Desire, Belief, and Conditional Belief

by

David Jeffrey Etlin

A.B., University of Chicago (1998)

Submitted to the Department of Linguistics and Philosophy
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2008

Author .....
Department of Linguistics and Philosophy
August 29, 2008

Certified by .....
Robert Stalnaker
Laurance S. Rockefeller Professor
Thesis Supervisor

Accepted by .....
Alex Byrne
Chairman, Department Committee on Graduate Studies

Desire, Belief, and Conditional Belief

by

David Jeffrey Etlin

Submitted to the Department of Linguistics and Philosophy
on August 29, 2008, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Philosophy

Abstract

This dissertation studies the logics of value and conditionals, and the question of whether they should be given cognitivist analyses. Emotivist theories treat value judgments as expressions of desire, rather than beliefs about goodness. Inference ticket theories of conditionals treat them as expressions of conditional beliefs, rather than propositions. The two issues intersect in decision theory, where judgments of expected goodness are expressible by means of decision-making conditionals.

In the first chapter, I argue that decision theory cannot be given a Humean foundation by means of money pump arguments, which purport to show that the transitivity of preference and indifference is a requirement of instrumental reason. Instead, I argue that Humeans should treat the constraints of decision theory as constitutive of the nature of preferences. Additionally, I argue that transitivity of preference is a stricter requirement than transitivity of indifference.

In the second chapter, I investigate whether David Lewis has shown that decision theory is incompatible with anti-Humean theories of desire. His triviality proof against "desire as belief" seems to show that desires can be at best conditional beliefs about goodness. I argue that within causal decision theory we can articulate the cognitivist position where desires align with beliefs about goodness, articulated by the decision making conditional.

In the third chapter, I turn to conditionals in their own right, and especially iterated conditionals. I defend the position that indicative conditionals obey the import-export equivalence rather than modus ponens (except for simple conditionals), while counterfactual subjunctive conditionals do obey modus ponens. The logic of indicative conditionals is often thought to be determined by conditional beliefs via the Ramsey Test. I argue that iterated conditionals show that the conditional beliefs involved in indicative supposition diverge from the conditional beliefs involved in learning, and that half of the Ramsey Test is untenable for iterated conditionals.

Thesis Supervisor: Robert Stalnaker
Title: Laurance S. Rockefeller Professor

Acknowledgments

I am grateful to my friends and family for their support, especially Thao. I would like to thank everybody who I have learned from while at MIT, most of all my committee: Robert Stalnaker, Vann McGee and Agustín Rayo.

Contents

1 Why Obey the Rationality Postulates? 11

1.1 Introduction 11

1.2 Humean versus Neo-Humean Theories of the Passions 12

1.3 Preference 14

1.3.1 Five Degrees of Preference Involvement 14

1.3.2 Troubles With Negative Transitivity 16

1.3.3 Fuzzy and Partial Rationality 17

1.4 Choice and Choiceworthiness 18

1.4.1 From Buridan's Ass to Choice Functions 18

1.4.2 Rationalizability 20

1.4.3 Platitudes of Preference and Indifference 21

1.5 Synchronic and Diachronic Path Independence 23

1.5.1 "All Things Considered" 23

1.5.2 Avoiding Cyclical Choice 25

1.5.3 Worse Off With Cyclical Preferences? 27

1.5.4 Putting the Money Back in the Money Pump? 28

1.5.5 Is it the Choice Rule? 29

1.5.6 Money Pumps Without Cyclical Preferences 31

1.5.7 Swapping Requirements versus Status-Quo Maintenance . . . 32

1.5.8 Swapping Permission and Life-Planning 34

1.5.9 Moving Beyond Path Independence 35

1.6 Constitutive Constraints and Fragmented Minds 36

1.6.1 Where Reasons Come to an End 36
1.6.2 Rational Choice Without Preferences? 38
1.6.3 Comparison With Deductive Rationality 40
1.6.4 Loading up the Consequences and Reflective Endorsement . . 42
1.7 Conclusion 45

2 Cognitivism and Two Kinds of Desire **47**
2.1 Introduction 47
2.2 Beliefs About Goodness and Their Alignment With Desire 50
2.2.1 Cognitivism 50
2.2.2 Externalism, Internalism, and Humeanism 51
2.2.3 Absolutism versus Relativism 53
2.3 The Setup and Proof of the Triviality Result 55
2.3.1 Background on the Formal Framework 56
2.3.2 Desire as Conditional Belief 58
2.3.3 Desire as Belief 59
2.3.4 The Triviality of Desire As Belief 63
2.3.5 Comparison With Triviality for Conditionals 63
2.3.6 Diagnosis: Conditionals and Negation 65
2.4 Attempts to Save Cognitivism from the Triviality Theorem 69
2.4.1 Qualms About the Decision Theoretic Idealizations 70
2.4.2 Abandoning Belief for Conditional Belief 71
2.4.3 The Relativity of Value 72
2.4.4 The Relativity of Indicative Conditionals 73
2.5 Cognitivism Saved by Causal Conditionals 75
2.5.1 From Evidential to Causal Decision Theory 76
2.5.2 Causal Decision Theory With Subjunctive Conditionals 78
2.5.3 From Causal Decision Theory to Desire as Belief 81
2.5.4 A Newcomb Problem Around Every Corner 82
2.5.5 Choiceworthiness versus Desire 85

