short, are there *thresholds* that prevent trumping? Intuitions may diverge. In the interest of setting forth the securest aspects of an account of subjective reasons grounded

in signatory desire, I will take a neutral approach. I will incorporate thresholds into my account, but make no assumptions as to where these thresholds should be set, or even whether they should be finite rather than infinite. So long as trumping occurs in cases like Walter's, where thresholds clearly do not come into play — given the near equality in both (a) the intensities of the signatory and the intrinsic desire and (b) the levels of confidence in the respective means-end beliefs — the outlines of a view involving trumping can be drawn.

Let's say that a subjective practical reason grounded in an extrinsic desire is *subordinate*, and a subjective practical reason grounded in an intrinsic desire is *superior*. Let the *weight* of a subjective practical reason grounded in a means-end belief and a desire be a non-decreasing function of the confidence the belief carries and the desire's intensity.[1] For any given option O, I assume there is some (possibly empty) set[2] containing all and only superior (/subordinate) reasons that favor O; I call this set *the* set of superior (/subordinate) reasons favoring O, and I assume the reasons in that set have a total weight which is a function of their individual weights.[3] I say that an agent's superior reasons favor some option O iff there is no O* such that the total weight of superior reasons favoring O* is greater than the total weight of superior reasons favoring O. Let every positive number be a *threshold*. Say that the set of subordinate practical reasons to do an option O is *trumped* if its total weight is below the appropriate threshold, and the superior subjective practical reasons favor

[1] I intend 'non-decreasing' in this sense: letting $w(c, i)$ be weight as a function of belief confidence and desire intensity, I assume that if n and m are non-negative then $w(c, i) \leq w(c + n, i + m)$.

[2] For convenience, I treat a one-membered set as interchangeable with its member and I speak collectively about the members of a set by speaking of the set.

[3] It is natural to think of this function as summation, but I don't assume this. I also don't assume favoring is one-one; a reason might favor two or more options equally — that is, it might favor doing any member of a set of options — or it might favor two or more options to different degrees.

some alternative option(s) distinct from O.[4][5] Say the set of subordinate reasons to do an option is *active* iff it isn't trumped.

With these definitions in hand, I propose

The Two-Level Account (TLA): If the set of subordinate subjective practical reasons $\mathbb{R}$ favoring an option O is active, its member(s) play the distinctive roles of subjective practical reasons. If $\mathbb{R}$ is trumped,

1. The member(s) of $\mathbb{R}$ do not contribute their weight to O; the weight, individually and collectively, of the member(s) of $\mathbb{R}$, is irrelevant to what it is rational for the agent to do.

2. The agent may not be rationally criticized for failing to act in accord with $\mathbb{R}$.

3. If the agent has suitably involved all their superior subjective practical reasons in their deliberation, failure to also consider, individually or collectively, the member(s) of $\mathbb{R}$, is not rationally criticizable.[6]

4. The agent cannot be rationally motivated by $\mathbb{R}$.[7]

TLA and MEP are logically independent. One could accept MEP only as a sufficient condition while also accepting ESC (cf. §8.4) and consistently conjoin both of those with TLA. Since I accept MEP as a necessary and sufficient condition and reject ESC, I prefer to take MEP and TLA as a package view, on which you can only have a superior subjective reason to do some option O if you

[4]Note that this is only a sufficient condition. There may be other circumstances in which a subordinate reason is trumped; I am currently uncertain about this.

[5]This is meant to be neutral on what thresholds are appropriate. It can be squared with the view that trumping always occurs by using the extended reals and taking the appropriate threshold to always be positive infinity. If appropriate thresholds should ever be finite, I make no assumptions about what determines them (beyond what I've already said), or about whether they should be the same for every agent at every time, or about whether they should be vague. What complications arise if they are treated as vague depends on what theory of vagueness is chosen (e.g., a trivalent approach requires dealing with reasons that are neither trumped nor not trumped). As far as I see, such complications are orthogonal to my claims.

[6]Note that this is only a sufficient condition. I tend to think there are other kinds of conditions under which it is rationally uncriticizable to ignore trumped reasons.

[7]This is not to say that the agent cannot be *irrationally* motivated by $\mathbb{R}$. Indeed I would say that they can.

believe that O is a means to the satisfaction of an intrinsic desire. I call the conjunction of TLA with MEP the *Two-level Means-end Account*.

Given the Two-level Means-end Account, Walter subjectively rationally ought to hang out with his friends. By MEP, Walter has a superior reason to hang out, and a subordinate reason to stay home. The latter reason is trumped (viz., its weight is below the appropriate threshold and Walter's superior reasons favor a different option). So, by TLA, Walter's subjective reason to stay home does not contribute its weight. Since his only other subjective reason favors hanging out with his friends, that's the rational choice.

There can be interesting cases where the subordinate reasons for one option are trumped but those for another option aren't. One way this can occur is if the weight of the former but not the latter is below the appropriate threshold. But it can also occur if superior reasons don't favor a unique option. Suppose there are subordinate reasons R1 to do A and R2 to do B, where R1 and R2 have weights below the threshold (so they are both "trumpable"), and superior reasons favor doing either B or C. In this case, R1 is trumped, but R2 is not. Let's look at a case that illustrates this.

The Dude intrinsically desires pleasure. He's certain that a Predictor mailed him a free day-pass to a spa iff she predicted he would bowl at Alley A, and eight dollars iff she predicted he would bowl at Alley B. The Dude wants to bowl at Alley A (/B), since he believes that would be a sign of an incoming spa pass (/eight bucks). The Dude thinks a spa day would be ten hedons; as for the eight bucks, he can use it to buy a White Russian, good for one hedon. Now, unlike Alley A, to which he's indifferent, The Dude likes Alley B; he digs the ambiance. He's certain going to Alley B is a means to nine hedons. If he stays home, Dude will take a long bath, also nine hedons. Like Walter, The Dude takes the Predictor to be 90% reliable. See Tbl. 9.1 below for a summary of The Dude's hedons in each action-prediction combination.

For simplicity and definiteness, let's assume that the weight of a reason grounded in a desire D and a belief B is equal to the strength of the desire times the confidence (understood as a percentage)

carried by the belief. Let's also assume that The Dude's desire for n hedons has an intensity of n.[8] The Dude has no superior reason to bowl at Alley A. He has a subordinate reason to bowl at Alley A with weight 9 = (desire for 10 hedon spa pass) × (90% confidence that bowling at Alley A is a sign of the spa pass). He has a subordinate reason to bowl at Alley B with weight 0.9 = (desire for 1 hedon white russian) × (90% confidence that bowling at Alley B is a sign of eight bucks, enough for one white russian). He has a superior reason to bowl at Alley B with weight 9 = (desire for 9 hedon experience at Alley B) × (100% confidence that bowling at Alley B is a means to 9 hedons). Lastly, he has a superior reason to stay home with weight 9 = (desire for 9 hedon bath soak) × (100% confidence that staying home is a means to 9 hedons).

Should Dude bowl at Alley A and get that sign of the spa pass? Or is he permitted to either bowl at Alley B or stay home, since either one is a means to more intrinsically desired pleasure than bowling at Alley A is a means to? None of the above! Given its equality in weight to the superior reason to bowl at Alley B, the Dude's subordinate reason to bowl at Alley A is trumped. But the subordinate reason to bowl at Alley B is active and breaks the tie with staying home. The Dude should bowl at Alley B: it's a means to just as much intrinsic desire as soaking in the tub, but it also satisfies a signatory desire.[9]

Table 9.1: The Dude's Hedons Depending on Which Prediction Was Made and What He Does

	A Predicted	B Predicted	Stay Home Predicted
Bowl at A	10	1	0
Bowl at B	18	10	9
Stay Home	18	10	9

[8]These assumptions could be relaxed to make the case more realistic but they're fine for purposes of this illustration.

[9]If we assume agents' utility functions over possible worlds ignore extrinsic desires and reflect intrinsic desire only (cf. Weirich (2001, 2015a,b)), we can observe an interesting decision theoretic result. Evidential Decision Theory says The Dude must bowl at either Alley A or Alley B, since according to EDT, ExpectedUtility(AlleyA) = $\sum_n n[\Pr(n \text{ hedons}|\text{AlleyA})] \approx 10$ = ExpectedUtility(AlleyB) = $\sum_n n[\Pr(n \text{ hedons}|\text{AlleyB})] \approx 10 >$ ExpectedUtility(Home) = $\sum_n n[\Pr(n \text{ hedons}|\text{Home})] \approx 9$. By contrast, Causal Decision Theory says he can either bowl at Alley B or stay home, since according to CDT, ExpectedUtility(AlleyA) + 9 ≤ ExpectedUtility(AlleyB) = ExpectedUtility(Home). Hence, by the lights of the Two-level Means-end Account, both Evidential and Causal Decision Theory are wrong, for The Dude must bowl at Alley B! For background on decision theory and the debate between the Evidential and Causal theories, see Jeffrey (1965), Nozick (1969), Gibbard and Harper (1978), Sobel (1978, 1986), Skyrms (1980, 1984), Lewis (1981b), and Joyce (1999).

Table 9.2: Summary of Weights of Reasons {#tbl:dude-table2}

	Bowl at A	Bowl at B	Stay Home
Total Weight of Superior Reasons		9	9
Total Weight of Subordinate Reasons	9	0.9	

9.1 Conclusion

As a general rule, when you want a mere sign, you do not also believe you can make that sign happen without undermining its significance. Cases like Donny's, Walter's, and the Dude's show that deviations from this rule are not inconceivable. I have argued that in such deviant cases, the agent has a subjective reason to bring about the sign they desire, and this reason can be accounted for by the straightforward means-end principle. What is more, such unusual cases make it clear that it is possible to have a subjective practical reason to do something despite believing it to be neither morally required nor a means to satisfying any of one's intrinsic desires. The view that subjective reasons can be grounded in signatory desires implies that they can be grounded in extrinsic desires; I have therefore responded to philosophers who argue that extrinsic desires are not rationally relevant as well as to those who argue that they simply do not exist. Last, I have argued that subjective reasons grounded in extrinsic desires can be trumped by ones grounded in intrinsic desires and have presented a Two-Level Account to express this idea. Conjoining this account with the aforementioned means-end principle yields an outline of a theory of desire-based subjective reasons: the Two-Level Means-End Account.

PART III: SUBJECTIVE DECISION THEORY

CHAPTER 10

STAGE SETTING

10.1 Introduction

In this part of the dissertation I present and defend a new approach to decision theory, based on the idea that expected utility is to be analyzed in terms of conditional thought, and that in the context of decision-making this sort of thought is not rightly understood as directed at public or contextually determined conditional propositions, but rather is governed by an agent's own idiosyncratic propensity to provisionally shift credence to supposed possibilities in their own subjective way subject to only a few weak constraints. The formalism I adopt to make this concrete is the possible world selection semantics of Robert Stalnaker (1975, 1984a), modified such that each possible agent at a time is represented as having their own subjective selection function. The general approach just outlined is what I most want to make plausible, but in much of the discussion to follow I will defend a specific theory — one making use of the probabilistic imaging of Peter Gärdenfors (1982, 1988). This Subjective Decision Theory (SDT) explains competing intuitions in Newcomb's Problem without entailing that either option is necessarily rationally required in that case, and offers reasonable verdicts in the more recent decision-making-under-determinism cases of Arif Ahmed (2013, 2014). One core contention I will press is that the line between the theory of practical rationality and epistemology needs to be drawn in such a way that much of what has been thought to belong to the former really belongs to the latter. In a nutshell, what my approach claims is that what it is rational for you to *do* is determined by what you suppose will be the case, for each of your options, if you do them, and what it is rational for you to *suppose will be the case* is the purview of epistemology, not decision theory.

In the remainder of this chapter I set things up with some motivating observations in the semantics of belief attributions and with some discussion of methodology in decision theory. In the next chapter, I present SDT and challenge some received wisdom about the nature of the

difference between EDT and CDT. In the final chapter, I survey cases and compare SDT to the recent Counterfactual Decision Theory of Brian Hedden (2023). I argue Hedden's intuitions are really reasons to embrace SDT.

10.2 Jackson and Harvey and Bert

Consider the following case, inspired by Frank Jackson (1977):

Edgy Jackson: Jackson is on the roof of a tall building. He walks to the edge and looking down sees an unobstructed drop one hundred plus feet straight to concrete below. The sight frightens him. His heartrate rises; his palms sweat; he gets goosebumps. He has an intrusive thought of flinging himself over the edge and falling screaming to his death. He silently reassures himself: he would never do such a thing. Jackson is disposed at that moment to answer whether he would die if he jumped by sincerely uttering, "Yes, certainly. If I jumped, I would die." Jackson stands there looking for a moment, then turns and leaves. Jackson is a reasonable guy who believes things that are plainly true on reflection and does not believe things that are plainly false on reflection.

Suppose we continue the story:

Harvey and Bert: These two guys hear someone say Jackson is at the edge of the building. Harvey says to Bert, "You know how sensible and cautious old Jackson is. He would only jump if he had previously set up a safety net to break his fall. If Jackson jumps, he won't die — a safety net will catch him!" Bert reflects a moment and says, "Yes, I agree. If Jackson were to jump, he would live. But you know what else? Right now old Jackson himself believes, heck, he *knows*, that if he were to jump he would live; he most certainly does *not* believe that if he jumped he'd die." "What do you reckon he'd say right now if you asked him whether he'd die if he jumped?" Harvey asks. To which Bert replies, "Why he'd say that he'd live if he jumped of course." Harvey smiles and nods.

What's your take on Harvey and Bert's conversation? In particular, how did Bert's assertions about Jackson sound to you? To me those assertions sound quite false, given what we know about Jackson's state of mind as he stands near the building's edge.

This clashes with a familiar philosophical understanding of conditionals and context. According to this familiar understanding, assertions like the ones Harvey opens the dialogue with, and that Bert accepts, modify conversational context in such a way as to presuppose the following counterfactual:

Backtracker: If Jackson were to jump, he would live.

Here is a typical vanilla formalism for this.[1] For any conversation, there is an abstract "scoreboard" that contains important parameters needed to determine what assertions in that conversation mean (cf. Lewis (1979b)). The scoreboard can be changed by the conversational contributions of the conversation's participants. One piece of the scoreboard for a context $\mathbf{c}$ is a *selection function* $f^{\mathbf{c}}$ such that for any w and P, $f^{\mathbf{c}}_w(P)$ is a set of worlds — intuitively the P-worlds that are most similar to w in whatever sense of "similar" is salient in $\mathbf{c}$.[2] We then suppose that a conditional *were* P *would* Q is true in $\mathbf{c}$ at w iff $f^{\mathbf{c}}_w(P) \subseteq Q$.[3] It is natural to say in addition that *were* P *would* Q expresses the proposition ϕ in $\mathbf{c}$ such that $\phi = \{w' : f^{\mathbf{c}}_{w'}(P) \subseteq Q\}$.[4] Thus, letting $\mathbf{h}$ be Harvey and Bert's context after their remarks and picking some world w, we can suppose that $f^{\mathbf{h}}_w$ is a function that ranks worlds as more similar to w, ceteris paribus, to the extent that in them Jackson is disposed to never perform impulsive acts of extreme self-harm. Thus $f^{\mathbf{h}}_{@}$(Jackson jumps) is a set of worlds where Jackson jumps and lives (because he lands in a net, or on a big cushion, or ...), which (1) makes Backtracker true in $\mathbf{h}$, and (2) makes Backtracker express a proposition in $\mathbf{h}$ that is true in the actual world.

So far, so good. But what we're interested in is not really Backtracker, but rather this:

---

[1] I would call what I am about to describe a semantics, but perhaps technically it's a pragmatics? The semantics-pragmatics distinction is a source of much consternation among experts and I do not pretend to authority on it. Also, I should say, this will just be a sketch based on Lewis (1979b). It is not intended to be a fully accurate summary of Lewis's views or the views of any other specific author.

[2] The salient sense of similarity can be changed by conversational contributions. For foundational literature on selection functions see Stalnaker (1999); Lewis (1973).

[3] A more serious account would use a sequence of sets of P-worlds, $f^{\mathbf{c}}_w(P) = (\Gamma_0, \Gamma_1, \dots)$, such that for any $0 < n \leq m$, $\Gamma_n \supseteq \Gamma_m$ and no worlds in Γ_n are more relevantly similar to w than any worlds in Γ_m. We would say that *were* P *would* Q is true in $\mathbf{c}$ iff there is some n such that for all $m \geq n$, $\Gamma_m \subseteq Q$. See Lewis (1973).