2.5.6 Rationalization and Two Kinds of Desire 86
2.5.7 Causal Utility as Motivating 89
2.6 How un-Humean is Decision Theory? 90

3 Modus Ponens Revisited **93**
3.1 Introduction 93
3.2 An Uncertain Inference 94
3.2.1 The Main Example 94
3.2.2 Some Lessons of the Example 95
3.3 Analyses of the Indicative Conditional 96
3.3.1 The Material Conditional 96
3.3.2 The Stalnaker Conditional 97
3.3.3 The Hybrid Theory 98
3.3.4 Conditional Probability and Adams' Thesis 100
3.4 The Import-Export Equivalence 101
3.4.1 The Logic of Iterated Conditionals 101
3.4.2 Modus Ponens as a Reasonable Inference 101
3.4.3 Trouble with Exportation 103
3.4.4 Importation and the Direct Argument 105
3.4.5 The Supplemented Equivalence Theory 107
3.5 Iterated Conditionals and the Ramsey Test 109
3.5.1 Triviality for Simple Ramsey Test Conditionals 109
3.5.2 Conditional-Factual Dualism and Iterated Conditionals 111
3.5.3 Import-Export Triviality Results for Revisions 113
3.5.4 Lessons of the Iterated Triviality Results 114
3.5.5 Semantics and Indexicalism 116
3.5.6 Modals 119
3.6 Subjunctive Conditionals 120
3.6.1 Extending the Analysis? 120
3.6.2 Are There Subjunctive Counterexamples to Modus Ponens? . 121

3.6.3 Counterfactual Counterexamples to Import-Export 123
3.7 Conclusion 125

Chapter 1

Why Obey the Rationality Postulates?

1.1 Introduction

Decision theory is often taken to be a precise version of a view of practical reason attributed to Hume: values are subjective, and reasoning about them is instrumental. Yet it is generally noted that decision theory goes beyond what is officially sanctioned by Hume, in that it places structural constraints on preferences where Hume did not find any. In a famous discussion of the subjectivity of the "passions", Hume says:

> 'Tis not contrary to reason to prefer the destruction of the whole world to the scratching of my finger. (*A Treatise of Human Nature* II.III.III "Of the Influencing Motives of the Will")

According to the requirement of transitivity, Hume should grant that if he also prefers the scratching of his finger to the brushing of his hair, then it is contrary to reason for him to fail to prefer to the destruction of the world to the brushing of his hair.

Hume nowhere concedes such a requirement, and explicitly denies that there are any rational constraints on passions other than what are imposed by causal beliefs in means-ends reasoning. Indeed, in laying out the principles of decision theory, the eminent statistician Leonard Savage presented them as "rationality postulates" which

he preferred to regulate his own preferences by. Sticking to his Humean scruples, Savage did little more to defend his postulates than to invite the reader to share this higher-order preference with him.

A more prevalent approach is to defend decision theory as a universal norm of practical reason, which fits Hume's overall doctrines more closely than it violates them. The canonical defense of the structural constraints on preferences—and here we abstract away from risk and uncertainty—is the money pump argument. This argument purports to show that if an agent violates the most basic constraints of decision theory, then the agent is disposed to be defeated in his actions in a way that is solely his own fault, not the world's. Such easily foretold doom would be a terrible failure of instrumental reason, so (it is claimed) decision theory's Humean credentials are secured.

I shall argue that the constraints of decision theory are constitutive principles of practical reason. In comparison with rules of logic, they form an ideal of rationality, which we often deviate from because of failures to unite our disparate mental states into a cohesive whole. Unlike the case of belief, where the norm of truth provides an external criterion, the principles of rational preference cannot be given a neo-Humean justification by the money-pump argument. Moreover, to the extent that there are requirements of rational preference, they are slightly weaker than the standard theory; while transitivity of preference is required, there is no requirement of transitivity of indifference.

1.2 Humean versus Neo-Humean Theories of the Passions

Here is a nice expression of the neo-Humean view, written by Leonard Savage in the 1950s car market:

> There is, of course, an important sense in which preferences, being entirely subjective, cannot be in error; but in a different, more subtle sense they

> can be. Let me illustrate by a simple example containing no reference to uncertainty. A man buying a car for $2,134.56 is tempted to order it with a radio installed, which will bring the total price to $2,228.41, feeling that the difference is trifling. But, when he reflects that, if he already had the car, he would certainly not spend $93.85 for a radio for it, he realizes that he has made an error. [90, p. 103]

Savage's customer violated the requirement that preferences should be *transitive.* In Savage's example, transitivity is violated since the customer prefers a car with a radio pre-installed to one without a radio, but prefers the latter to a car with a radio installed after purchase. However, he is indifferent between a car with a radio pre-installed and one with it installed after the fact.

Savage's example has persuasive intuitive force. But is the nature of the problem with the preferences in the example, and why should they be altered by somebody who has them? Here is what Savage has to say:

> Suppose someone says to me, "I am a rational person, that is to say, I seldom, if ever, make mistakes in logic. But I behave in flagrant disagreement with your postulates, because they violate my personal taste, and it seems to me more sensible to cater to my taste than to a theory arbitrarily concocted by you." I don't see how I could really controvert him, but I would be inclined to match his introspection with some of my own. I would, in particular, tell him that, when it is explicitly brought to my attention that I have shown [a trio of intransitive preferences], I feel uncomfortable in much the same way that I do when it is brought to my attention that some of my beliefs are logically contradictory. Whenever I examine such a triple of preferences on my own part, I find that it is not at all difficult to reverse one of them. In fact, I find on contemplating the alleged preferences side by side that at least one of them is not a preference any at all, at least not any more. [90, p. 21]

Savage is by no means idiosyncratic, but neither is his traditional Humean interlocutor.

Are intransitive preferences irrational, and in correcting them is an agent altering his preferences or better expressing his true preferences? A mix of all attitudes can be found in the experimental literature. Tversky, in his famous study of intransitive preferences, reported that subjects regarded themselves as having made a mistake akin to a logical mistake when they were alerted to their intransitive choices. [105, p. 455] But Raiffa reports that in his experiments, even while acknowledging the normative authority of transitivity, some subjects resisted altering their "true" preferences despite their being "illogical". [84, p. 75] In a different kind of experimental setup where subjects were not alerted to their own intransitivities, Davidson reports:

> It was found that as time went on, people became steadily more consistent; intransitivities were steadily eliminated ... Apparently, from the start, there were underlying and consistent values which were better and better realized in choice. [23, pp. 235-6]

Davidson adds, "I found it impossible to construct a formal theory that could explain this, and gave up my career as an experimental psychologist." Following Davidson's move to philosophy, let us turn from descriptive concerns to the more conceptual and normative issues. In order to better interpret the range of intuitions and experiments, let us take a closer look at the nature of preference and its relation to choice.

1.3 Preference

1.3.1 Five Degrees of Preference Involvement

Here are five properties of a binary relation P, in increasing strength.

P1. *asymmetric*: if xPy then not yPx

P2. *acyclic*: no finite preference chain $xPy...Px$

P3. asymmetric, plus *transitive*: if xPy and yPz then xPz

P4. asymmetric, plus *negatively transitive*: if not xPy and not yPz, not xPz

P5. asymmetric, plus *strongly complete*: either xPy or yPx or $x = y$.

The letter 'P' is used as a mnemonic for "x is preferred to y" (in the sense of "strictly preferred to"). But it is doubtful that all of these might be eligible properties of a genuine preference relation, which isn't to say that each hasn't had its defender. Plato would have endorsed P4, as expressing his doctrine in the *Protagoras* that all values should be commensurable. Commensurability is captured by negative transitivity, as is revealed by its contraposed form: if xPz then xPy or yPz. (Whenever preference holds between some pair of items, any third item can be "commensurated" with them by being placed on the same ordering, by standing in that same relation with respect to at least one of them.) Orthodox economic theory (such as Savage) employs follows Plato in the employing P4, underlying the famous maxim of maximizing utility.[1] Heterodox decision theorists have employed preference relations as weak as P1, while going in the other direction some have opted for the strongest P5.

One might try to fix on the correct level of rational preference by appeal to intuitions. If one accepts the intuition that Savage's car shopper is structurally irrational in having intransitive preferences—and recall that many subjects in experiments to hold themselves to such standards, even when they refuse to obey them—then preference should be at least as strong as P3.

[1]It is common to discuss the weak preference relation xRy, which can be defined in terms of the strict preference relation as: not yPx. This definition entails the completeness of R. (If xPy then not yPx by the asymmetry of P, so xRy. If not xPy then yRx. So either xRy or yRx.) The principle corresponding to P4 is then transitivity of R.

Many economists, especially those working in the tradition of choice functions (see below), take R as the primitive concept. Strict preference xPy can be defined as $xRy \wedge \neg yRx$. As noted by Rechenauer, since completeness of R is a substantive assumption on this approach, one must modify one half of the equivalence with P4, resulting in: if P is asymmetric and negatively transitive, then if R is complete then it is transitive. Rechenauer maintains that R is the better conceptual primitive since one should not assume the completeness of R. But the problems with doing so arise in connection with the negative transitivity of P, and in the next section we argue against that as a condition of rationality.

1.3.2 Troubles With Negative Transitivity

However, the move from transitivity (P3) to negative transitivity (P4) has been viewed with much suspicion. This is largely due to two kinds of counterexamples, the first turning on incommensurabilities of value having a good deal of currency in moral and political philosophy, the second stemming from issues of indiscriminability in psychology and the philosophy of mind. In the classic incommensurability example, a child prefers a bike with a bell to one without a bell. But the child has no preference between either of them versus a pony. In the classic indiscriminability example, a person prefers two lumps of sugar in his coffee to one, and prefers each to zero. But the coffee consumer has no preference between differences of a single grain of sugar, since he cannot taste the difference in pairwise comparisons.[2]

The defender of the classical P4 condition might resist the force of these intuitions. For instance, Maher argues that it is mistaken to think that the lumps of sugar differing by only a small amount really taste the same even though they are pairwise indiscriminable; the difference comes out in other comparisons. So even though you only care about how sweet your coffee is, Maher claims that your preferences for sugar can indeed obey P4, by strictly preferring larger amounts of sugar to smaller amounts. Note that this is compatible with Humeanism about value, as it is a case of satisfying your higher order preferences that your lower order preferences track sweetness in sugar.