[4] We will continue to merrily ignore the multitude of objections to the possible worlds account of propositions. The literature is too vast to summarize. For a taste of these objections, I recommend (Merricks, 2015, Ch. 3) and (Lycan, 2019, Ch. 10). For a few classic defenses of the possible worlds account, I recommend Stalnaker (1984a), Stalnaker (1987), (Stalnaker, 1999, Part IV), and (Lewis, 1986, §1.4).

Backtracker Believed (BB): # Jackson (at the edge) believes if he were to jump, he would live.

This is not a conditional — it's an attribution of belief in a conditional, to Jackson. Is this attribution true in Harvey and Bert's context, after their remarks? Intuitively, no — BB is false, even in Harvey and Bert's context. Just consider what's going on with Jackson as he stands on the edge of the building — the sweaty palms, the fear, the intrusive thoughts, the racing pulse. Jackson simply does *not* believe that if he jumped he would live. When Bert says Jackson believes he'd live, Bert is speaking falsely.

But as mentioned, this grates against the formalism for conditionals in context just sketched. The proposition Backtracker expresses in **h** according to the vanilla formalism is one it is natural to assume Jackson does indeed believe. After all, there is no problem with thinking that Jackson, as he stands there at the edge, believes that he survives the jump in the nearest jump-worlds where he is disposed to only jump knowing he'll live.[5]

Here is a natural proposal to fix the formalism. Instead of using the contextually determined selection function, let us suppose that each agent's psychological state at a given time t somehow fixes a subjective selection function that should be used to determine the truth of ascriptions of conditional belief-at-t to them.[6] That is, to a first approximation, we can suppose that ⌜ S believes that if P then Q ⌝ is true in w at t iff S in w at t believes $\{w' : f_{w'}^{S,w,t}(P) \subseteq Q\}$.[7] Here the set of worlds is a conditional proposition determined in part by the *subjective selection function* $f^{S,w,t}$, which, intuitively, is just the selection function such that given a world w' and proposition P, $f_{w'}^{S,w,t}(P)$ is the set of P-worlds most similar to w' according to the sense of world-similarity subjectively salient to the agent S in w at t.[8] Applying this to **BB**, we can see easily why that belief attribution would

[5]More carefully: there is no problem with thinking that Jackson believes a proposition that is true iff he survives the jump in the nearest worlds where he jumps and yet is disposed to only jump if he knows he'll live.

[6]That such is the case is already suggested by the vanilla formalism, for it is natural to think the selection functions settled on in conversational context are a function of individual selection functions of the participants. I will not rely on this naive explanation in what follows, but it seems worth pointing out.

[7]One way in which this is an approximation is that it makes no mention of context. Context might indeed be required to determine the propositions expressed by P and Q in the ascription, but the idea is that it is not required to determine the selection function. We have, tacitly, been loose with our notation; on the left-hand-side P and Q stand for sentences, and on the right, for propositions expressed by those sentences in context.

[8]The following bears mentioning: we are not imagining that all agents with conditional beliefs are thinking explicitly about possible worlds or world-similarity relations as such. These things are parts of the model we use to

be false, since given what we know about Jackson, the worlds where he jumps that are subjectively salient to him are worlds where he dies.

Recently Grano and Phillips-Brown (2022) have defended an approach to desire ascriptions that makes use of precisely these subjective selection functions. They formulate their account in terms of *conditional belief sets*. A conditional belief set is a set determined by an agent in a world at a time together with a proposition.[9] The basic idea is that in addition to having a *belief set* $\text{Bel}_{S,w,t}$, which contains all and only the possible worlds compatible with their beliefs, an agent S in w at t has, for any possibly true proposition P, a *conditional belief set* $\text{ConBel}_{S,w,t}(P) = \{w' : \exists w'' \in \text{Bel}_{S,w,t}, f_{w''}^{S,w,t}(P) = w'\}$ that represents what they believe will (would) be the case if P is (were) true.[10] I need not rehearse the details of Grano and Phillips-Brown's paper here.[11] One interesting central aspect of their proposal bears highlighting, however. They derive their results for desire ascriptions using a principle that asserts a rather splendid analogy between belief sets and context sets (cf. Stalnaker (1975, 1978)). On the standard Stalnakerian account of conditionals in conversation, the public selection function operative in a conversation always selects worlds inside the context set if such there be — and this fact corresponds to how only indicatives are felicitous when the antecedent is compatible with what is presupposed, and subjunctives, or rather counterfactuals specifically, are felicitous only when the antecedent is contrary to the conversation's presuppositions. Grano and Phillips-Brown claim that the agent's subjective selection

---

describe the contents of agents' beliefs (and the truth conditions of belief ascriptions) — they are not, according to our model, ingredients represented in those contents.

[9] Though they do not draw the connection explicitly, the concept not only dovetails with the Stalnakerian account of conditional assertions in context, but with a familiar possible worlds approach to belief and belief ascription of the sort first formulated and defended by Hintikka (1969) and later embraced in various forms by Stalnaker (1999) and Lewis (1986) among many others.

[10] On the possible worlds approach to belief, an agent S in w at t believes P iff $\text{Bel}_{S,w,t} \subseteq P$, so it is easy to see that on this account S in w at t will believe $\{w' : f_{w'}^{S,w,t}(P) \subseteq Q\}$ just in case $\text{ConBel}_{S,w,t}(P) \subseteq Q$, that is, just in case their conditional belief set (for antecedent P) entails Q.

[11] The desire ascriptions that Grano and Phillips-Brown are interested in are those where the agent either believes that what they desire is the case, or believes it is not the case — e.g., cases where an agent wants it to be the case that P and either believes P or believes $\neg P$. The standard approach to desire ascriptions analyzes them in terms of the agent's preferences over worlds in their belief set, but this breaks when the relevant worlds are not in the belief set because the agent believes or disbelieves the content of the desire. Grano and Phillips-Brown propose using conditional belief sets instead. Very roughly, instead of using preferences on worlds in the agent's belief set to give truth conditions for desire ascriptions, they propose using preferences on worlds in the union of $\text{ConBel}_{S,w,t}(P)$ and $\text{ConBel}_{S,w,t}(\neg P)$, where P is the content of the desire.

function will always select worlds within their belief set when given an antecedent that is compatible with the agent's beliefs. That is, they propose that for any S, w, and t, if $\text{Bel}_{S,w,t} \cap P \neq \varnothing$, then $\text{ConBel}_{S,w,t}(P) \subseteq \text{Bel}_{S,w,t}$. The conditional belief set thus represents what S in w believes at t is the case if P is true just in case $P \cap \text{Bel}_{S,w,t} \neq \varnothing$, and otherwise represents what they believe would be the case if P were true.[12] The theory implies that no deep psychological difference separates indicative from counterfactual beliefs; both are a matter of selecting possibilities compatible with the antecedent. The difference amounts to a consequence of a principle — which could be read as constitutive or as normative — that selection is confined to worlds in the belief set whenever such worlds are compatible with the antecedent. In other words, the theory implies that when we consider antecents compatible with our beliefs, we consider — or we ought to consider — only possibilities we haven't ruled out.

It is interesting to ask what this principle, which Grano and Phillips-Brown call the *Stalnakerian Belief Set Constraint* (SBSC), implies about Jackson. Whether Jackson believes — or ought to believe — that if he jumped he would live turns on whether his beliefs have ruled out the possibility that he jumps and lives (e.g., because he lands on a cushion). If he jumps at no worlds in his belief set, then whether he believes he would live if he jumped depends on what possibilities outside his belief set — counterfactual by his lights — he subjectively selects when he considers his jumping. On the other hand, if he isn't sure he won't jump, and neither believes he will or that he won't, then we must look strictly at what happens in his belief worlds where he jumps. Now one not unreasonable possible explanation of his fear, his intrusive thoughts, etc., is that as he stands at the edge, he doesn't truly believe he won't jump, and in the worlds compatible with his beliefs where he jumps, he dies. If this is the case, then given that he has these belief worlds, it is a consequence of the SBSC that Jackson does not — or ought not — believe that if he jumped, he would live.[13] If we accept a basic norm connecting agents' belief sets to their knowledge or evidence, we could strengthen this and say that if it is incompatible with Jackson's knowledge and evidence that he

[12] Later in §11.2 we will see reason to relax this — specifically the first part — in view of future-directed conditionals.

[13] It seems we can express this (lack of) belief with either the subjunctive or the indicative mood. This seems to be just a quirk — and not at all an insignificant one for my purposes — of future-directed conditionals (cf. (Gibbard, 1981, p. 228) and DeRose (2010)).

jumps and lives, then this epistemic norm and SBSC jointly imply he epistemically ought not believe that if he jumped he would live.

10.3 Newcomb's Problem and Philosophical Methodology

As previously discussed (§1.12), Newcomb's Problem and its many offshoots divide philosophical decision theorists into two loose schools of thought. One side intuitively judges that the agent in the case, Robert, should one-box. The other side judges that he should two-box. This disagreement has been thought to bear directly on the philosophy of subjectively rational choice, and hence on decision theory. Both sides have for years assumed that intuitions about what it is rational for Robert to do in Newcomb's Problem motivate substantively different definitions of expected utility. The intuitions of one-boxers are thought (by many one-boxers) to support EDT, whereas the intuitions of two-boxers are thought (by many two-boxers) to support CDT. There seems thus to be a certain methodological premise driving philosophical debate in decision theory, namely that the ultimate solution to Newcomblike cases lies in formulating a definition of expected utility that implies that the intuitively rational choices indeed maximize expected utility. This methodological premise merits scrutiny. Decision theory should be, as far as possible, a neutral formal framework regimenting uncontroversial norms of subjectively rational choice. The persistent debate over the proper formulation of expected utility, and the substantive philosophical bent that the various formulations on offer seem to possess, looks to me very much like a result of failure to pursue this kind of methodological neutrality.

What might a return to methodological neutrality mean in decision theory? It would mean somehow validating both one-boxer and two-boxer intuitions at least so far as subjectively rational choice is concerned. How can this be done? My suggestion, to preview in somewhat loose terms, will be that one- and two-boxing can each be rationally required for different agents, depending on what their conditional beliefs are like. Consider a simplified example. Suppose that Robert has no beliefs about what he will do; hence, there are worlds in his belief set where he one-boxes, and worlds where he two-boxes. Suppose, however, that at all belief worlds where he one-boxes he gets the million, and at all belief worlds where he two-boxes he only gets the thousand. The SBSC implies that his conditional belief sets concerning his options must be confined to his belief

worlds, and thus Robert believes — or ought to believe — that if he one-boxes he will get the million and if he two-boxes he will only get the thousand. We may put the same thing in terms of subjunctive conditionals and say that Robert believes — or ought to believe — that if he were to one-box he would get the million and if he were to two-box he would only get the thousand. (The difference between the moods of the conditionals on this analysis is a superficial matter; it's the underlying belief and conditional belief sets that count.) I think we must say that given his conditional beliefs, Robert's subjective practical rational obligation is to one-box. Is it possible to criticize Robert's beliefs on epistemic grounds? Maybe! But that, I claim, should be the purview of epistemology, not of decision theory. Decision theory should take Robert's beliefs, insofar as they are consistent, as a given — and at the very least, there is nothing inconsistent in the description I have made of Robert's beliefs just now. Now let us suppose Robert's beliefs are different. Suppose that, like last time, Robert has no beliefs about what he will do; at some belief worlds he one-boxes, at others he two-boxes. But unlike last time, at some one-box belief worlds he gets the million and at others he does not, and ditto for two-box belief worlds. In this case Robert's beliefs make room for four possibilities (one-box&million, one-box&nothing, two-box&million+thousand, and two-box&thousand) rather than just two (one-box&million and two-box&thousand). In this case, Robert's conditional beliefs depend on his subjective selection function substantively. We will consider two possibilities. Let us suppose first that the sort of similarity that is salient to Robert is similarity in the laws and in the current state of the room, whatever it is, aside from himself.[14] So, if the million is on the table, then, ceteris paribus, worlds that evolve forward in accordance with the actual laws of nature and where the million is on the table are more similar by Robert's lights to the actual world — and mutatis mutandis if the million is not on the table. Now, Grano and Phillips-Brown's account tells us nothing about Robert's degrees of belief in conditionals, but there is an obvious way to extend it so as to derive such information. Just suppose that Robert's conditional belief sets come with probability functions controlled by his subjective selection function. That is,

[14]Intuitively this seems like enough to fix a similarity relation that should do the trick, but if we need to we can go the full distance to Lewis's "default" relation (cf. Lewis (1979a)), with its carefully ranked criteria and use of the "miracle" concept. Despite the case Lewis makes, I remain skeptical that anything as heavy-duty as that is presupposed in ordinary conditional thought.

suppose that when Robert makes a conditional supposition, the prior probability he assigned to any world w provisionally moves to whatever antecedent-world his subjective selection function maps w to. The result will be a probability function normalized on the antecedent proposition which it is natural to take to represent Robert's suppositional credences on the antecedent proposition. Now, combine this quick theoretical sketch with the above assumptions concerning Robert's beliefs and his subjective selection function. On either supposition (i.e., one-box, two-box), credence flows in a way that preserves the state of the room in the prior worlds and the laws. Hence, Robert's credence that the million is (would be) present on the supposition that he one-boxes equals his credence that it is (would be) present on the supposition that he two-boxes. Hence, the intuitively rational thing for him to do is to two-box and (by his own lights) grab the extra thousand. Consider a third variation. Suppose once more that Robert's belief set divides into the four above-mentioned possibilities. But this time suppose that the similarity subjectively salient to Robert is in the Predictor's success — a world where the Predictor is right (wrong) about Robert's choice will get mapped to a similar world in that respect.[15] Further, suppose that among his belief worlds, almost all of his credence goes to worlds where the Predictor's prediction is correct. Given these assumptions, almost all of Robert's credence on the supposition that he one-boxes goes to worlds where he receives the million, and by the same token, almost all his credence on the supposition that he two-boxes goes to worlds where he only gets the thousand. Here again, as with the first variation, it seems intuitive to me that Robert subjectively rationally ought to one-box — he is after all highly confident that if he one-boxes he will (would) receive the million, and highly confident that if he two-boxes he will (would) receive only the thousand.

There are details to be ironed out and objections to be addressed, but we have before us the basic gist of my proposal. Whether an agent is subjectively rationally required to one-box or two-box should not be something we expect to just fall out of the formal constraints of subjective practical rationality; it should be something we take to substantively depend on their suppositional credences, which depend on the feature of their psychology represented by the subjective selection function. Debate over what is rational to do in Newcomb's Problem should not be understood as

[15] See Horgan (1981) for a defense of the view that this sort of similarity is what's relevant in Newcomb's Problem.

concerning the nature of practical rationality, but rather the epistemology of suppositional credence. Those insisting that one-boxing (two-boxing) is required should see themselves as concerned with subjective rationality "all-things-considered" — where to be all-things-considered rational is to be not only subjectively practically rational but subjectively epistemically rational — and should be concerned to specifically argue that only suppositional credences that recommend one-boxing (two-boxing) are epistemically rational.

CHAPTER 11

SUBJECTIVE DECISION THEORY

11.1 Generalized Subjective Imaging and SDT

The preceding sketch of an extension of Grano and Phillips-Brown's theory of conditional belief is a proposal to adopt a specific sort of imaging theory (§1.10 and §5.2.2.4). Previously I dealt with imaging functions at a high level of abstraction, treating them schematically as components of decision theories. Now I wish to defend the idea that a specific sort of imaging function should be used as the basis for expected utility. The basic idea is, of course, to make use of a subjective selection function to define imaging, and to combine this with an appropriate version of the Stalnakerian Belief Set Constraint (§10.2). This can be accomplished in different ways depending on how one wants to understand subjective selection and imaging. As I've suggested, my proposal is meant to be programmatic. I am less attached to specifics than to these general guiding ideas. Nevertheless, I will put something specific on the table that represents my current favorite distillation of the approach I'm championing.

I favor Peter Gärdenfors's (1982; 1988) generalization of the selection function for its flexibility. On this sort of account, instead of a function that yields a possible world as value, we use a function that yields a probability function. So, instead of our subjective selection function $f^{S,w,t}$ we will have a probabilistic subjective selection function $\sigma^{S,w,t}$, where for any world w' and proposition P, the function $\sigma^{S,w,t}_{w'}(\bullet \parallel P)$ is a probability function such that (1) $\sigma^{S,w,t}_{w'}(P \parallel P) = 1$ and (2) if $w' \in P$ then $\sigma^{S,w,t}_{w'}(w' \parallel P) = 1$. The idea here is that on the supposition of P, the credence agent S in world w at time t assigns to w' is moved to worlds compatible with P in proportion to the probability $\sigma^{S,w,t}_{w'}(\bullet \parallel P)$ — that is, the prior probability of w' can, under the supposition of P, smear over multiple P-worlds, but with the caveat that if the world w' is a P-world, then its probability stays put, concentrated on itself.