But even if Maher's claim about the psychology of sensation is correct, so there are differences in taste that you cannot pairwise discriminate, we haven't yet established that the agent would be more rational if cared about them (much less that he would be irrational if he didn't care about them). And the transitive "correction" of the sugar preferences seems to be an artifact of the example, where the preferences track a natural continuum the agent imprecisely perceives.

Say that a set of preferences P' extends P if $xP'y$ whenever xPy. As Lehrer

[2]The first example is due to the economist W. Armstrong; for philosophical discussion see Raz and the introduction and essays in Chang. The second example is due to Luce, who proposes a formal model of thresholds of discriminability; his semi-orders lie between P3 and P4. Further formal and psychological discussion is in Ng.

and Wagner note, there are five distinct extensions of the preference structure the bike-pony example, where xPz is the only preference over x, y, z.[3] There is no "natural" way of resolving the failure of negative transitivity in the case of the child's preferences, unless we suppose that there is a real question the child needs to decide of where the pony fits into a complete ranking with the bikes without and without bells. That would be not only an anti-Humean, it would also fly in the face of even a realist theory of value which allows that there are some issues that are simply a matter of personal taste.

1.3.3 Fuzzy and Partial Rationality

An ecumenical position treats rationality as coming in degrees. This is the position taken by Suzumura, who quotes approvingly from Lionel Robbins:

> there is a sense in which the word rationality can be used which renders it legitimate to argue that at least some rationality is assumed before human behavior has an economic aspect – the sense, namely, in which it is equivalent to "purposive" ... But to say this is not to say in the least that all purposive action is completely consistent. It may indeed be urged that the more that purposive action becomes conscious of itself, the more it necessarily becomes consistent. ([88], quoted in [104, p. 18])

Another more tolerant position allows for failures to live up to the full standard of rationality without deeming them less rational.

The key concept for articulating the more tolerant position is the extension of preferences, which was introduced in the last section. The extendibility theory of rationality settles on a certain level (say, P4) as the correct ideal of rationality, and says that an agent's preferences are rational as long as they meet a specified threshold which is capable of being extended to reach that ideal. The purest version of such a theory says that the threshold just is the minimal level which is extendible. This

[3] One is $xPyPz$, but other extensions are given by $yPxPz$ and $xPzPy$, as well as by $xIyPz$ and $xPyIz$ (where I expresses that neither is preferred to the other).

places the bar at P2, since a preference cycle preferences such as $xPyPzPx$ cannot be extended to be both transitive and asymmetric. But in his defense of the extendibility interpretation of P4, Maher raises the bar to P3 as the minimal condition for rationality. This is hard to resist once you have bought into the extendibility theory, because for any preferences meeting P2, there is a unique extension of them satisfying P3 (whenever xPy and yPz, let $xP'z$ hold). As we saw in the last section, there is no unique extension to preferences satisfying P4 from those merely satisfying P3.

But even if we hold that the above properties of P express increasingly stronger degrees of rationality, or that they lie within the fuzzy region of a vague concept or rationality, we can't just stop there. We still need to get a better grasp of where the boundaries of the concept of rational preference is. If P4 is the correct ideal, why is this is so? To answer this question, let us take a closer look at the connection between preference and rational choice.

1.4 Choice and Choiceworthiness

1.4.1 From Buridan's Ass to Choice Functions

If asked what the relation between preference and choice is, an answer that slips easily off the tongue is that the chosen option is the one which is preferred to all the alternatives. However, most of us are readily persuaded that this can't be literally correct even in the case of fully rational choices. After all, there might be several alternatives which are equally good. Or speaking even more strictly, so as to bypass worries about incommensurabilities of value, there might be several alternatives such that none is preferred to the other.

The original answer which we have abandoned was defended by Liebniz. He thought that whenever something was chosen, this revealed that the mind found (however unconsciously) some feature to tip the balance in favor of the chosen option. Reid derided this view as an unfounded metaphor comparing the operations of the mind to a physical system of equilibrium. He contended that the active powers of

the mind allowed it to pick from amongst options where none was preferred to the other. Buridan's Ass, who dithered perpetually between two equal bails of hay, failed to exercise its power of choice rather than find a hidden betterness-maker.[4]

A useful tool to formalize these matter is the economic model of a "choice function" (which as we will see is not the best name, apart from sharing its name with a different concept in pure math). We shall confine ourselves to a finite universe of objects U. A function C is a choice function just in case for every non-empty set S, $C(S)$ is a nonempty subset of S. The intended interpretation is that the function represents an agent's disposition to chose from a menu of options S. $C(S)$ represents the options that agent regards as choiceworthy, although perhaps only one of them will be chosen.

Choice functions are often approached with the foundational methodology of behaviorism (of the Rylean dispositional sort), where preference is to be constructed out of choice. However, one might regard preferences as conceptually prior to choice (or sharing equal billing), and simply treat choice functions as a useful way of characterizing the behavior induced by different kinds of preference relations (or the lack thereof).[5]

In the language of choice functions, the Liebnizian approach is to always treat $C(S)$ as a singleton set. The more general (and generally accepted) approach follows in the footsteps of Reid, and treats choiceworthiness as being *undominated*: $C(S) = \{x : \neg(\exists y \in S)xPy\}$.[6] But this formula is unquantified, so let us now consider how to turn it into a statement.

---

[4]I owe these references to Morgenbesser and Ullman-Margalit [106].

[5]One might also regard as overly restrictive the interpretation of S as a genuine *choice*, especially since we will usually consider arbitrary sets from S, some of which might not make sense as genuine menus of options. We can go partway to relieving this by moving further from behaviorism, and treating $C(S)$ as that which the agent would like to learn to be obtainable out of S. This is a kind of halfway move to Jeffrey's theory of "news value", which attaches to propositions (sentences) rather than commodities. Moving all the way to Jeffrey's theory might have conceptual advantages, but here I shall employ the constructions familiar from standard economics.

[6]It is here that R has struck most economists as more elegant, since in terms of it we can say that an option is choiceworthy when it is (weakly) preferred to everything: $C(S) = \{x : \{(\forall y \in S)xRy\}$. Although this formulation is superficially similar to the Liebnizian position, they are only equivalent when $C(S)$ is always a singleton set.

1.4.2 Rationalizability

The question of the order of quantifiers for our choice principle is anticipated in a dispute Aristotle had with Plato. Aristotle claimed that in choosing the mind makes a unity of several images. However, he denied Plato's view (from the *Protagoras*) that the mind needed to conceive of a unity between all choices. In our framework, Aristotle's position seems to be that for each choice there is a preference relation determining choiceworthiness, while Plato argued that there should be a preference relation which holds for all choices.[7]

The Aristotelian condition $[\forall S \exists P] C(S) = \{x : \neg(\exists y \in S) xPy\}$ is simply too weak to be a substantive norm for rational choice. This is because *any* choice function will satisfy it, and indeed there will be a preference relation satisfying P4. Let xPy hold just in case $x \in C(S), y \notin C(S)$. Rationality should require more than that.

On the face of it, what is missing from this supposed unity within a choice is some level of unity between choices. Taking a hint from Plato, we shall reverse the order of the quantifiers. We shall take as a necessary condition of a genuine preference relation that it satisfy the condition known in the choice theory literature as "rationalizability":

$$[\exists P \forall S] C(S) = \{x : \neg(\exists y \in S) yPx\}.$$

As a normative principle, given a preference function meeting its strictures, the intended rule is: any option which is undominated is worthy of choice. As a descriptive principle, it places conditions on what dispositions can be explained in terms of maximizing or optimizing by a preference relation.

A choice function is rationalizable just in case it is "binary":

$$x \in C(S) \equiv \forall y \in S : x \in C(\{x, y\}).$$

However, not only is this not sufficient for realizing the full Platonic ideal of commensurable preferences, it doesn't even suffice for transitivity of preference. Rationalizability can be satisfied by a preference relation which is merely acyclic (P2), as

[7] See Wiggins for discussion, and arguments for the Aristotelian view.

in Savage's car example. (This isn't to say that acyclic pairwise choice is sufficient for rationalizability, since what is chosen out of larger sets might not be determined by the pairwise choices.) Something has been left out.[8]

1.4.3 Platitudes of Preference and Indifference

A quick diagnosis of the problem is that we were too hasty in abandoning Liebniz for Reid, and followed common economic practice of treating the relation between choiceworthiness and preference as simply being undominance. True, it is a necessary condition of an option being choiceworthy that it is undominated. By focusing on the cases where multiple options are choosable, it overlooks the simple insight that when an option is *uniquely* choiceworthy, that option is preferred to its rivals. This seems to be as much a platitude about preference and choice as the undominance principle.[9]

Given the binariness (and thus rationalizability) of a choice function, this platitude is captured by this condition on choice functions:

$$[C(S) = \{x\} \wedge y \in S] \supset C(\{x, y\}) = \{x\}$$

(It should not be confused with the stronger Liebnizian condition that the fact that an option was *chosen* shows this it is preferred.) Given acyclic but intransitive preferences $aPb, bPc, \neg aPc$, the rule to pick something undominated says $C(\{a, b, c\}) = (\{a\})$. This would be ruled out by the platitude under consideration, according to which such a choice shows that aPc after all.

[8]Critics from the other direction criticize rationalizability for being too strong. Schwartz points out that there are lots of ways a choice function can be related to a binary relation, and places the blame on the principle of undominated choice. He claims it is too restrictive, and gives a more general equation which agrees with it in the case where preferences are acyclic, but still is defined for cyclic preferences. (Schwartz's rule says that choiceworthy items are those from the smallest subsets of the menu such that for each subset, no thing could be removed from the set without it being preferred to something left over.) I do not accept Schwartz's rule, because it allows that some things might both be choiceworthy from a set even though each is "preferred" to the other.