I will assume the Stalnakerian Belief Set Constraint (§10.2) in a suitably updated form and hold that if $c_{S,w,t}(P) > 0$ then $\sigma^{S,w,t}_{w'}(w'' \parallel P) > 0$ only if $c_{S,w,t}(w'') > 0$.[1] Equivalently, if we let $\text{Open}_{S,w,t}$ stand for the set of worlds such that $w \in \text{Open}_{S,w,t}$ iff $c_{S,w,t}(w) > 0$, then we can say that if $P \cap \text{Open}_{S,w,t} \neq \varnothing$ then $\sigma^{S,w,t}_{w'}(\bullet \parallel P) = \sigma^{S,w,t}_{w'}(\bullet \parallel P \cap \text{Open}_{S,w,t})$. I will call this constraint the *Levian Constraint* inasmuch as it seems to first appear in print in Gärdenfors (1982) as a reported suggestion made by Isaac Levi to Gärdenfors for how generalized imaging could make for a "conservative" method of credence updating, in the sense that if a prior assigns a proposition credence one (i.e., certainty), then so will the image of that prior on any proposition to which the prior assigns a nonzero value.

For an agent S in w at t and propositions P and Q let us define the *subjective imaging function*:

$$\mathcal{I}^{S,w,t}(Q \parallel P) \stackrel{\text{df}}{=} \sum_{w'} \sum_{w'' \in Q} c_{S,w,t}(w') \sigma^{S,w,t}_{w'}(w'' \parallel P) = \sum_{w'} c_{S,w,t}(w') \sigma^{S,w,t}_{w'}(Q \parallel P). \tag{11.1}$$

Thus a reasonable concrete instance of the sort of decision theory I wish to defend will define expected utility like so:

$$\forall S, w, t, \forall O \in \mathcal{O}_{S,w,t}, \text{EU}_{S,w,t}(O) \stackrel{\text{df}}{=} \sum_{w'} \mathcal{I}^{S,w,t}(w' \parallel O) u_{S,w,t}(w'). \qquad \textbf{(Subjective EU)}$$

There is a natural way to extend this definition to a partition invariant form. Define subjectively imaged expected value like so: $\mathcal{V}^{S,w,t}(P \parallel O) \stackrel{\text{df}}{=} \sum_{w' \in P} \frac{\mathcal{I}^{S,w,t}(w' \parallel O)}{\mathcal{I}^{S,w,t}(P \parallel O)} u_{S,w,t}(w')$.[2] Then we have:

$$\forall S, w, t, \forall O \in \mathcal{O}_{S,w,t}, \text{EU}_{S,w,t}(O) \stackrel{\text{df}}{=} \sum_{S \in \mathcal{S}_{S,w,t}} \mathcal{I}^{S,w,t}(S \parallel O) \mathcal{V}^{S,w,t}(S \parallel O). \qquad \textbf{(Subjective EU)}$$

One might be initially unsure whether what I am calling a subjective imaging function is indeed a bona fide imaging function — that is, will such functions possess the constitutive characteristics stated in §1.10? Answer: they will, since we assume probabilistic subjective selection functions obey

[1] All the variables are universally quantified. Here $c_{S,w,t}(\bullet)$ is the credence function of S in w at t. We proceed according to the usual idealizing assumption that agents come with probabilistically coherent credence functions. Also, as usual, we conflate $\{w\}$ with w when convenient in our notation.

[2] Cf. §5.2.2.4.

the principle of centering (Grano and Phillips-Brown, 2022, p. 659), where for any S, w, t, we have $\sigma^{S,w,t}_{w'}(w' \parallel P) = 1$ for any w', P where $w' \in P$.[3] For if probabilistic subjective selection functions are centered, then it trivially follows that $\mathcal{I}^{S,w,t}(Q \parallel P) \geq c_{S,w,t}(QP)$ for all S, w, t, Q, P. Centering ensures that prior credence in antedent worlds stays put.

To give it a name, let us call the theory that says subjective practical rationality consists in maximizing subjective expected utility *Subjective Decision Theory* (SDT). I will defend this theory with the caveat, stated above, that I am more wedded to the idea that expected utility should be understood in terms of subjective imaging than I am to this specific formulation.

Those who accept SDT wish to separate the question of what subjective practical rationality requires from the question of what the rational epistemic perspective towards one's options is during deliberation. As the SDTer sees things, what is at stake in Newcomb's Problem is not what it is rational for the agent to do given her credences and desires — indeed, the SDTer says, the case does not actually directly pertain to the proper interpretation of expected utility or to what sort of expected utility maximization rule is correct. Once we have the agent's credences in practical conditionals (understood via her subjective selection function) and her utilities, we know what it is subjectively rational for her *to do*, regardless of whether her credences are *epistemically* rational. To an SDTer, what Newcomb's Problem does pertain directly to is precisely the epistemic rationality of one's subjective imaging function. From this viewpoint, debates between one- and two-boxers depend for their substantivity on the non-trivial assumption that whatever the epistemic norms governing subjective imaging may be, they cannot license characteristic imaging functions of both one- and two-boxers. This way of separating the definition of expected utility from the norms of conditional thought in deliberation runs counter to the conventional wisdom in decision theory. Hitherto, all sides have tended to agree that what is at stake in Newcomb's Problem is precisely how we should define expected utility.

[3] Centering here means roughly that the subjectively selected world given an antecent proposition that holds at that world is just that very same world.

11.2 Indicative, Subjunctive, and Modes of Supposition

SDT diverges from conventional wisdom in another way. It has been thought that one- and two-boxer accounts of subjectively rational choice align neatly with competing accounts of the conditionals of deliberation and with competing accounts of the correct deliberative mode of supposition. To an SDTer like me, these alignments are misconceived. For a sense of the orthodox perspective, here is James Joyce:

> The disparity between [CDT and EDT] has to do, rather, with a difference of opinion about the correct *epistemic perspective* from which rational agents should evaluate their options. [...] Evidential and causal decision theorists both maintain that actions should always be evaluated *under the supposition that they are performed*, but they have rather different notions of supposition in mind when they say this.
>
> Proponents of the evidential theory are speaking about supposition in the *indicative* or *matter-of-fact mode*. This process involves provisionally accepting some proposition as true in a way that preserves its evidential relationships. The hallmark of indicative supposition is that it never forces the supposer to revise any of her views about the "facts," that is, propositions that she is certain are true. Causal decision theorists, in contrast, believe that people should evaluate their actions by *subjunctively* supposing that they will be performed. Here the person is guided not by her antecedent views about "the facts" as they actually happen to be, but by her judgments of similarity and difference among merely possible situations. The contrast between these two modes of supposition is just the difference between asking, "What is the world like if I do perform A?" and asking, "What would the world be like were I to perform A?" (Joyce (1999))

For a more recent echo of this take, here is Eva et. al.:

> The most popular alternative to evidential decision theory — causal decision theory — replaces the use of indicative suppositions in the calculation with subjunctive suppositions. The debate between evidentialists and causalists in decision theory boils down to a dispute about which type of supposition is relevant for *ex ante* evaluations of options. (Eva et al. (2022))

These characterizations, which fit with well-known earlier framing of Lewis (1981b), must to some extent be right, and indeed, from the perspective of SDT, they are. To an SDTer like myself, fundamentally, one- and two-boxers disagree about the epistemology of subjective imaging, about the epistemology of deliberative supposition. But there are two important ways in which, from the perspective of SDT, such characterizations are not quite right. First, to view the one- versus two-boxer debate in terms of the indicative-subjunctive difference is misguided; the characteristic thoughts of one- and of two-boxers can be equally well captured using indicative or subjunctive conditionals. Second, it is somewhat confused to say that one- and two-boxer intuitions originate

from distinct "modes of supposition." There is a lot packed into this idea that merits critique. For starters, while it is true that one-boxer intuitions can derive from a certain unique "mode" of supposition, it is also true (1) that they do not strictly have to originate from the mode in question, and (2) that this mode is not really an entirely separate species of supposition from that which two-boxers use — rather, it is best treated as a kind of limiting case of subjective imaging, requiring no separate apparatus to model. Further, one-boxers have no exclusive claim to "matter of fact" supposition or to "antecedent views about the facts," and two-boxers are not the sole proprietors of "judgments of similarity and difference" among worlds. Rather, two-boxers have just as much of a right to consider their characteristic suppositions as proceeding from their antecedent views about the facts as one-boxers, and one-boxers can, and in some cases must, make use of supposition that non-trivially depends on "judgments of similarity and difference" among possible worlds. I turn now to discuss these points in more detail.

Start with conditionals. In an illuminating paper, Keith DeRose (2010) argues for the thesis that the *conditionals of deliberation* are indicative rather than subjunctive. DeRose stipulates that he classifies as "indicatives" whatever should be grouped for semantic treatment with the first of the following two well-known examples (from Adams (1975)):

(A) If Oswald didn't shoot Kennedy, someone else did.

(B) If Oswald hadn't shot Kennedy, someone else would have.

To my mind, DeRose's case gains a lot of its force from his exploiting the fact that when it comes to future directed conditionals, the same content seems expressible in either mood. Here is a pair DeRose uses to illustrate this:

- If I offer Eve £2,000 for her car, she will accept.
- If I were to offer Eve £2,000 for her car, she would accept.

As DeRose notes, it sounds contradictory to conjoin one of these with the other's complement:

- (#) If I offer Eve £2,000 for her car, she will accept; but, if I were to offer Eve £2,000 for her car, she would not accept.

- (#) If I were to offer Eve £2,000 for her car, she would accept; but, if I offer Eve £2,000 for her car, she will not accept.

The communicative purpose of the 'were'ed-up conditionals, DeRose suggests, is just to signal to the hearer that the antecedent is improbable.[4] Thus both future-directed indicative conditionals with action antecedents and their 'were'ed-up counterparts should be grouped with (A) above, making them both indicative in the stipulated sense. There is much more to the paper than this, but I will not summarize it. The important point for my purposes is just that, as DeRose argues (and as others[5] have noticed), not much separates future-directed conditionals in the indicative from those in the subjunctive mood.[6] Indeed, there seems little more than a pragmatic difference between them in most cases. Certainly, they are not by virtue of mood alone so divergent in meaning as the Adams conditionals (A) and (B) above. It seems quite natural to think that if either sort of future-directed conditional is appropriate to deliberation, then the other should be as well; and indeed, one wonders what sort of case could possibly be made for thinking either of the two conditionals above (about "Eve") might be off-limits or deprecated in rational deliberation. But there seem to me two further perhaps even more critical points. The first is that in Newcomb's Problem, nothing prevents us from stating conditionals a one-boxer should intuitively accept in the subjunctive and conditionals a two-boxer should intuitively accept in the indicative. The second is that if the choice between one- and two-boxing were about choosing between moods of conditionals, it should be possible to find paradigmatic one-boxer indicative conditionals and paradigmatic two-boxer subjunctive conditionals that everyone will intuitively assent to. We would expect this insofar as the competing intuitions correspond to the meanings of the conditionals themselves.[7] But this is not what we find on inspection. There are no uncontroversial one-boxer indicatives or uncontroversial two-boxer subjunctives. Consider:

---

[4]DeRose's term for a subjunctive version of a future-directed indicative conditional is ''were'ed-up conditional'.

[5]See, e.g., (Gibbard, 1981, p. 228).

[6]I quite like DeRose's classification scheme for "indicative" versus "subjuntive," but I will revert to the more expected usage.

[7]That is, there should be such uncontroversial examples if the dispute concerns not what conditionals best express the thoughts of deliberators in Newcomb's Problem but rather the appropriateness to practical deliberation of one or the other sort of thought — where the two sorts of thought are semantically encoded by the two different moods.

1. If I were to one-box, the million would be in the opaque box.

2. If I were to two-box, the million would not be in the opaque box.

3. If I one-box I will make less money than if I two-box.

4. If I two-box I will make more money than if I one-box.

5. If I one-box the opaque box will contain the same thing that it will contain if I two-box.

6. If I one-box, the million will be in the opaque box.

7. If I two-box, the million will not be in the opaque box.

8. If I were to one-box the opaque box would contain the same thing that it would contain if I were to two-box.

One-boxers should happily assent to the first two conditionals, subjunctives.[8] Two-boxers should happily assent to the next two, indicatives — one-boxers will reject them.[9] Moving down, I expect one-boxers to reject the fifth conditional, an indicative, and two-boxers to accept it. The sixth and seventh conditionals are characteristic one-boxer indicatives that as a pair are intuitively inconsistent with the fifth; it is natural thus to expect two-boxers to reject or at least withhold assent from one or both of those two. Lastly, one-boxers should reject the eighth conditional, a characteristic two-boxer subjunctive that is intuitively inconsistent with the first two, which as mentioned the one-boxer should accept. By contrast two-boxers will accept the eighth conditional, and therefore, should reject or at least withhold assent from one or both of the first two conditionals. Thus, we see that there are subjunctives a one-boxer should intuitively accept and a two-boxer should intuitively reject or at least withhold assent from one if not both of (1. and 2.) and indicatives a one-boxer should intuitively reject and that a two-boxer should accept (3., 4., and 5.). We see the indicatives (6. and 7.) that are paradigmatically associated with one-boxing do not command universal assent — a two-boxer will reject or at least withhold assent from one if not both of them. Lastly, we see

[8]The one-boxers I've chatted have all been happy to do so.

[9]These two are what Korzukhin (2014) calls *dominance conditionals*.

that a subjunctive (8.) paradigmatically associated with two-boxing does not command universal assent — a one-boxer will reject it.

Does one need to adopt SDT to make the above observations about conditionals? Certainly not. But SDT does seem particularly aptly equipped to predict and account for all of these intuitions, following as it does in the vein of Stalnakerian approach to conditionals, on which the mood of conditionals, particularly future directed conditionals like the ones relevant to decision-making, is a superficial feature by comparison with the underlying selection or imaging function. The way in which subjective imaging accounts for one- and two-boxers' endorsements of conditionals like those above will be made increasingly clear as we proceed.[10]

Now consider "modes" of supposition. That some difference separates one- and two-boxers' practical suppositions seems unarguable — something must account for the divergence in their intuitions after all, and where else is it to be found? But is this difference just that the former camp draws on a "matter of fact mode of supposition" and the latter on a "subjunctive mode of supposition"? If this were so, two things would seem to follow. First, there would of course be a question as to which of the two modes — if either — is more germane to rational choice and why. But second, it would seem to follow that the intuitions of either camp should be readily accessible in their full vivacity to everyone simply by switching from one mode to the other — for surely it is not to be assumed that matter of fact supposition is a faculty possessed only by one-boxers or that subjunctive supposition is the mental endowment of two-boxers alone. Nor is it terribly plausible that an entire mode of supposition is more readily accessible to one group as opposed to another. Rather, if distinct modes exist, every thinking adult is, presumably, quite familiar with both, if not under such labels. Thus, the first piece of evidence I would point to as telling against the propriety of positing separate modes of supposition is that intuitions in Newcomb's Problem do not seem flippable via some mental switch. It has been my experience that those drawn to one or the other

[10]It is natural to ask: how did the idea of a clean linkage between indicative conditionals and one-boxing and subjunctive conditionals and two-boxing originate and propagate, if it is so entirely misguided? I will not dive into the intellectual history, but it seems to have emerged in an era of work on conditionals heavily influenced by Ernest Adams (1975). My conjecture is that this imagined tight association was inferred from the undeniable affinity between one-boxing and conditional probability and Adams's having persuasively drawn a connection between the assertability of an indicative and the probability of its consequent conditional on its antecedent. It seems that this inference was not a sound one.

school of thought often do profess an ability to see things from the point of view of the other side, but they very often will qualify this by saying something to the effect that they consider that opposing point of view to be illusory or fundamentally confused or insufficiently responsive to reason. If intuitions in Newcomb's Problem originated from two fully separate modes of supposition, however, this is not the sort of thing one would expect to see. One would expect less devotion to particular intuitions, less assurance by those on either side that the others have it wrong. Instead, one would expect a shared perplexity in all those who grasp the basic details of the case — something such as we observe in reactions to certain optical illusions. But if, on the other hand, the diverging intuitions were a symptom not of two fully separate species of supposition accessible to all, but of a single underlying mode of supposition whose behavior tends to follow a certain idiosyncratic, personal, subjective tendency, it would seem natural that we would see reactions divide into stable groups according to what subjective tendencies were possible. It would also seem natural to find members of the groups subjecting each another's subjective tendencies to rational critique; seeking to defend what they consider to be the right view on balance, rather than recognizing the alternatives as simply two ways of looking at it. This is indeed how I think the SDTer, equipped with subjective imaging, should perceive the terrain.

Let us back up a bit. Aside from their supposed association with one- and two-boxing, with EDT and CDT, what are the two modes of supposition supposed to be like? Following Joyce (1999), we can say, first, that the "matter of fact" mode of supposition involves (1) provisionally conjoining the proposition supposed with one's beliefs, while (2) "preserving evidential relationships," such that (3) one is never required to provisionally revise one's beliefs about "the facts" to make the supposition coherent.[11] Second, subjunctive supposition involves (1) engaging in a sort of consideration that lies on a gradient whose lower endpoint consists in contemplating subjunctive conditionals with the supposition as antecedent, and whose upper endpoing consists in contemplating the negations of

[11]The second condition I take to mean that the provisional degree of confidence in any world in one's belief set induced by the supposition reflects the evidential bearing of the conjunction of the supposition with one's beliefs on the likelihood that that world is actual. The third condition I take to mean that to matter of fact suppose a proposition, that proposition must be compatible with one's current belief set — compatible, that is, with all the propositions one currently believes.

such conditionals, where (2) evidential relationships need not be preserved, and (3) the supposition need not be compatible with one's current beliefs.[12]

The two modes are thus thought to differ in three main respects. Let's consider the first respect: the logical relation of what is supposed to what is believed. One-boxers are supposed to suppose by conjoining with prior beliefs, two-boxers not so — they consider subjunctive conditionals, or something like that. One fact that seems of central importance to a critical assessment of this supposed difference is that very often, though indeed not always, all of a deliberating agent's options are doxastically open — the deliberating agent has no belief as to what they will do. In Newcomb's Problem in particular it is natural to consider the case from the perspective of one who is unsure what he will do. What difference can there then be between the logical relation of the one-boxer's options, as propositions, to what the one-boxer believes, and the logical relation of the two-boxer's options to what *he* believes? Are we to suppose that the two-boxer, when making his practical suppositions, is considering worlds he believes will not obtain, despite the fact (naturally assumed) that he has some belief worlds where he one-boxes and some where he two-boxes? Are we to suppose that the two-boxer is uniquely the sort who straightaway forms a belief as to what he will do, whereas the one-boxer is uniquely the sort to withhold such belief and thereby preserve belief worlds for either option to conjoin with? I find it difficult to credit any such difference, and am convinced there is none. For both one- and two-boxers, I will assume that practical suppositions conjoin with belief worlds.

There remain two other putative differences. First, one-boxers' characteristic mode of supposition is claimed to uniquely merit being called "matter of fact," to "preserve evidential relationships," and to be somehow undefined or impossible to deploy unless the supposition is compatible with the user's beliefs. Second, it is claimed not to involve "judgments of similarity and difference" among worlds. By contrast, two-boxers' characteristic mode of supposition is said to have the contrary features: it does not deserve to be called "matter of fact"; it does not necessarily preserve evidential relationships; it does not constitutively require that the supposed proposition be compatible with

[12]There is something odd in the idea that subjunctive supposition is to be understood in terms of the contemplation of conditionals rather than vice versa. One would expect one's account of supposition to shed light on what it means to contemplate a conditional rather than the other way around.

the agent's beliefs; and it *does* involve judgments of similarity and difference among worlds. I think this taxonomy of differences between these two supposed modes of supposition is confused in several ways, which I will now endeavor to explain.

What is "evidential" about one-boxers' mode of supposition? Talk of "evidential relationships" is invariably cashed out in terms of conditionalization.[13] Setting aside the question of whether conditionalization is the right way to understand the rational bearing of evidence on credence in decision situations, let us take for granted that in the sort of supposition most intimately associated with one-boxing, the provisional distribution of credence induced by a practical supposition is identical to the distribution that would result from conditionalizing on the supposed option. It does not follow from this that one-boxers' most characteristic suppositions are not to be accounted for via relations of similarity among worlds, that is, via Stalnakerian selection. Take for instance the case made by Andrew Bacon (2015) that under Stanakerian semantics the probability of a conditional can be predictably equal in any given context to the conditional probability of the consequent on the antecedent if we suppose that the selection function selects worlds at random from the context set.[14] He writes:

> A potentially illuminating way to think of the selection function is as picking out an antecedent world at random from the epistemically accessible worlds. In Stalnaker's theory a world is selected from the accessible antecedent worlds with an overriding preference for more similar worlds. On my understanding, however, the selection process has no preference for more similar worlds: we can think of it as having a preference for worlds that are more probable on the evidence, but this preference is not overriding but proportional to the probability.

To be sure, Bacon's way of understanding world selection is has little to do with any intuitively familiar concept of similarity, and he spells out his account's divergence from stereotypical similarity accounts in clear terms. But his account nevertheless constitutes proof of concept that imaging can track conditional probability if we relax our understanding of world selection. To show how one-boxers' evidential supposition can be modeled within our broadly Stalnakerian subjective imaging

[13] While this has become commonplace, it remains non-trivial. In the years since Lewis (1981c) introduced the concept of imaging, there have been at least a few who have suggested that credence should perhaps respond to evidence via imaging, at least in decision scenarios (see e.g., Fusco (2025)).

[14] By "at random" Bacon has in mind randomness in accord with prior probability.

framework, I will make a proposal similar in spirit to Bacon's. To introduce it, let me first say a bit about how I understand subjective selection.

One way to take an agent's probabilistic subjective selection function is as encoding aspects of the world that the agent assumes will tend to be conserved under any eventuality — these are what I will term *projectible* factors.[15] Taking there to be such factors seems to me intuitively bound up with viewing oneself as an inhabitant of a real moment of time in a concrete universe rather than as a sort of God's-eye looking down on the space of possibility from a removed and indifferent perspective. To take certain features to be projectible is to take the conservation of these features to guide the movement of credence into possibilities under supposition. On this construal, the probabilistic subjective selection function works in tandem with the agent's credence function to represent their doxastic state. Since the agent's selection function tells us what they take to be projectible, their credences do more than just tell us which possible worlds they regard as likely candidates to be the actual world — they also tell us what they take the current state of the projectible features of the world to be. If I take my cat Alice's mood to be a projectible feature and I put most of my credence on worlds where Alice is happy on Aug 21, 2025, then when I entertain a supposition my credence tends ceteris paribus to move to worlds where Alice is happy on Aug 21, 2025.[16] Given some range of projectible features how does probability move under a supposition? The natural thought, it seems to me, is that in moving from a given world w into a supposed proposition, it distributes evenly over worlds compatible with that proposition that match w in the projectible features, and if no world is a perfect match it moves more to worlds to the extent that they are more similar to w with respect to those features. Thus, the fewer features the agent takes to be projectible, ceteris paribus, the more the probability moved from a given world incompatible with a supposed proposition will tend to distribute uniformly over a greater measure of worlds compatible with the supposition. At the limit, where the agent takes no features to be

[15] I will also speak of projectible features; this is just for terminological variety. Also, while I prefer projection talk, we can — and I sometimes will — also use similarity talk and say projectible factors are the respects of similarity according to which credence is moved when the agent makes suppositions. I realize my characterization is a bit tendentious. While I do indeed doubt on familiar Humean grounds that a plausible notion of projectibility could ever be defined in purely probabilistic terms, I do not to presuppose that. The anti-Humean can thus take my talk as somewhat stipulative.

[16] This feature is a particular matter of fact, but projectible features could also be patterns, e.g., the color of swans.

projectible, probability moved into a supposed proposition from a given incompatible world should spread evenly over all the worlds compatible with that proposition. Here then we have identified a clear way that our subjective imaging framework can mimic conditionalization. Imaging mimics — or perhaps just is — conditionalization when the agent takes zero features of their world to be fixed, settled, projectible. When such an agent makes a supposition, the contours of her current distribution of credence to worlds compatible with the supposition remains unchanged because the credence she has in worlds compatible with the supposed proposition stays put[17] and the credence she has in worlds contrary to that supposition gets spread evenly over those prior contours. Is this a "different mode of supposition"? Yes and no. It is not an altogether different species of supposition, requiring different modeling tools than those afforded by an imaging framework; it is rather an edge case within that framework. Is it more "matter of fact" than supposition that involves non-trivial projection? I see no reason to think this; indeed, in one respect at least it appears rather less "matter of fact", inasmuch as it corresponds to a sort of God's-eye-view doxastic state where all features of the world are up for grabs, so to speak — one's confidence in various hypotheses as to the state of all features of the world around oneself can be completely obviated or inverted by the right supposition. But be that as it may.

Is this the only possible way to conduct one-boxer suppositions — i.e., by taking no features of the universe one inhabits to be projectible? No. A one-boxer might fixate on the reliability of the Predictor as a projectible feature, something to be preserved in their suppositions.[18] It's a bit of a non-trivial question what "reliability" is, but it seems not implausible to think it can be analyzed in terms of objective chance. To say that the Predictor is "highly reliable" can be taken to mean that the Predictor at the time of their prediction had a feature such that the objective chance of their prediction being correct is near one.[19] If a one-boxer places almost all of their credence in worlds where the Predictor is highly reliable in this particular sense, then when they suppose they will one-box, most of the credence from two-box worlds will flow to one-box worlds where the

[17] This is a consequence of centering.

[18] See Horgan (1981). Cf. note 15.

[19] For an approach to chance that would license this sort of talk, see Meacham (2005). See also Meacham (2014).

Predictor's prediction proves right and the million is present. Likewise, when they suppose they two-box most of their one-box credence will go to worlds where the Predictor is right and the million is absent. And of course in both cases most of their prior credence will already be in worlds where the Predictor is right. Here we have a way to non-trivially image one's credences and yet wind up with one-boxer suppositions.[20]

In order to generate one-boxer verdicts do both options have to be doxastically open? If the agent is certain that they will one-box, for instance, are all modes of supposition on which they will not get the million supposing they two-box somehow closed off? I do not think so. Intuitively one might think, *I am certain I will one-box, and if I were to two-box, the million wouldn't be there.* In our imaging framework this just requires, again, that one take the correctness or the reliability of the Predictor to be projectible. Hence when one images into two-box worlds — all of which, we assume, lie outside one's belief set and receive zero prior credence — one's credence clusters on worlds where the Predictor is correct, and the million is absent.[21] The same kind of thought is possible after the fact; one can say to oneself, while hand-counting the million, *just think, if I had two-boxed I'd only have the thousand.*[22] Whether we take such conditional thoughts to be rational is beside the present point; they present no problem to our imaging framework as attitudes capable of being modeled and understood.

So, one-boxer supposition, in various forms, is entirely intelligible from within the imaging framework, and indeed some instances of such supposition call for the imaging framework — how else to understand one-boxer suppositions that are contrary-to-belief?

What about two-boxer supposition? Naturally it is supposed to be at home in the imaging framework. Is it somehow less "matter of fact"? At this point perhaps it should be clear that this is

---

[20]Nor is it the only way. Another would be to take the correctness of the Predictor's prediction about this particular action to be projectible, which would lead to much the same place assuming most prior credence is in worlds where the Predictor is correct about what one will do.

[21]If one were to take no features to be projectible one could still image into the two-box worlds, but one's credence would distribute over them uniformly, as opposed to on worlds where the million is absent and the Predictor's prediction is correct. I am bracketing the issue of how to understand uniform distribution over infinite sets of worlds, but see Rubenstein (2025).

[22]The fact that this thought is possible ex post is an ancillary reason to think it's possible ex ante, in the way previously described.

not the case. The two-boxer, as I understand them, takes the current state of the boxes, whatever it is, to be a projectible feature. When they suppose they one-box, the credence that moves to one-box worlds does not flow to worlds where the million is present. Since it is mostly coming from worlds where the million is not present, it goes to one-box worlds where the Predictor got it wrong and the million is absent. We see the characteristic dominance mechanics in that whatever level of confidence the two-boxer has in the hypothesis that the million is present, that level of confidence gets more or less conserved whether they suppose they will one-box or that they will two-box; hence two-boxing is bound to look like the way to take home more money. The reasoning, in indicative and "matter of fact" terms, looks like this: *whatever is now in the opaque box will be there if I one-box and will be there if I two-box — so, if I two-box I will get strictly $1000 more.* There is no need to frame this as counterfactual reasoning; both options are doxastically open, the agent is considering what to do. Indeed, if we hold to our Levian Constriant as I think we must, and maintain that selection must be confined to doxastically open worlds whenever possible, we have no choice but to take the two-boxer's suppositions as entirely "matter of fact" whenever both options are doxastically open for them.

In this section my aim has been to challenge some prevailing wisdom concerning the linkage of one- and two-boxer intuitions with the moods of conditionals and with two supposedly separate modes of supposition. My main contentions have been: (1) there is no reason to think one-boxers (two-boxers) characteristically reason or express their deliberative conditional thoughts with indicative (subjunctive) conditionals — rather, one-boxers and two-boxers can both make use of conditionals of either mood; (2) one-boxers are not drawing on an entirely separate mode of supposition from two-boxers — rather, at most, they are using a limiting case of the same sort of supposition two-boxers use by taking their universe to have zero projectible features; (3) one-boxers' intuitions can be supported by non-trivially imaged credences — non-trivial similarities among possible worlds can guide one-boxers' suppositions, and in some cases must do so, e.g., when they consider ex post facto what would have happened had they acted differently and they maintain their characteristic one-boxer sense of reality; (4) there is no reason to consider two-boxers' suppositions as being less "matter of fact" than one-boxers — thinkers of both sorts make their characteristic suppositions by provisionally moving credence into possibilities for action that are doxastically open, and indeed

given our Levian Constraint (§11.1) we must take this movement of credence to be confined to doxastic worlds.

CHAPTER 12

APPLICATION TO SOME CASES

12.1 Newcomb's Problem

What is it rational to do in Newcomb's Problem? For an SDTer, the correct answer will vary from person to person depending on their credences in practical conditionals, that is, on their subjective imaging function. Speaking loosely, an SDTer simply says that it is rational to do what you are confident will make you the most money. If you are convinced (or close enough) that you will get the million if you one-box but not if you two-box, then the rational thing to do is to one-box. If on the other hand you are convinced you will make more money if you two-box than if you one-box, then the rational thing to do is to two-box. Formally the expected utilities look like this (making the usual assumptions about utilities):

$$\begin{aligned} \mathrm{EU}^{\mathrm{SDT}}_{S,w,t}(\mathrm{OB}) &= \mathcal{I}^{S,w,t}(\mathrm{M} \parallel \mathrm{OB})\mathcal{V}^{S,w,t}(\mathrm{M} \parallel \mathrm{OB}) + \mathcal{I}^{S,w,t}(\overline{\mathrm{M}} \parallel \mathrm{OB})\mathcal{V}^{S,w,t}(\overline{\mathrm{M}} \parallel \mathrm{OB}) \\ &= \mathcal{I}^{S,w,t}(\mathrm{M} \parallel \mathrm{OB}) \times 1,000,000 \end{aligned} \tag{12.1}$$

$$\begin{aligned} \mathrm{EU}^{\mathrm{SDT}}_{S,w,t}(\mathrm{TB}) &= \mathcal{I}^{S,w,t}(\mathrm{M} \parallel \mathrm{TB})\mathcal{V}^{S,w,t}(\mathrm{M} \parallel \mathrm{TB}) + \mathcal{I}^{S,w,t}(\overline{\mathrm{M}} \parallel \mathrm{TB})\mathcal{V}^{S,w,t}(\overline{\mathrm{M}} \parallel \mathrm{TB}) \\ &= \mathcal{I}^{S,w,t}(\mathrm{M} \parallel \mathrm{TB}) \times 1,001,000 + \mathcal{I}^{S,w,t}(\overline{\mathrm{M}} \parallel \mathrm{TB}) \times 1,000 \end{aligned} \tag{12.2}$$

Whether there are any epistemological constraints on $\mathcal{I}^{S,w,t}(\mathrm{M} \parallel \mathrm{OB})$, $\mathcal{I}^{S,w,t}(\mathrm{M} \parallel \mathrm{TB})$, and $\mathcal{I}^{S,w,t}(\overline{\mathrm{M}} \parallel \mathrm{TB})$, the SDTer sets to one side when determining what it is rational for agent S in w at t to do. The subjectively rational thing to do is to act on one's beliefs as to what will most likely yield maximal utility, regardless of what those beliefs are. Naturally, there is *something* irrational, broadly speaking, about acting on irrational beliefs, but it is a different species of irrationality from

that which occurs when one fails to maximize expected utility, given what one's credences and desires in fact are. It is the latter sort of rationality which in the eyes of the SDTer should strictly be the target of decision theory.