[9]Davidson *et al* provide such a disjunctive criterion: "a rational choice (relative to a given set of alternatives and preferences) is one which selects the alternative which is preferred to all other alternatives; if there are several equivalent alternatives to which none is preferred, than any one of these is selected." [27, p. 145] However, even this is not adequate to the criterion of choice they intend (as is clear from the rest o their article), which is meant to also capture the intransitivity of indifference (which they are apparently presupposing).

If the move from P2 to P3 can be supported by platitudes connecting rational preference and choice, can we also support a similar move to P4? The relevant formal condition on choice functions is:

$$[C(\{x,y\}) = \{x,y\} \wedge \{x,y\} \subset S] \supset [x \in C(S) \equiv y \in C(S)].^{10}$$

Given binariness (and thus rationalizability), this gives us that if an option a is choiceworthy and another b isn't, then a is preferred to b. But unlike our earlier platitude, this principle is incorrect as a strict requirement of rationality. For we have no objection to its failing in the case of the indiscriminable sugar sensations, and in the case of incommensurable values of the bikes and the pony.

Its appeal comes from overlooking the other other choiceworthy options c in the counterexamples, whose preference to b is responsible for the latter not being choiceworthy. One might hold that just as with preference, transitivity is a requirement of indifference. Despite initial impressions, it is not necessarily true that if an agent is equally disposed to pick either of a or b out of $\{a,b\}$, then he is genuinely indifferent between them.

How can we make sense of this lopsided treatment of preference and indifference, with respect to transitivity? We have seen that transitivity of preference is a prerequisite for transitivity of indifference, in that (assuming acyclicity) the latter is captured by P4 while the former is captured by the weaker P3. Moreover, while transitivity of preference follows from a unique extension of acyclic preferences, transitivity of indifference does not. So we can treat a choice function meeting merely P2 as an erroneous expression of the agent's true preferences, which would yield a P3 choice function. But we cannot in general say that a choice function meeting only P3 is an erroneous expression of preferences which yield a P4 choice function. So if having rationalizable choices is a condition of rationality, we can see how transitivity of preferences would be a stricter requirement than transitivity of indifference.

But could it be that transitivity of indifference is no requirement of rationality

[10]This is a pairwise (and thereby weaker) version of Sen's β (called the "dual Chernoff axiom" by Suzumura). It entails the previous principle (corresponding to transitivity), which is a weaker version of Sen's δ (compare the Superset axiom in Suzumura).

at all? In the next section, we shall examine a famous argument which purports to show that transitivity of indifference is in fact a strict requirement of rationality. The argument turns out, in the end, to be fallacious unless one grants an implausible picture of what even idealized rational agency consists in.

1.5 Synchronic and Diachronic Path Independence

1.5.1 "All Things Considered"

According to Darwall, decision theorists will fail to account for the transitivity of preferences so long as they try to divorce preference from beliefs about betterness. As Darwall points out, an essential feature of preference that is brought out by the transitivity requirement is that it is an "all things considered" relation. It is not surprising that a is better than b in one respect, b better than c in another respect, while a fails to be better than c. What is surprising is that a is better than b all things considered, b better than c all things considered, and yet a fails to be better than c. This is something which would be explicable if preferences were held accountable to judgments (beliefs) about what is better than what.

Darwall argues that it is inexplicable on a Humean view, where the only constraints on preferences are those of internal consistency. Darwall focuses on the money-pump argument, discussed in the next section. However, there is a way in which the Humean can explicate the sense in which transitive preferences express an "all things considered judgment" without basing them in beliefs.

To see how, let us return to choice functions. Another property of choice functions is the condition of "path independence":

$$x \in C(S \cup T) \equiv x \in C(C(S) \cup C(T)).$$

Path independence is independent of binariness; a choice function might satisfy either without satisfying the other. For instance, the choices in Savage's car example satisfy binariness, but not path independence.[11] But taken together, binariness and

[11] An example satisfying path independence but not binariness is given by Plott, [81, p. 1081]

path independence give necessary and sufficient conditions for rationalizability by a preference relation which is transitive (obeys P3).[12]

Path-independence, I claim, makes a good explication of the notion of preference representing an "all things considered judgment". Path independence represents the idea that a choice may be subdivided into a series of smaller choice problems, without regard to how the problem is so divided. When the choice function is also binary, it expresses that the choice is represent by a (transitive) preference relation representing choiceworthiness in a way that considers all ways of arriving at the decision.

Path independence is an attractive property of a choice procedure. Unfortunately, I do not think it will satisfy the critic of decision theory or its Humean foundations, because it will not *justify* transitivity. That is, it does not show to somebody antecedently skeptical about decision theory why somebody is subject to criticism (by himself or others) for failing to have a transitive preference relation. For neither path independence nor binariness will be acceptable to somebody who thinks that values are "context dependent" or "emergent" from the options in the choice situation. However, the sorts of examples which count against path-independence also can be used against binariness, so at least they are of no use to somebody who thinks that rationality requires acyclicity of preferences but not transitivity.