Now, it may be that the agent is such that

$$1 \approx \mathcal{I}^{S,w,t}(\text{M} \parallel \text{OB}) >> \mathcal{I}^{S,w,t}(\text{M} \parallel \text{TB}) \approx 0. \tag{12.3}$$

In this state, they have one-boxer credences. It is subjectively rational for them to one-box. On the other hand, it may be that the agent is such that

$$\mathcal{I}^{S,w,t}(\text{M} \parallel \text{OB}) \approx \mathcal{I}^{S,w,t}(\text{M} \parallel \text{TB}). \tag{12.4}$$

In this state, they have two-boxer credences. It is subjectively rational to two-box.

Is this the "solution" to Newcomb's Problem? *How can it be,* one might well ask, *if it tells us nothing about what credences are rational?* If a "solution" must resolve every question that arises out of a paradoxical (or quasi-paradoxical) case, then indeed this is no solution. But I suggest it as a solution at the level of decision theory. If correct, it shows that most of the work required to "solve" Newcomb's Problem is in the epistemology of credence in conditionals. It shows that what's truly at issue is not what to do in the situation, so much as what to believe (or to be confident in) concerning the million. For an SDTer the question of what it is rational to do in Newcomb's Problem is trivial; the question of what it is rational to believe (or to be confident in) is non-trivial, and should take center stage.

12.2 Lewis's Pauper Problem

In a well-known letter to Wlodek Rabinowicz (2009), David Lewis describes a hypothetical case that he considers to be more of a problem for CDT than Newcomb's Problem. Lewis writes:

> I believe in the logical possibility of time travel, precognition, etc., and I see no reason why suitable evidence might not convince a perfectly rational agent that these possibilities are realized, and in such a way as to bring him news from the future. My worry is a different one. It seems to me completely unclear what conduct would be

> rational for an agent in such a case. [...] You know that spending all you have on armour would greatly increase your chances of surviving the coming battle, but leave you a pauper if you do survive; but you also know, by news of the future that you have excellent reasons to trust, that you will survive. (The news doesn't say whether you will have bought the armour.) Now: is it rational to buy the armour? I have no idea — there are excellent reasons both ways. And I think that even those who have the correct two-box'ist intuitions about Newcomb's problem may still find this new problem puzzling. That is, I don't think that the appeal of not buying the armour is just a misguided revival of [one-box'ist] intuitions that we've elsewhere overcome.

The trouble this case raises for CDT arises from the way the evidence from the future does not bear on the expected causal efficacy of buying armor, though it does intuitively seem to bear on what it is rational to do. In other words, the trouble Lewis evidently sees (despite his remark that "there are excellent reasons both ways"), is that it looks clearly irrational to make the purchase but not at all clear how CDT can permit that.

What does SDT say about this case? Notably, SDT claims rational choice depends not on one's credences concerning the specifically causal efficacy of one's options, but rather concerning practical conditionals. Moreover, SDT implies that an agent's credences in such conditionals are governed by the Levian Constraint (§11.1) such that when one makes a practical supposition, any credence moved must be confined to doxastically open worlds so far as possible. Thus, it is not difficult to see how SDT can imply that it is irrational to make the purchase; indeed, this implication does not depend on any one delicate assumption. For simplicity assume the utility of life-and-purchase-worlds is one, of life-and-no-purchase worlds is one plus δ, and of death-worlds is zero. Here are the expected utility according to SDT:

$$\begin{aligned} \mathrm{EU}^{\mathrm{SDT}}_{S,w,t}(\mathrm{Buy}) &= \mathcal{I}^{S,w,t}(\mathrm{L} \parallel \mathrm{Buy})\mathcal{V}^{S,w,t}(\mathrm{L} \parallel \mathrm{Buy}) + \mathcal{I}^{S,w,t}(\overline{\mathrm{L}} \parallel \mathrm{Buy})\mathcal{V}^{S,w,t}(\overline{\mathrm{L}} \parallel \mathrm{Buy}) \\ &= \mathcal{I}^{S,w,t}(\mathrm{L} \parallel \mathrm{Buy}) \end{aligned} \tag{12.5}$$

$$\begin{aligned} \mathrm{EU}^{\mathrm{SDT}}_{S,w,t}(\overline{\mathrm{Buy}}) &= \mathcal{I}^{S,w,t}(\mathrm{L} \parallel \overline{\mathrm{Buy}})\mathcal{V}^{S,w,t}(\mathrm{L} \parallel \overline{\mathrm{Buy}}) + \mathcal{I}^{S,w,t}(\overline{\mathrm{L}} \parallel \overline{\mathrm{Buy}})\mathcal{V}^{S,w,t}(\overline{\mathrm{L}} \parallel \overline{\mathrm{Buy}}) \\ &= \mathcal{I}^{S,w,t}(\mathrm{L} \parallel \overline{\mathrm{Buy}}) + \mathcal{I}^{S,w,t}(\mathrm{L} \parallel \overline{\mathrm{Buy}})\delta \end{aligned} \tag{12.6}$$

It is rather hard to see how an agent in this case could fail to be such that

$$\mathcal{I}^{S,w,t}(\mathrm{L} \parallel \mathrm{Buy}) \approx \mathcal{I}^{S,w,t}(\mathrm{L} \parallel \overline{\mathrm{Buy}}). \tag{12.7}$$

If we suppose that the agent, having received this time-traveler's evidence, is now certain that they will live, then Eq. 12.7, replacing '≈' with equality, is an immediate consequence of the Levian Constraint. If we suppose the agent is not certain that they will live, it is natural to expect them to take the pattern of wearing armor's being protective as projectible, and the question comes down to whether their confidence that they will live if they buy the armor is greater than δ times their confidence that they will live if they do not. This may be close. SDT can still imply that the agent rationally must forgo the purchase, for instance, if they are all but certain they will live, and the armor is costly. Regardless, the case remains problematic for (some formulations of) CDT, at least insofar as according to that theory (or those formulations) it is *never* close, so to speak, whether the purchase is rationally required.

12.3 Elga on Betting on the Laws

Adam Elga (2022) has recently raised a challenge for all suppositional decision theories — a class to which SDT belongs along with EDT and CDT — based on an influential species of hypothetical cases introduced by Arif Ahmed (2013, 2014) as putative counterexamples to CDT. I will follow Elga's presentation, which adheres to the spirit if not the exact letter of Ahmed's cases. The blast radius, so to speak, of Elga's argument is wide. His aim is to show that either the intuitive verdict in at least one of two Ahmed-style cases must be rejected or suppositional theories as a whole must be deemed incapable of licensing those intuitions. Elga's argument is valid. The response I will recommend on behalf of SDT is that one of the two intuitive verdicts ceases to be intuitive when you hold up a magnifying glass. So, I claim, we reject that verdict and all is well. Now, Ahmed has argued that CDT is bound to endorse both of these verdicts; he denies that this as a problem for all suppositional theories, since one of the intuitive verdicts seems to be intuitive mainly for CDTers like Elga, and Ahmed is an EDTer to the marrow. It is indeed this disputed verdict that I will deny as well. But I disagree that this response is tantamount to embracing EDT. The right explanation for the verdicts is not that rational decision is governed by evidential considerations, but rather that it is a matter of subjective practical supposition.

Elga's first case is his version of Ahmed's *Betting on the Laws*:

Betting on the Laws: Let D be the proposition that the Laws of Nature imply Newtonian physics. Suppose agent S in w at t has credence greater than 0.5 in D. Suppose this agent is facing a choice whether to raise his hand or not. Raising his hand is an assertion that D is true; not raising his hand is an assertion that D is not true. This agent only cares about this assertion's truth. Worlds where his assertion is true get utility one; worlds where his assertion is false get utility zero.

The only rational thing for this agent to do is raise his hand, right? This is Elga's first intuitive verdict. I agree, provided we make the additional assumption that this agent believes D is no more likely to be true if he raises his hand than it is if he doesn't. Elga suggests that the intuition is dependent on a similar but orthogonal claim, namely that the agent believes the laws are outside his causal influence. As an SDTer, I don't think it strictly matters what the agent thinks about causal influence in any sense; rather, all that matters is what he thinks will be the case if he raises his hand versus if he doesn't.[1] We will get this intuitive verdict from SDT straightforwardly provided that $\mathcal{I}^{S,w,t}(D \parallel \text{Raise}) = c_{S,w,t}(D) = \mathcal{I}^{S,w,t}(D \parallel \overline{\text{Raise}})$, and indeed nothing prevents an agent from imaging in this way on our account.[2]

So far, so good. Now for the other case. This is Elga's variant on Ahmed's *Betting on the Past*:

Newcomb on the Past: We're with a doxastic twin of the previous agent. This new agent S^* in w^* at t^* has the same credences and (we will assume) the same imaging function as the agent in the previous case. He is faced with qualitatively the same options: to raise his hand or not.[3] His utilities are different. This new guy doesn't care about making a true assertion; the only things he cares about are the truth of a proposition H and avoiding the effort of raising his hand. H is a disjunction of all propositions P about a specific date in the past, say Jan

[1] It bears reiterating that I both reject the conventional wisdom that talk of credence in indicatives has to be cashed out in terms of conditional probability and that I embrace a subjective imaging approach instead. Cf. §11.2.

[2] This is not the only way to get this result from SDT, but it's the most obvious.

[3] We can even take S and S^* to be the same guy — or at least modal counterparts of one another; we can do the same for the times t and t^*. These assumptions allow us to say the agents' options are literally the same. I believe this is Elga's intention.

1, 1900, such that PD implies he raises his hand.[4] (Thus, not only does HD imply he raises his hand; $\overline{H}D$ implies he doesn't.) Worlds where H is true get utility M and worlds where he doesn't raise his hand get utility T, and worlds where both are true get $M+T$. We are to assume that $M >> T > 0$.

Elga calls this Newcomb on the Past because, he suggests, the same sort of causal dominance reasoning that CDTers think we must employ in Newcomb's Problem should work here. If we suppose that the agent is certain the truth of the proposition H is out of his control, he can disregard the M; he knows he gets the T if and only if he chooses not to raise his hand, so that's what he rationally must do. Elga gives a short and clear proof that no suppositional theory can imply both that the agent in *Betting on the Laws* must raise his hand and that the agent in *Newcomb on the Past* must not raise his hand.[5]

It is Elga's intuitive verdict in *Newcomb on the Past* that Ahmed rejects, and that I too reject. But our reasons are, I suspect, not aligned. For Ahmed and for EDT, it is natural to think that the agent in Newcomb on the Past should raise his hand, for given his confidence in D raising his hand is evidence of H, and his overriding desire is for the truth of H. This is of course of a piece with EDTers' verdict in the original Newcomb. For the SDTer, there is a more basic argument to be made. As Elga himself shows, if D is suppositionally independent of hand raising and the agent has over 0.5 credence in D, it follows that H is suppositionally dependent on hand raising, on *any* suppositional theory. It is easy to see why. D and raising entails H; D and not raising entails $\overline{H}$. Hence the agent's credence in H given the supposition that they raise must be greater than 0.5 (ex hypothesi their prior in D is preserved under supposition); but by the same token, their credence in $\overline{H}$ given the supposition that they do not raise must also be greater than 0.5. Hence if this agent thinks D is just as likely if they raise their hand as if they don't, and they have over half credence in D, they are logically bound to think H is more likely than not if they raise and $\overline{H}$ is more

[4] D is the same as before.

[5] I think Elga's proof is sound, but his spin on it is perhaps slightly misleading. A crucial step involves assuming that M can be set to an arbitrarily larger value than T. Assuming this is tantamount to assuming that S^* in w^* at t^* cares about the truth of H infinitely more than about keeping his hand down. If we reject this assumption Elga's conclusion is blocked. But the case appears no less problematic given that assumption, so we cannot rely on this point to deal with his argument.

likely than not if they do not. For an SDTer, this is a decisive argument in favor of raising, given these assumptions. *But the agent believes H is totally outside their causal influence,* protests the CDTer.[6] "So what," replies the SDTer. "We are not causation fetishists." What matters is what the agent thinks *will/would be the case* if they raise their hand versus what they think *will/would be the case* if they do not. Given our assumptions, the agent is bound by logic to think something they overridingly (if rather insanely) want will/would likely be true if they raise their hand, and will/would likely not be true if they do not. So raising is rational in *Newcomb on the Past*, says the SDTer, *pace* Elga. Indeed, for the SDTer, Elga's label for the case is misleading; for in the original Newcomb's Problem, the presence and the absence of the million are each logically compatible with both of the agent's options.

12.4 Hedden's Counterfactual Decision Theory

SDT can deliver the sort of verdicts one expects from CDT when an agent's subjective imaging function treats aspects of their surroundings that evidentially depend on their choice as projectible (§12.1), but it is not a "causal" theory — it has no built-in sensitivity to intuitive relations of causal dependence or influence. It is a theory built to be sensitive to agents' credences in practical conditionals — more literally, to their practical suppositions. This insensitivity to causality per se is at its most evident in cases like the Pauper's Problem (§12.2) and Newcomb on the Past (§12.3). A theory according to which subjective rationality tracks agents' expectation of causal influence over desired outcomes would presumably enjoin buying armor in the Pauper's Problem and keeping one's hand down in Newcomb on the Past. To reiterate my stance on those verdicts: such fixation on causation seems pathological.

I am not the first to combine a willingness to dismiss causation as such from decision theory with rejection of EDT. Brian Hedden (2023) has proposed that anti-evidentialists embrace a decision theory of counterfactuals, where the interpretation of these is guided not by any commitment to the supposed centrality of causation to decision, but rather by independently motivated semantic

[6] This is not to say that it falls out of any specific formulation of CDT that one should keep one's hand down in this case; rather, it is (in part) precisely this persistent intuition that seems to many CDTers (e.g., Elga) to be the nagging problem with this case and others like it.

or metaphysical considerations.[7] Given the formal proximity of Hedden's theory to SDT and the spirit, highly congenial to SDT, of certain intuitions he wishes to marshal in support of his theory, some comparison seems warranted between the two theories. In this section I'll undertake this comparison and argue on behalf of SDT.

Hedden champions the formalism of Gibbard and Harper (1978) and Stalnaker (1981), which defines expected utility like so:

$$\mathrm{EU}^{\mathrm{CNDT}}_{S,w,t}(O) \stackrel{\mathrm{df}}{=} \sum_{X \in \mathcal{X}_{S,w,t}} c_{S,w,t}(O \mathrel{\square\!\!\rightarrow} X) \times \mathcal{V}_{S,w,t}(X). \tag{12.8}$$

Historically, this has been classed with CDT as one of its several main formulations. Hedden rejects this grouping as a mistake and argues that a theory based strictly on Eq. 12.8, i.e., a *Counterfactual Decision Theory* (CNDT), diverges wholesomely from CDT in cases of causal overdetermination, constitution, and in Ahmed-style cases.

Hedden's discussion seems to me to point in different directions at different places in his paper. In his discussion of causal overdetermination and constitution, he seeks to show that the theory of Gibbard and Harper is not really a "causal" theory, and that indeed, on the most plausible construal of their intuitive content, the leading forms of CDT are best understood not as "causal" theories per se but as theories centered on *counterfactual dependence*.[8] He does not, in this portion of his paper, challenge any of the verdicts of CDT. His point seems to be about the intuitive meaning

[7] I am summing up what I take to be Hedden's proposal; he doesn't put it in exactly these terms.

[8] For a case of causal overdetermination of the sort Hedden envisions, suppose you want a disposible cup removed from your desk. You could throw it away in the waste paper basket next to you at no cost to yourself (let us assume), or do nothing, in which case the janitor who is right this moment passing by your desk will remove it. Intuitively, either option is rational, Hedden says, but a truly "causal" decision theory would have to imply strangely that you rationally must remove it yourself, as that's the only way for your action to be a cause of what you desire. [Bracketing abstruse complications to do with causally efficacious omissions. Side-note: Hedden uses a structurally analogous (and more familiar) case involving throwing rocks at windows.] Cases of constitution, as Hedden understands them, are cases where you want to do an action for its own sake; his example going for a walk. Suppose you want to want to walk not to feel good or stimulate your mind or exercise, but just to walk. The alternative is to stay indoors, which you believe will cause you to sit and watch YouTube shorts, which will please you slightly. Surely so long as we assume your desire to walk is sufficiently strong, the intuitively rational option can be to walk rather than stay in; but, Hedden says, a truly "causal" theory will never imply this since causation is irreflexive and thus walking will not cause anything you desire.

we should assign to the formalism. I have nothing much to say about this.[9][10] In his discussion of Ahmed-style cases, on the other hand, he seems not to want to intuitively reframe an existing theory so much as offer an interesting new theory. It is this portion of his paper that will be our focus as I continue my explanation and sales pitch on behalf of SDT.

Hedden's first case is a version of Ahmed's *Betting on the Past and the Laws* (cf. §12.3):

Hedden's Bet: You must take one of two bets. Bet 1 nets \$1 if Q is true and −\$10 if Q is false. Bet 2 nets \$10 if Q is true and −\$1 if Q is false. Q is the proposition that the initial state of the universe and the laws entail that you will take Bet 1. You are certain the laws are deterministic. Hence you are certain that if Q is false, the initial conditions and the laws entail you will take Bet 2.

We note that in this case, Bet 2 seems to causally dominate Bet 1, insofar as Q is not under the agent's control — let us assume the agent is certain of this fact — and whether Q is true or false Bet 2's net is higher. Thus a causalist like Elga (cf. §12.3) will presumably intuit that taking Bet 2 is rationally obligatory in this case. But, Hedden says, intuitively Bet 1 is the only rational choice here. I agree, and so does SDT.

We make the usual assumption about utilities matching monetary values. We also suppress the usual notation for agent, world, and time. To see that SDT entails taking Bet 1, consider:

[9]See Gallow (2024) for a thoroughly unfavorable critique of Hedden's arguments and (Ahmed, 2014, §5.3.3) for some argumentation parallel to Hedden's with a very similar gist.

[10]Hedden seems to presume that Gibbard and Harper's theory is meant to be neutral with respect to the semantics of counterfactuals, but this is incorrect. In note 2 of their paper, Gibbard and Harper (1978) write, "Although the rough treatment of counterfactuals we propose is similar in many respects to the theories developed by Stalnaker and Lewis, it differs from them in some important respects. Stalnaker and Lewis each base their accounts on comparisons of overall similarity of worlds. On our account, what matters is comparative similarity of worlds at the instant of decision. Whether a given a-world [viz. antecedent world] is selected as W_a depends not at all on how similar the future in that world is to the actual future; whatever similarities the future in W_a may have to the actual future will be a semantical consequence of the laws of nature, conditions in W_a at the instant of decision, and actual conditions at that instant. (Roughly, then, they will be consequences of laws of nature and the similarity of W_a to the actual world at the instant of decision.) We consider only worlds in which the past is exactly like the actual past, for since the agent cannot now alter the past, those are the only worlds relevant to his decision." They then cite Lewis (1979a) as offering a theory on which the similarity relation they have in mind is supposed to be a "deep consequence of general facts about the world" rather than a stipulation.

$$\begin{aligned} \mathrm{EU}^{\mathrm{SDT}}(\mathrm{Bet1}) &= \mathcal{I}(\mathrm{Q} \parallel \mathrm{Bet1})\mathcal{V}(\mathrm{Q} \parallel \mathrm{Bet1}) + \mathcal{I}(\overline{\mathrm{Q}} \parallel \mathrm{Bet1})\mathcal{V}(\overline{\mathrm{Q}} \parallel \mathrm{Bet1}) \\ &= 1 \end{aligned} \tag{12.9}$$

$$\begin{aligned} \mathrm{EU}^{\mathrm{SDT}}(\mathrm{Bet2}) &= \mathcal{I}(\mathrm{Q} \parallel \mathrm{Bet2})\mathcal{V}(\mathrm{Q} \parallel \mathrm{Bet2}) + \mathcal{I}(\overline{\mathrm{Q}} \parallel \mathrm{Bet2})\mathcal{V}(\overline{\mathrm{Q}} \parallel \mathrm{Bet2}) \\ &= -1 \end{aligned} \tag{12.10}$$

Eq. 12.9 holds given the Levian Constraint (§11.1). The agent is certain the laws are deterministic; hence they are certain there are no $\overline{Q}$-worlds where they take Bet 1; hence regardless of what their subjective selection function looks like, their credence imaged on Bet 1 must go to worlds where they win \$1. Eq. 12.10 holds because Q is logically incompatible with taking Bet 2; hence, again, regardless of what their subjective selection function looks like, their credence imaged on Bet 2 must go to worlds where they lose \$1. The intutive story here is much the same as the one I told concerning Elga's *Newcomb on the Past*. Here the agent is certain that if they take Bet 1, they will/would win \$1, and certain that if they take Bet 2, they will/would lose \$1. For the SDTer, that's a knockdown argument for taking Bet 1.

Hedden's rationale for taking Bet 1 differs from that of SDT and of EDT. He joins with SDTers in rejecting EDT's characteristic support for Bet 1 on the grounds that taking it yields better evidence of a desirable outcome than does taking Bet 2. The support he gives for taking Bet 1 is supposed to fall out of reasoning about the semantics for counterfactual assertions. Hedden has in mind two competing accounts. On both accounts, a counterfactual $A \boxrightarrow B$ is true at a deterministic $\overline{A}$-world w when the "nearest" A-worlds to w are B-worlds.[11] On the first sort of account, which we may call the *Lewis-style*[12] account, the "nearest" A-worlds to w will have initial conditions matching those of w but will have different laws. On the other sort of account, which we may call the *Dorr-style*[13] account, it's the opposite: the "nearest" A-worlds to w will have laws of nature matching w but will have different initial conditions. Hedden reasons that whichever sort of account is correct, his CNDT will recommend Bet 1. Is he right? Let us try to reconstruct the reasoning. The agent divides their credence between the hypothesis that they will take Bet 1

[11] We gloss over hairy irrelevant details of these accounts.

[12] See Lewis (1979a).

[13] See Dorr (2016). See also Bennett (1984, 2003), Loewer (2007), and Goodman (2015).

and the hypothesis that they will take Bet 2. As noted, they are certain that all Bet 1 worlds are Q-worlds where they win a dollar and certain that all Bet 2 worlds are $\overline{Q}$-worlds where they lose a dollar. Now, Bet 2 $\boxrightarrow Q$ is logically false. Hence, let us assume the agent assigns zero credence to that counterfactual. Hedden tells us the agent is certain their world is deterministic. This implies all the Bet 1-worlds they assign credence to are Bet 1 $\boxrightarrow Q$-worlds, but does it imply that all the Bet 2-worlds to which they assign credence are Bet 1 $\boxrightarrow Q$-worlds? These worlds are deterministic worlds where the initial conditions and laws entail that the agent takes Bet 2. Now, if a Dorr-style account of counterfactuals is true, then it seems these worlds must all be Bet 1 $\boxrightarrow Q$-worlds, since the nearest Bet 1-world to any one of them will have deterministic laws but initial conditions that yield Bet 1, making it a Q-world. But suppose instead that a Lewis-style account of counterfactuals is true. In that case, the nearest Bet 1 worlds to any one of these deterministic Bet 2 worlds will have the same initial conditions, but different laws. It is not logically forced that these laws must be deterministic. Indeed, for all that's assumed, they may all be indeterministic worlds where Bet 1 happens as a matter of chance, making them $\overline{Q}$-worlds. Suppose that is so. It follows that $\mathrm{EU}^{\mathrm{CNDT}}(\text{Bet 1}) = c(\text{Bet 1}) - 10c(\text{Bet 2}) = 11c(\text{Bet 1}) - 10$. Since $\mathrm{EU}^{\mathrm{CNDT}}(\text{Bet 2}) = -1$, CNDT will require taking Bet 2 iff $c(\text{Bet 1}) < 9/11$.

Let us inspect Hedden's own summary of how the reasoning for Bet 1 is supposed to proceed for CNDT when a Lewis-style account is assumed. Hedden suggests we model the relevant space of possibility over which the agent's credences are distributed as a partition (§1.3) with four cells: $\{A \stackrel{\mathrm{df}}{=} H_1L_1; B \stackrel{\mathrm{df}}{=} H_1L_2; C \stackrel{\mathrm{df}}{=} H_2L_1; D \stackrel{\mathrm{df}}{=} H_2L_2\}$. Here H_1 and H_2 are disjoint initial states, and L_1 and L_2 are disjoint sets of laws. He says Bet 1 is entailed by A and by D, and Bet 2 is entailed by B and by C. He also says $Q = A \vee D$ and $\overline{Q} = B \vee C$. He writes:

> Suppose first that H_1 is actual. In that case, if you were to take Bet 1, then L_1 would be true (along with H_1), and so Q would be true, and so you would win \$1. If you were to take Bet 2, then L_2 would be true (along with H_1), and so $\overline{Q}$ would be true, and so you would lose \$1. Suppose next that H_2 is actual. In that case, if you were to take Bet 1, then L_2 would be true (along with H_2), and so Q would be true, and so you would win \$1. If you were to take Bet 2, then L_1 would be true (along with H_2), and so $\overline{Q}$ would be true, and so you would lose \$1. So no matter how things are with the initial state of the universe, if you were to take Bet 1, you would win \$1, while if you were to take Bet 2, you would lose \$1. So you should take Bet 1.

The problem with this is the assumption that $\overline{Q} = B \vee C$ and thus entails Bet 2. From the assumption that $Q = A \vee D$ it follows that L_1 and L_2 are deterministic sets of laws, but $\overline{Q}$ contains worlds with indeterministic laws, since it is just the negation of the proposition that the initial conditions and the laws entail Bet 1. Hedden seems to have conflated the assumption that the agent is certain the world is deterministic with the claim that there are no indeterministic possible worlds where Bet 1 is true. These are not the same, and this matters insofar as part of the agent's credence in the counterfactual Bet 1 $\Box\!\!\rightarrow \overline{Q}$ can be credence in deterministic Bet 2-worlds whose nearest Bet 1-worlds are indeterministic.

Thus, CNDT can go wrong in *Hedden's Bet*. What CNDT recommends is yoked to the semantics of counterfactuals and to whether the "objectively correct" similarity relation ranks indeterministic Bet 1 worlds as most similar to deterministic Bet 2 worlds. SDT lacks this catch — it is not beholden to any semantics of counterfactuals. The "similarity relation" it uses is the agent's own general subjective selection function. More importantly for this case, SDT makes the assumption that agents' suppositions obey the Levian Constraint; hence, credence provisionally shifted from Bet 2 worlds must go to Bet 1 worlds with prior credence, and these ex hypothesi are all deterministic worlds, and hence Q-worlds. It should be said, if Hedden were to adopt a view on belief in counterfactuals along the lines sketched by Grado and Phillips-Brown (§10.2), with their Stalnakerian Belief Set Constraint, he could plug up the undesired looseness in CNDT in this case, even if he did not go the full way to embracing subjective world selection. But there are other reasons why we should not stop part-way and should indeed embrace subjective selection and hence SDT in some form, as Hedden's next two cases will illustrate.

Hedden next considers a pair of Ahmed-stye cases — versions of *Betting on the Laws* and *Betting on History*. His intent with this pairing to show that CNDT's verdicts in some deterministic cases can depend on whether a Lewis-style or a Dorr-style account of counterfactuals under determinism prevails. *Pace* Hedden, there are some mistakes in his handling of his cases, but if we modify the presentation slightly we can see his point, and this will afford us a chance to observe the theoretical benefit of SDT's subjective imaging.

Let us first consider one of the two cases as he presents it. I will modify the details in a way that I take to be superficial in order to aid analysis.

Hedden's Bet on the Laws: You are very confident in D, a deterministic set of laws. All of your credence in $\overline{D}$ is in D^*, another set of deterministic laws contrary to D. Your credence also divides over the pair of initial state propositions $\{H, H^*\}$, where HD entails you take Bet 3 and H^*D^* entails you take Bet 4. You can take Bet 3 which pays \$1 if D and \$0 if $\overline{D}$, or you can take Bet 4 which pays \$1 if D^* and \$0 if $\overline{D^*}$. You have no other options and all you care about is winning a dollar.

We have the following CNDT expected utilities:

$$\begin{aligned} \mathrm{EU}^{\mathrm{CNDT}}(\text{Bet 3}) &= c(\text{Bet 3} \boxdot\!\!\rightarrow D)\mathcal{V}(D \wedge \text{Bet 3}) + c(\text{Bet 3} \boxdot\!\!\rightarrow \overline{D})\mathcal{V}(\overline{D} \wedge \text{Bet 3}) \\ &= c(\text{Bet 3} \boxdot\!\!\rightarrow \mathrm{D}) \end{aligned} \tag{12.11}$$

$$\begin{aligned} \mathrm{EU}^{\mathrm{CNDT}}(\text{Bet 4}) &= c(\text{Bet 4} \boxdot\!\!\rightarrow D^*)\mathcal{V}(D^* \wedge \text{Bet 4}) + c(\text{Bet 4} \boxdot\!\!\rightarrow \overline{D^*})\mathcal{V}(\overline{D^*} \wedge \text{Bet 4}) \\ &= c(\text{Bet 4} \boxdot\!\!\rightarrow \mathrm{D}^*). \end{aligned} \tag{12.12}$$

Hedden claims CNDT's verdict in this case is beholden to what theory of counterfactuals under determinism is correct.[14] He says that if Lewis is right, it is permissible to take either Bet 3 or Bet 4. Here is his argument (with only the notation for propositions changed):

> Suppose that H is true. Then, if you were to take Bet 3, D would be true, winning you \$1, while if you were to take Bet 4, D^* would be true, likewise winning you \$1. Now suppose that H^* is true. Then, if you were to take Bet 3, D^* would be true, winning you \$0, while if you were to take Bet 4, D would be true, likewise winning you \$0. So regardless of how things are that are counterfactually independent of your choice — i.e. regardless of whether H or H^* is true — both bets would yield the same payoff. So it's permissible to take either.

Is this right? The aim is to prove the agent's credence in H is simultaneously all in Bet 3 $\boxdot\!\!\rightarrow D$ and all in Bet 4 $\boxdot\!\!\rightarrow D^*$ while all his credence in H^* is in Bet 3 $\boxdot\!\!\rightarrow D^*$ and in Bet 4 $\boxdot\!\!\rightarrow D$; hence

[14] In fact, his words are that it "depends on *your views* about the correct theory of counterfactuals under determinism," which certainly sounds rather like SDT, but he gives no indication that this is anything more than a loose manner of putting the claim that what CNDT says depends on what theory of counterfactuals under determinism is in fact correct and hence what one thinks CNDT says depends on what one thinks that theory is. (Emphasis added to the quote.)

both expected utilities above reduce to $c(H)$ and there's a tie. Hedden seems to tacitly assume something like the Stalnakerian Belief Set Constraint of Grano and Phillips-Brown (§10.2). But this constraint is by no means universally adopted; the more expected approach is to treat belief in a counterfactual $A \boxrightarrow B$ as belief in a proposition that holds at a world w iff the "nearest" A-worlds to w are B-worlds. Given this expected approach, it does not follow that the H-worlds in which the agent has credence — which are all Bet 3-worlds — are (Bet 4 $\boxrightarrow D^*$)-worlds on a Lewis-style account of "nearness" for counterfactuals under determinism. For all logic requires, the nearest Bet 4-worlds to these ($H \wedge$ Bet 3)-worlds might be worlds with indeterministic laws where Bet 4 occurs purely by chance.[15] They will, in any case, be worlds where Lewisian miracles[16] take place, and hence not worlds of exceptionless regularity. Whether we wish to grant that the laws at some such worlds can still be classified as deterministic takes us into hairy taxonomic metaphysics; but in any case, there is no clear reason to think laws at such worlds, whether deterministic or not, must be D^* (cf. Lewis (1981a)).[17]

What is the intuitive verdict here? The kneejerk thought is surely that given the agent's high confidence in D, they should take Bet 3.[18] What does SDT say? It can yield this verdict, depending on how the agent images in their practical suppositions. If the agent treats the laws of nature as projectible then on the supposition that they take Bet 3 the credence from a Bet 4-world that shifts

[15] The "nearest" in the Lewisian-ish sense — where the initial conditions are the same.

[16] Cf. Lewis (1979a, 1981a).

[17] In fact, on a Lewis-style account, it can happen in *Hedden's Bet on the Laws* that for either option, the agent is required by CNDT to take it. Let's see how this can happen for Bet 4. Once we do this it should be trivial for the reader to work out how it can happen for Bet 3. The agent divides all their credence between their options: Bet 3 and Bet 4. Suppose they're nearly certain they'll take Bet 4 — hence, given the hypothesis of the case, they are nearly certain of ($D \wedge$ Bet 4). Let their credence in ($D^* \wedge$ Bet 4) be $a > 0$ (we know it is non-zero since ex hypothesi the agent has some credence in D^*). Let their credence in ($D \wedge$ Bet 3) be b where $a >> b > 0$. [Hence, we assume $c(\text{Bet } 3)/c(\text{Bet } 4) << c(D^*)/c(D)$.] Is their credence in ($D \wedge$ Bet 4)-worlds — which per our assumptions is where the overwhelming bulk of their credence lies — credence in (Bet 3 $\boxrightarrow D$)-worlds? No. For credence in these worlds is credence in the disjunction of initial states P such that DP entails Bet 4. On a Lewis-style account, initial states are preserved counterfactually, which means credence in worlds with these initial states must be credence in Bet 3 $\boxrightarrow \overline{D}$. Is credence in ($D^* \wedge$ Bet 4)-worlds credence in Bet 3 $\boxrightarrow D$? This does not seem logically forced, for we have not assumed that the initial conditions that given D^* produce Bet 4 are such as produce Bet 3 given D. [If H^* and H are contraries then it follows that on a Lewis-style account, credence in ($D^* \wedge$ Bet 4)-worlds is credence in Bet 3 $\boxrightarrow \overline{D}$, since the agent's credence in Bet 4 is in H^*-worlds. But Hedden makes it less than obvious whether H and H^* are meant to be contrary.] Hence we are free to assume this credence is also credence in Bet 3 $\boxrightarrow \overline{D}$. Given our assumptions, CNDT entails Bet 4 is obligatory.