An example against path independence, as well as binariness, can be found in this example of a 1950s gourmand, from Luce and Raiffa [70, p. 288]. A new customer in a restaurant is offered a chooses salmon at \$2.50 over his other option of steak at \$4.00, considering the first dish harder to spoil than the second. But when the waiter informs him that frog's legs are also available at \$4.50, he changes his order to the steak. The customer detests frog's legs, but regards them as evidence that

---

against an apparent suggestion by Arrow that path independence is sufficient for transitivity. Let $C(\{x,y\}) = \{x,y\}, C(\{y,z\}) = \{y,z\}, C(\{x,z\}) = \{x,z\}, C(\{x,y,z\}) = \{x,y\}$. Plott thought that path independence was a condition on rational choice of interest apart from representability by a binary relation.

[12]Jointly, the conditions feature some redundancy, in that the left to right directions of either (but not both) could be dropped. The shared component is the "Chernoff Condition" ("independence of irrelevant alternatives", Sen's α): if $S \subset T$ and $C(T) \cap S \neq \emptyset$ then $S \cap C(T) \subset C(S)$. This is equivalent to the left to right part of path independence, and is is equivalent to that half of binariness in the presence of its right to left direction (the "Generalized Condorcet" condition, called $\gamma 2$ by Sen).

the restaurant is high quality, and so the steak will be to his liking. These choices are intuitively acceptable, even though they violate both halves of the conditions in question.[13]

Later we will consider how to deal with the force of this kind of counterexample to even the most basic conditions of rationality. But next we turn to *the* classic defense of transitivity (and negative transitivity) of preferences. It involves another kind of path independence, about sequential choice, rather than the static situation of a one-off choice considered above.

1.5.2 Avoiding Cyclical Choice

Davidson, McKinsey and Suppes provide a nice example of cyclical preferences.

> Mr. S. is offered his choice of three jobs by a cynical department head (never mind what department): He can be a full professor with a salary of \$5,000 (alternative a), an associate professor at \$5,500 (alternative b) or an assistant professor at \$6,000 (alternative c). Mr. S. reasons as follows: aPb since the advantage in kudos outweighs the small difference in salary; bPc for the same reason; cPa since the difference in salary is now enough to outweigh the rank. [27, p. 145]

(Note that this article was published after Savage's book and before Luce and Raiffa's, so you can compare the salaries in terms of their buying power for a car or a meal in the examples above.)

Davidson *et al* attempt to illustrate the irrationality of Mr. S's preferences by means of this elaboration, the notorious money pump.[14] With Mr. S. starting off

[13]Indeed, the only consistency condition from the literature satisfied in this example is the extremely weak Stability Condition (mentioned in Suzumura): $C(S) = C(C(S))$. Here is an example violating that too, which plays on causal relations between choices. Out of a choice between two Democrats you prefer one over the other; but when a third Republican candidate is added as an option, you would choose either of the Democrats instead (since supporting one over the other weakens both relative to the Republican).

[14]They attribute this to Norman Dalkey. The phrase "money pump" appears in Raiffa [84] and Tversky [105]. In their early writings on Bayesianism, De Finnetti provided a Dutch Book arguments for why degrees of belief should conform to the probability calculus; in hinting at a similar argument, Ramsey highlighted the transitivity axiom. As noted by Schick, the Dutch Book

with one of the options, the department head offers him the option to upgrade to the option he prefers to that, for a payment of $25, which Mr. S. accepts. Exhibiting the full extent of his cynicism, the department chair keeps repeating this offer, and Mr. S. pays more money to end up back where he started and start the cycle over again.

Intuitively something has gone badly wrong. Let us mention two apparent problems afflicting Mr. S. First, he has engaged in a cycle of choices. The path independence principle we considered above can be understood ad treating a single choice as if it were determined by a sequence of choices. Here we consider treating a sequence choices as if they were a single choice. What we want to avoid cyclical choice is that the end result of any series of pairwise choices is contained in what would be choiceworthy from the set consisting of all the items that figured in those choices.

Second, if the process continues, Mr. S can be bilked for an arbitrarily large amount of money. The latter seems especially dramatic, and I think it is what persuades many neo-Humeans that they have a defense of decision theory that meets Hume's scruples about restricting practical reason to instrumental reason. For upon a bit of self-conscious reflection, it should be apparent to Mr. S that he is on the road to ruin. And so surely he is suffering a failure of instrumental reason in his dispositions to respond to the offers of the department chair (who, after all, need know nothing more of the world than his victim does in order to take advantage of him).

Although this money pump argument and variants considered below are quite dramatic, some think they don't show anything at all about rationality. Broome says of the victim of a money pump, "It is as though you stole his shirt and then sold it back to him. ... Rationality cannot protect [somebody] from that sort of sharp practice." [16, p. 75] Here Broome sells rationality short, as it can protect you from money pump schemes. But it turns out that there are other protection plans than the one offered by proponents of the money-pump argument.