[18] This is what true blue EDTers will say.

to the supposed proposition will stay within whichever of D or D^* it is in prior to supposition — and the same thing will hold for the supposition that they take Bet 4.[19] This will make it the case that according to SDT the expected utility of Bet 3 is near one and that of Bet 4 is near zero.[20] Bet 3 will be required for such an agent, who we might say images in the Dorr-style. On the other hand, what if our agent images in the Lewis-style? Applying what we worked out in note 17, we can see how SDT might imply they are rationally obligated to take Bet 4.[21] Finally, let me mention another way SDT can deliver the result that Bet 3 is obligatory.[22] If the agent treats neither the laws nor the history as projectible features, when they suppose Bet 3, their credence should distribute evenly over D and D^*, maintaining the prior ratio between the conjunctions of those propositions with Bet 3 and hence in effect yielding credence conditional on Bet 3.[23] Of course, if this is how the agent images, Bet 3 will have expected utility close to one according to SDT and Bet 4 will have expected utility close to zero.[24] But okay, let's move on. We have not canvassed all the possible verdicts SDT can issue in this case, but what's been said should give a sense of the possibilities and their attendant rationales.

There is a case similar to *Hedden's Bet on the Laws* that exhibits the features Hedden meant to highlight and discuss. Here it is:

Certain Sam Bets on the Laws: Sam can take Bet 5 which pays \$1 if D and \$0 if $\overline{D}$, or he can take Bet 6 which pays \$1 if $\overline{D}$ and \$0 if D. Sam has no other options and all he cares about is winning a dollar. Sam is certain of the deterministic set of laws D. He is also certain of H,

[19] That is, if the agent treats the laws as projectible, but not the initial conditions.

[20] Remember, they're nearly sure of D.

[21] Naturally if such an agent is sufficiently confident in Bet 3, then SDT will say they must take that bet.

[22] This will not be the only remaining meaningfully different way to yield this verdict, as we will see.

[23] This will at least be the case if the agent treats *no* features of the world as projectible.

[24] Is one or more of these agents being epistemically irrational? I have no steady intuitions about epistemology in these strange deterministic cases. But as I've stressed in previous sections, I think we must separate questions of epistemic rationality from questions of practical rationality. I will at least say that none of these verdicts seems unacceptable to me. Each reflects a conceivable agent's attitudes as to what will/would happen if they take one bet as opposed to the other.

where H disjoins all initial-condition propositions P such that PD entails his taking Bet 5. Hence, Sam is certain he will take Bet 5.

In this case, CNDT will entail that either option is permissible if a Lewis-style account of counterfactuals under determinism is true. Obviously, the CNDT expected utility of Bet 5 for Sam is one. The CNDT expected utility of Bet 6 is also one, since all of Sam's credence is in H, which is counterfactually preserved on a Lewis-style account; hence, on a Lewis-style account, all the worlds in which he places credence are Bet 6 $\Box\!\!\rightarrow \overline{D}$ worlds. On a Dorr-style account, it is D that is preserved counterfactually; hence on this account, all his credence is in Bet 6 $\Box\!\!\rightarrow D$ worlds. Thus, on a Dorr-style account, the CNDT expected utility of Bet 6 is zero, and Bet 5 is rationally required. So, in *Certain Sam Bets on the Laws*, what CNDT recommends will be dictated by what theory of counterfactuals under uncertainty is correct.

Hedden argues that the dependence of CNDT's verdicts on the semantics of counterfactuals is unproblematic.[25] He offers two defenses that are instructive to set up alongside our comparison between CNDT and SDT. I will argue the intuitions he marshals in support of CNDT are in fact inimical to that theory, and are at home with SDT.

Here is Hedden's first defense:

> . . . if you really are certain that the laws would vary depending on how you act, then it does seem permissible to bet against the laws you think are most likely to actually obtain. For you are certain that for each proposition about the laws, it would be true if you were to bet on it, and so you would win the same amount of money either way.

I agree to the letter with this intuitive sentiment, but I do not think CNDT is the best theory to uphold it. If a Dorr-style account of counterfactuals is correct, then we are all of us (insofar as we are probabilistically coherent) certain that the laws would not vary depending on how we act, since that is a logical truth.[26] If a Lewis-style account is correct, then we are all certain the laws would vary, by the same token. Insofar as we reject this, and maintain instead that different people can and do have different beliefs about whether the laws would vary depending on their actions, and

[25] He is thinking not of *Certain Sam* but rather of *Hedden's Bet on the Laws* and an analogous case involving initial conditions. The details of that case aren't important to the points I wish to make, as it is a straightforward structural analog of *Hedden's Bet on the Laws*.

[26] Assuming a default context. See note 28 below.

that such interpersonal variance is not a symptom of conversational context per se but an inherent aspect of the psychology of conditional belief (§10.2), we need an account that treats the selection of antecedent possibilities as a subjective phenomenon. We need a theory like SDT. I will unfold these claims in more detail below.

Here is Hedden's second defense:

> Second, and more tenuously, perhaps you ought to be uncertain about whether it is the past, or instead the laws, that would be different depending on how you were to act. If so, then counterfactual decision theory tells you to bet on the laws and the initial state that you think most likely obtain .... After all, you're plausibly certain that whether it is the laws or the initial state that would vary is itself counterfactually independent of your choice. And for each pair of bets, we have a situation where conditional on it being the one thing (the laws, or the initial state) that would vary, they're equally good, while conditional on it being the other thing that would vary, the first is better than the second.

Here I more or less agree[27] to the letter with the intuitions Hedden is voicing, but it is hard to see how to model the kind of uncertainty he's talking about in terms of credences in counterfactuals, since, again, it is not a contingent matter of fact on standard accounts whether the laws or the past vary counterfactually. If a Lewis-style account is right, then *the correct* similarity relation ranks worlds with the same initial conditions closer than those with the same laws. This similarity is supposed to be a feature of logical space itself. The opposite holds if a Dorr-style account is right. Let me be more painstaking. For any function ρ from worlds to propositions and any proposition P let the proposition $P \boxrightarrow !\rho$ be true at a world w iff $\rho(w)$ obtains at the nearest P-worlds to w. Let ℓ be a function from worlds to propositions such that $\ell(w)$ is the conjunction of all the laws of nature at w. Let h be the same sort of function but for initial conditions. As modeled within the standard framework, then, uncertainty about whether the laws would vary depending on which option in $\mathcal{O}$ is done is uncertainty about whether the conjunctive proposition $\bigwedge_{O \in \mathcal{O}} O \boxrightarrow !\ell$ obtains. Similar uncertainty about initial conditions is modeled on the standard framework as uncertainty about whether $\bigwedge_{O \in \mathcal{O}} O \boxrightarrow !h$ obtains. Now, uncertainty about whether a proposition obtains is modeled via a credence function assigning some credence to worlds where the proposition holds and to others where it does not. But if a Lewis-style account of counterfactuals holds, then for

[27] I see no reason why the intuition has to be couched in terms of counterfactuals as opposed to just future-directed practical conditionals (cf. §11.2), but otherwise I agree.

any partition $\mathcal{O}$, the proposition $\bigwedge_{O \in \mathcal{O}} O \boxrightarrow !h$ holds at every deterministic possible world, since at any such w and $O \in \mathcal{O}$, the nearest O-worlds to w are worlds where $h(w)$ holds. That is, assuming a Lewis-style account, if $\mathscr{D}$ is the proposition that the world is deterministic and $\mathcal{O}$ is any partition, $\mathscr{D} \rightarrow (\bigwedge_{O \in \mathcal{O}} O \boxrightarrow !h)$ is a logical truth; hence assuming probabilistic credences, we have $c(\bigwedge_{O \in \mathcal{O}} O \boxrightarrow !h \mid \mathscr{D}) = 1$. The same is true, mutatis mutandis, if a Dorr-style account holds. So, on the standard approach to uncertainty in counterfactuals, we cannot make sense of the suggestion that an agent in one of these deterministic cases is unsure whether the laws or the past would vary with their options. If a Lewis-style account is correct the agent is *certain* — as a matter of pure logic — that the past would not vary; if a Dorr-style account is correct, they are *certain* — again, by pure logic — the *laws* would not vary.[28]

Stepping back, what we see here is a creaky framework showing us its limits. There is a better way to model the kind of uncertainty Hedden has in mind. Viewed via probabilistic subjective selection (§11.1), it is simply uncertainty about what is projectible — uncertainty about whether it's the laws or the past that's projectible. If I am unsure of this, then for any deterministic world $w \in \mathscr{D}$ and any one of my options O, we have both $0 < \sigma_w(\ell(w) \parallel O) < 1$ and $0 < \sigma_w(h(w) \parallel O) < 1$, where σ is my current probablistic subjective selection function. In other words, if I am uncertain in this way, then for any such deterministic world w, if I image on one of my options O then whatever credence I assign to w will go to some O-worlds that share their initial conditions with w and to other O-worlds that share their laws with w.[29]

With this in hand, does Hedden's intuitive reasoning — with which I am in sympathy — work? Specifically, does it work in *Hedden's Bet on the Laws*, where it is meant to apply? Does perfect

---

[28] Assuming a default context, of course. The role of context in the standard approach to counterfactuals to my mind only adds a further layer of irrelevant strangeness to the representation in the standard framework of the kind of uncertainty Hedden has in mind, which should intuitively be simple.

[29] Of course, is nothing special about the role of options here. This sort of thing can happen when imaging on any sort of proposition. I would point out that a notable independent takeaway here is that within a subjective selection framework we can model uncertainty about which of a class of features are projectible even in cases where it may seem to us somehow necessary that one and only one feature in the class is *objectively* projectible. What might it mean to be objectively projectible? This takes us beyond the scope of the present discussion, but intuitively, the appropriate idea seems to be that some particular selection function corresponds to answers to questions like *what would/will be the case if I do such-and-such*, as opposed to answers to questions like *what do I think would/will be the case if I do such-and-such*.

uncertainty about whether the laws or the past are projectible imply that it is uniquely rational to bet on the laws one is most confident in — Bet 3 if it's D, Bet 4 if it's D^*?

Let's apply SDT again, supposing this time the agent has this kind of uncertainty — in particular, suppose that for any $w \in \text{Bet } 3 \wedge (D \vee D^*)$, we have $\sigma_w(\ell(w) \parallel \text{Bet } 4) = 0.5$ and for any $w \in \text{Bet } 4 \wedge (D \vee D^*)$, we have $\sigma_w(\ell(w) \parallel \text{Bet } 3) = 0.5$.[30] As a first pass, suppose the agent is all but certain of D. It follows[31] that

$$\begin{aligned} \text{EU}^{\text{SDT}}(\text{Bet } 3) &= \mathcal{I}(D \parallel \text{Bet } 3)\mathcal{V}(D \parallel \text{Bet } 3) + \mathcal{I}(\overline{D} \parallel \text{Bet } 3)\mathcal{V}(\overline{D} \parallel \text{Bet } 3) \\ &= \mathcal{I}(D \parallel \text{Bet } 3) \\ &= c(D \wedge \text{Bet } 3) + 0.5c(\text{Bet } 4)^{\dagger} \\ &\approx c(\text{Bet } 3) + 0.5c(\text{Bet } 4)^{\ddagger} \end{aligned} \tag{12.13}$$

$$\begin{aligned} \text{EU}^{\text{SDT}}(\text{Bet } 4) &= \mathcal{I}(D^* \parallel \text{Bet } 4)\mathcal{V}(D^* \parallel \text{Bet } 4) + \mathcal{I}(\overline{D^*} \parallel \text{Bet } 4)\mathcal{V}(\overline{D^*} \parallel \text{Bet } 4) \\ &= \mathcal{I}(D^* \parallel \text{Bet } 4) \\ &= c(D^* \wedge \text{Bet } 4) + 0.5c(\text{Bet } 3)^{\dagger} \\ &\approx 0.5c(\text{Bet } 3)^{\ddagger} \end{aligned} \tag{12.14}$$

Indeed, given our assumptions Bet 3 is uniquely rational per SDT. Naturally the reverse holds if the agent is all but certain of D^*.

More generally, if the agent is merely more confident in D than in D^* does it follow that only Bet 3 is rational? Note that because $DD^* = \bot$ and $\text{Bet } 3 \wedge \text{Bet } 4 = \bot$ and $c(D \vee D^* \vee \text{Bet } 3 \vee \text{Bet } 4) = 1$,

$$\text{EU}^{\text{SDT}}(\text{Bet } 4) = 1 - c(D) - c(D^* \wedge \text{Bet } 3) + 0.5c(\text{Bet } 3). \tag{12.15}$$

Hence Bet 3[32] is obligatory according to SDT iff

[30] Superscripts suppressed.

[31] The steps marked with † depend on our Levian Constraint together with the assumptions just made about the agent's probabilistic subjective selection function. Those marked with ‡ depend on the agent's near certainty in D.

[32] Recall that $\text{EU}^{\text{SDT}}(\text{Bet } 3) = c(D \wedge \text{Bet } 3) + 0.5c(\text{Bet } 4)$.

$$c(D) > 1 - c(\text{Bet 3}) + 0.5[c(\text{Bet 3}) - c(\text{Bet 4})] = 1 - 0.5c(\text{Bet 3}) - 0.5 + 0.5c(\text{Bet 3}). \qquad (12.16)$$

That is, iff $c(D) > 0.5$, which lines up perfectly with Hedden's intuition.[33]

Returning to *Certain Sam Bets on the Laws* for a moment, I would like to say that in that case, given the inapplicability of Hedden's defenses, the inadequacy of CNDT and of CDT and EDT seem to me rather plain. EDT cannot be applied, since the Sam is certain of one of his options.[34] CDT and CNDT insist that what is rational for Sam depends on some objectively special interpretation of counterfactual conditionals.[35] But my intuition in this case aligns with Hedden's remarks quoted above (as his first defense of his theory). What it is rational for Certain Sam to do depends on what he thinks would be the case[36] if he were to take Bet 6, and intuitively, he could think the laws would be the same, or that the past would be the same. Neither of these is forced on him by logic itself, as CNDT would have it, nor is either dictated by practical rationality itself, as CDT would have it.[37] We want to know what Sam himself thinks would happen, whatever that might be. Once we know that we can determine whether both options are rationally permitted or just the one he is certain he will perform (i.e., Bet 5). As we've seen, SDT seems thus uniquely equipped to honor this intuitive assessment.[38]

12.5 Decision Instability Is Epistemology

Decision instability cases are thought by some to present a problem for CDT. These are cases, roughly speaking, where one's options are evidence against their own practical rationality. Take for instance a well-known case by Egan (2007):

---

[33] It is no small virtue that SDT allows us to model Hedden's intuition via probabilistic subjective selection, it seems to me, given how natural it is to suppose a person in an Ahmed-style decision situation might be unsure if the laws or the past would be the same or different.

[34] At least on the standard approach to conditional credence (§1.5). Cf. note 24 from §1.13.

[35] Standardly, when formulated in accordance with Eq. 12.8, CDT makes use of the so-called non-backtracking interpretation of counterfactuals — i.e., the Lewis-style interpretation.

[36] Here counterfactual talk is uniquely appropriate, since Sam can only contemplate Bet 6 counterfactually.

[37] Is one of them epistemically irrational? Separate matter altogether! See note 24, etc.

[38] I will forgo a repetitious presentation of the details.

Murder Lesion: Mary can shoot her rival Alfred or not. She is a top sharpshooter with a reliable and deadly rifle. Mary's utility is highest at worlds where Alfred is murdered, it is lowest at worlds where she shoots and misses, and it is in the middle at worlds where she does not shoot. Mary is nearly certain that most murder attempts are the result of brain lesions that also cause bad aim at the critical moment — thus she thinks most murder attempts are failures due to bad aim.

The more confident Mary is that she will try to shoot Alfred, the more confident she is that she has a lesion, and hence the more practically irrational it looks to her to follow through on the murderous plan. Many have the intuition in this case that shooting is always irrational, regardless of how likely Mary thinks it is that she will shoot. EDTers seize on this intuition, since EDT entails that shooting is irrational, being nearly certain evidence of the worst outcome, while not shooting is nearly certain evidence of the middling outcome. CDT seems to entail that what is rational for Mary to do varies with her confidence in what she will do. The more confident she is that she will do nothing, the more it looks like she should shoot (since by her lights she's probably lesion-free); the more confident she is that she will shoot, the more it looks like she shouldn't (since she's probably got herself a lesion). For CDTers, the case is thus one of "decision instability." What will SDT say about cases like this?

SDT says the verdict depends on what Mary supposes will happen if she shoots, and what this is and whether it varies with her confidence in what she will do are subjective features of Mary. It could be that no matter how confident she is in what she will do, Mary will think if she shoots she will miss, because she will have a lesion. If this is how Mary is, then she is imaging (and is disposed to image) in a way that mimics conditionalization (§11.2), and hence SDT will side with EDT and say she should not shoot. If Mary is disposed to take the physical state of her brain as a projectible feature of reality, then the more confident she is that she will shoot, the more she will suppose that she will miss if she does. In this case Mary is imaging in a way that mimics the machinery of CDT, and she will be caught in decision instability.

Stepping back, we observe that according to SDT, decision instability is not a characteristic of expected utility maximization, but of the psychology of certain decision-makers. Whether Mary

thinks like an EDTer or like a CDTer, SDT puts all the focus on what she supposes will happen if she shoots, and does not look with prejudice on what she might thus suppose; this seems exactly right at the level of subjective practical rationality. It also seems, indirectly, to rightly situate debate over such cases in the realm of epistemology. Should Mary suppose that she will miss, regardless of how confident she currently is that she has the lesion? Should her practical suppositions instead treat her brain state as projectible? These are interesting questions for the epistemology of supposition, to be approached, in my view, via the general framework of probabilistic imaging. They are not questions for expected utility theory, which should just say, roughly: shoot if you think you'll hit, don't if you think you won't.

BIBLIOGRAPHY

Adams, E. (1975). *The Logic of Conditionals*. Dordrecht: D. Reidel.

Ahmed, A. (2013). Causal decision theory: A counterexample. *The Philosophical Review*, 122(2).

Ahmed, A. (2014). *Evidence, Decision and Causality*. Cambridge University Press.

Anscombe, G. E. M. (1963). *Intention*. Harvard University Press, second edition edition.

Arpaly, N. and Schroeder, T. (2014). *In Praise of Desire*. Oxford University Press.

Audi, R. (2001). *The Architecture of Reason: The Structure and Substance of Rationality*. Oxford University Press.

Bacon, A. (2015). Stalnaker's thesis in context. *Review of Symbolic Logic*, viii(1):131–63.

Bennett, J. (1984). Counterfactuals and temporal direction. *Philosophical Review*, 93(1):57–91.

Bennett, J. (2003). *A Philosophical Guide to Conditionals*. Oxford University Press.

Braddon-Mitchell, D. and Nola, R. (2009). Introducing the canberra plan. In Braddon-Mitchell, D. and Nola, R., editors, *Conceptual Analysis and Philosophical Naturalism*, chapter 1, pages 1–20. MIT Press.

Brandt, R. B. (1979). *A Theory of the Good and the Right*. Clarendon Press, Oxford.

Buchak, L. (2017). *Risk and Rationality*. Oxford University Press.

Chan, D. K. (2004). Are there extrinsic desires? *Noûs*, 38(2):326–350.

Dancy, J. (2012). Response to mark schroeder's slaves of the passions. *Philosophical Studies*, 157:455–462.

Davidson, D. (1963). Actions, reasons, and causes. *The Journal of Philosophy*, 60(23):685–700.

DeRose, K. (2010). The conditionals of deliberation. *Mind*, 119(473):1–42.

Dorr, C. (2016). Against counterfactual miracles. *The Philosophical Review*, 125(2).

Easwaran, K. (2011). Bayesianism i: Introduction and arguments in favor. *Philosophy Compass*, 6(5):312–320.

Egan, A. (2007). Some counterexamples to causal decision theory. *The Philosophical Review*, 116(1):93–114.

Elga, A. (2022). Confessions of a causal decision theorist. *Analysis*, 82(2):203–213.

Eva, B., Fitelson, B., and Shear, T. (2022). Four approaches to supposition. *Ergo an Open Access Journal of Philosophy*, 8(26).

Finlay, S. (2007). Responding to normativity. In *Oxford Studies in Metaethics*, volume 2, chapter 10, pages 220–239. Oxford University Press.

Finlay, S. (2008). Motivation to the means. In Chan, D. K., editor, *Moral Psychology Today: Essays on Values, Rational Choice, and the Will*, Philosophical Studies Series, vol. 110, pages 173–192. Springer.

Fusco, M. (2025). Imaging and the diachronic dutch book. *Philosophy and Phenomenological Research*, pages 1–33.

Gallow, J. D. (2024). Counterfactual decision theory is causal decision theory. *Pacific Philosophical Quarterly*, 105(1):115–156.

Gärdenfors, P. (1982). Imaging and conditionalization. *The Journal of Philosophy*, 79(12):747–760.

Gärdenfors, P. (1988). *Knowledge in Flux*. MIT Press.

Gerken, M. (2011). Warrant and action. *Synthese*, 178:529–547.

Gibbard, A. (1981). Two recent theories of conditionals. In Harper et al. (1981), pages 211–248.

Gibbard, A. and Harper, W. L. (1978). Counterfactuals and two kinds of expected utility. In Hooker, C. A., Leach, J. J., and McClennon, E. F., editors, *Foundations and Applications of Decision Theory*, pages 125–162. Reidel. An abridged version appears in *Paradoxes of Rationality and Cooperation*, ed. R. Campbell and L. Sowdon (The University of British Columbia Press, 1985), 133–158. Citations refer to the abridged version.

Goodman, J. (2015). Knowledge, counterfactuals, and determinism. *Philosophical Studies*, 172(9):2275–2278.

Grano, T. and Phillips-Brown, M. (2022). (counter)factual want ascriptions and conditional belief. *The Journal of Philosophy*, CXIX(12):641–672.

Hájek, A. (2003). What conditional probability could not be. *Synthese*, 137(3):273–323.

Hammond, P. (1976). Changing tastes and coherent dynamic choice. *Review of Economic Studies*, 43:159–173.

Harman, G. (2000). *Explaining Value and Other Essays in Moral Philosophy*. Clarendon Press, Oxford.

Harper, W. L., Stalnaker, R., and Pearce, G., editors (1981). *Ifs: Conditionals, Belief, Decision, Chance, and Time*. D. Reidel, Dordrecht.

Hawthorne, J. and Stanley, J. (2008). Knowledge and action. *The Journal of Philosophy*, 105(10):571–590.

Hedden, B. (2023). Counterfactual decision theory. *Mind*, 132(527):730–761.

Hey, J. D. and Lotito, G. (2009). Naive, resolute or sophisticated? a study of dynamic decision making. *Journal of Risk and Uncertainty*, 38(1):1–25.

Hintikka, J. (1969). Semantics for propositional attitudes. In *Models for Modalities*, pages 87–111. D. Reidel Publishing Company.

Horgan, T. (1981). Counterfactuals and newcomb's problem. *The Journal of Philosophy*, 78(6):331–356. Reprinted in *Paradoxes of Rationality and Cooperation*, ed. R. Campbell and L. Sowdon (The University of British Columbia Press, 1985), 159–182.

Hume, D. (1739). *A Treatise of Human Nature*. Clarendon Press, 1978, Oxford, second edition.

Jackson, F. (1977). A causal theory of counterfactuals. *Australasian Journal of Philosophy*, 55(1):3–21.

Jeffrey, R. C. (1965). *The Logic of Decision*. McGraw-Hill, New York. Second edition published by The University of Chicago Press, 1983.

Joyce, J. M. (1999). *The Foundations of Causal Decision Theory*. Cambridge University Press.

Joyce, J. M. (2000). Why we still need the logic of decision. *Philosophy of Science*, 67:S1–S13. Supplement: Proceedings of the 1998 Biennial Meetings of the Philosophy of Science Association. Part II: Symposia Papers.

Joyce, R. (2001). *The Myth of Morality*. Cambridge University Press.

Karttunen, L. (1977). Syntax and semantics of questions. *Linguistics and Philosophy*, 1(1):3–44.

Korzukhin, T. (2014). Dominance conditionals and the newcomb problem. *Philosopher's Imprint*, 14(9). http://hdl.handle.net/2027/spo.3521354.0014.009.

Kripke, S. (1980). *Naming and Necessity*. Harvard University Press.

Lewis, D. (1973). *Counterfactuals*. Harvard University Press.

Lewis, D. (1979a). Counterfactual dependence and time's arrow. *Noûs*, 13:455–76.

Lewis, D. (1979b). Scorekeeping in a language game. *Journal of Philosophical Logic*, 8:339–59.

Lewis, D. (1981a). Are we free to break the laws? *Theoria*, 47:113–121.

Lewis, D. (1981b). Causal decision theory. *Australasian Journal of Philosophy*, 59(1):5–30.

Lewis, D. (1981c). Probabilities of conditionals and conditional probabilities. In Harper et al. (1981), pages 129–147.

Lewis, D. (1981d). Why ain'cha rich? *Noûs*, 15(3):377–380.

Lewis, D. (1986). *On the Plurality of Worlds*. Blackwell, Oxford.

Littlejohn, C. (2009). Must we act only on what we know? *The Journal of Philosophy*, 106(8):463–473.

Loewer, B. (2007). Counterfactuals and the second law. In *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*, pages 293–326. Oxford University Press.

Lord, E. (2010). Having reasons and the factoring account. *Philosophical Studies*, 149:283–296.

Lycan, W. G. (2019). *The Philosophy of Language*. Routledge, New York, third edition.

Mackie, J. L. (1977). *Ethics: Inventing Right and Wrong*. Penguin, London.

Marks, J. (1986). The difference between motivation and desire. In Marks, J., editor, *The Ways of Desire*, pages 133–147. Routledge.

McDaniel, K. and Bradley, B. (2008). Desires. *Mind*, 117(466):267–302.

Meacham, C. J. G. (2005). Three proposals regarding a theory of chance. *Philosophical Perspectives*, 19:281–307.

Meacham, C. J. G. (2014). Autonomous chances and the conflicts problem. In Wilson, A., editor, *Chance and Temporal Asymmetry*, pages 45–67. Oxford University Press.

Meacham, C. J. G. (2015). Understanding conditionalization. *Canadian Journal of Philosophy*, 45(5-6):767–797.

Meacham, C. J. G. and Weisberg, J. (2011). Representation theorems and the foundations of decision theory. *Australasian Journal of Philosophy*, 89(4):641–663.

Merricks, T. (2015). *Propositions*. Oxford University Press, Oxford.

Nozick, R. (1969). Newcomb's problem and two principles of choice. In Rescher, N., editor, *Essays in Honor of Carl G. Hempel*, pages 114–146. Springer, Dordrecht.

Parfit, D. (2001). Rationality and reasons. In Egonsson, D., Josefsson, J., Petersson, B., and Rønnow-Rasmussen, T., editors, *Exploring Practical Philosophy: From Action to Values*, chapter 2, pages 17–40. Routledge.

Parfit, D. (2011). *On What Matters, Volume One*. The Berkeley Tanner Lectures. Oxford University Press.

Rabinowicz, W. (2009). Letters from long ago: On causal decision theory and centered chances. In Johansson, L. G., editor, *Logic, Ethics, and All That Jazz - Essays in Honour of Jordan Howard Sobel*, volume 56 of *Uppsala Philosophical Studies*, pages 247–273.

Rawls, J. (1971). *A Theory of Justice*. The Belknap Press of Harvard University Press.

Rubenstein, E. (2025). The necessary uniformity of physical probability. *Philosophy and Phenomenological Research*.

Savage, L. (1972). *The Foundations of Statistics*. Dover, second revised edition. The original edition was published by John Wiley & Sons in 1954.

Savulescu, J. (1994). Rational desires and the limitation of life-sustaining treatment. *Biolethics*, 8(3):191–222.

Scanlon, T. M. (1998). *What We Owe to Each Other*. The Belknap Press of Harvard University Press.

Schroeder, M. (2007). *Slaves of the Passions*. Oxford University Press.

Schroeder, M. (2008). Having reasons. *Philosophical Studies*, 139:57–71.

Schroeder, M. (2018). Getting perspective on objective reasons. *Ethics*, 128:289–319.

Sinhababu, N. (2009). The humean theory of motivation reformulated and defended. *The Philosophical Review*, 118(4):465–500.

Skyrms, B. (1980). *Causal Necessity*. Yale University Press.

Skyrms, B. (1984). *Pragmatics and Empiricism*. Yale University Press.

Skyrms, B. (1990). *The Dynamics of Rational Deliberation*. Harvard University Press.

Smith, M. (1994). *The Moral Problem*. Blackwell, Malden, MA.

Smith, M. (1997). In defense of "the moral problem": A reply to brink, copp, and sayre-mccord. *Ethics*, 108(1):84–119.

Smith, M. (2004). Instrumental desires and instrumental rationality - i. *Proceedings of the Aristotelian Society, Supplementary Volumes*, 78:93–109.

Sobel, J. H. (1978). Probability, chance, and choice: A theory of rational agency. Unpublished. Presented in part at a workshop on Pragmatics and Conditionals at the University of Western Ontario in May 1978.

Sobel, J. H. (1986). Notes on decision theory: Old wine in new bottles. *Australasian Journal of Philosophy*, 64(4):407–437.

Stalnaker, R. (1975). Indicative conditionals. *Philosophia*, 5:269–86. Reprinted in Stalnaker (1999).

Stalnaker, R. (1978). Assertion. In Cole, P. and Morgan, J., editors, *Syntax and Semantics, Volume 9: Pragmatics*, pages 315–32. Academic Press, New York. Reprinted in Stalnaker (1999).

Stalnaker, R. (1981). Letter to david lewis. In Harper, W. L., Stalnaker, R., and Pearce, G., editors, *Ifs: Conditionals, Belief, Decision, Chance, and Time*, pages 151–152. D. Reidel Publishing Company.

Stalnaker, R. (1984a). *Inquiry*. MIT Press.

Stalnaker, R. (1984b). Possible worlds. In *Inquiry*, chapter 3, pages 43–58. MIT Press. Reprinted in Stalnaker (2003).

Stalnaker, R. (1987). Semantics for belief. *Philosophical Topics*, 15(1):177–90. Reprinted in Stalnaker (1999).

Stalnaker, R. (1999). *Context and Content: Essays on Intentionality in Speech and Thought*. Oxford University Press, Oxford.

Stalnaker, R. (2003). *Ways a World Might Be*. Oxford University Press, Oxford.

Stalnaker, R. (2012). *Mere Possibilities*. Princeton University Press.

Stanley, J. and Williamson, T. (2001). Knowing How. *The Journal of Philosophy*, 98(8):411–444.

Sylvan, K. (2015). What apparent reasons appear to be. *Philosophical Studies*, 172:587–606.

Sylvan, K. (2020). Respect and the reality of apparent reasons. *Philosophical Studies*.

Thomson, J. J. (1986). Imposing risks. In Parent, W., editor, *Rights, Restitution, & Risk*. Harvard University Press.

Velleman, J. D. (1989). *Practical Reflection*. Princeton University Press.

Verdejo, V. M. (2020). Norms for pure desire. *Theoria*, 35(1):95–112.

Vogelstein, E. (2012). Subjective reasons. *Ethical Theory and Moral Practice*, 15:239–257.

Wedgwood, R. (2011). Instrumental rationality. In Shafer-Landau, R., editor, *Oxford Studies in Metaethics*, volume 6, chapter 12, pages 280–309. Oxford University Press.

Wedgwood, R. (2013). Gandalf's solution to the newcomb problem. *Synthese*, 190:2643–2675.

Weirich, P. (2001). *Decision Space: Multidimensional Utility Analysis*. Cambridge University Press.

Weirich, P. (2015a). Intrinsic utility's compositionality. *Journal of the American Philosophical Association*, pages 545–563.

Weirich, P. (2015b). *Models of Decision-Making: Simplifying Choices*. Cambridge University Press.

Whiting, D. (2014). Keep things in perspective: Reasons, rationality, and the a priori. *Journal of Ethics and Social Philosophy*, 8(1).

Williams, B. (1981). Internal and External Reasons. In *Moral Luck*. Cambridge University Press.

Wodak, D. (2017). Can objectivists account for subjective reasons? *Journal of Ethics and Social Philosophy*, 12(3).

Wodak, D. (2019). An objectivist's guide to subjective reasons. *Res Philosophica*, 96(2):229–244.