---

argument for probability presupposes that values are structured in the way purportedly established by the money pump argument.

1.5.3 Worse Off With Cyclical Preferences?

An attraction of the money pump argument is that it appears to provide a purely instrumental justification for the basic structural constraints of decision theory. By showing that an agent who doesn't have structurally correct preferences is disposed to lose out by his own lights (no matter what happens outside his interactions with the exploiter), the agent exhibits a gross failing of instrumental reason. If this argument were successful, it would be of great importance to the neo-Humean's attempt to cling true to his Humean heritage, by showing that decision theory really is nothing more than a kind of instrumental reasoning. Unfortunately, it is from the Humean vantage point that it is easiest to see that the money pump argument doesn't show this kind of flaw in cyclical preferences.

For a Humean (neo or otherwise), any explanation of how the money pumped agent ends worse off must appeal to the agent's own values. But no such appeal can be made, if the agent actually has the supposedly irrational preferences which form the hypothesis of the money pump argument! After all, the explanation of why the agent makes every transaction leading to the cycle of choice is that he chooses what he prefers.

One might feel like protesting to Mr. S: "Look, you started with c and ended up getting it back, even though you could have had b, which you prefer to it." But then Mr. S can just as well reply, "Sure, but if I had stuck with b, then I would have failed to act rationally by not choosing a when I had the chance, since I prefer the latter; but once I had a, then I should have traded to c, which is exactly where I am now." This reasoning has the infuriating nature of a seemingly coherent defense of the flatness of the Earth, but what can the Humean appeal to in order to show that Mr. S is really losing out by his own lights? Thanks to the stipulated preference cycle, any attempt to show that Mr. S. is getting what he doesn't want is counterbalanced by a claim that he is too getting what he wants.

The lesson to be drawn from the money-pump argument against cyclical preferences is not that the agent loses out by his lights. Rather, it is that we cannot make

sense of genuine preferences which are cyclical. Mr. S. neither loses nor wins in making his choices. So this kind of money-pump does not do anything more than provide a vivid illustration of the already mentioned fact that acyclicity of preferences is a necessary condition for rationalizability.

1.5.4 Putting the Money Back in the Money Pump?

This is where the "money" appears to give the appearance of an answer. Although perhaps we cannot directly explain how the agent loses out by his own lights in terms of his choices over a, b, and c, we can explain how he loses out overall in terms of his indefinitely extendible monetary losses. Of course, it doesn't matter that the good in question is money. Anything could play the role in the trades, so long as the agent prefers more to less.

Following Hansson and Grünne-Yanoff, [41] let us call the objects of cyclical preference the *primary alternatives*, and the money-like objects the *auxiliary commodity*. The *composite alternatives* are pairs of a primary alternative possessed by the agent and the total cost paid in the auxiliary commodity. As they note, a main controversy over the money-pump argument is whether it follows from the agent having transitive preferences over the auxiliary commodity that he will choose as if he has preferences over the composite alternatives such that: $< c, -3 > P < b, -2 > P < a, -1 > P < c, 0 >$.

What is the problem with thinking that the agent will choose that way over the composite alternatives, based on his preferences over the primary and auxiliary commodities? I suggest there are two fundamental problems here, each of which go beyond the specific case of cyclical primary preferences. The first problem, which I discuss in this section, is that any considerations about the auxiliary and composite commodities are irrelevant to assessing the rationality of the primary alternatives. The second problem, which I discuss in the next section, is that there are difficulties in going from claims about preferences to claims about how an agent ought to rationally choose based on those preferences.

If Mr. S. is a typical person, then he will happen to prefer not to lose money.

Indeed, by the assumption that he always prefer more money to less, his monetary preferences obey P5. Hence, he may acknowledge that he is engaging in a pattern of choice that violates some of his preferences (for the auxiliary commodity of money), even if it seems to satisfy his other preferences (over the primary and composite alternatives). And even in his somewhat muddled state, we may imagine that Mr. S. would treat an arbitrarily large monetary loss as a calamity compared to the benefits enjoyed by his cyclical preferences over a, b, c.

However, although this shows that something is wrong with Mr. S. since he happens to care about money, it doesn't show anything about somebody who simply has intransitive preferences over some a, b, c but no better-behaved preferences over an auxiliary commodity. That is, the money pump argument does not show that somebody is disposed to lose out by his own lights, in virtue of having cyclical preferences simpliciter.

1.5.5 Is it the Choice Rule?

A different diagnosis of what goes wrong in the money pump argument is offered by Schwartz,[93] one of its earliest critics. Schwartz claims that it is not the preferences which lead to the odd behavior, but a careless application of the principle for choice. Especially, Schwartz argues that one should not choose irrespective of one's past choices or anticipation of future choices. Schwartz pushes this line as far he can, concluding that there is nothing structurally irrational with even cyclical preferences.[15]

In order to make his case, Schwartz argues that a context-free application of principles for choice will lead to unacceptable consequences even when the preferences are as rational as you please. Suppose an agents preferences are aPb, bPc, aPc. Now

[15]Many have followed Schwartz's negative argument in diagnosing that the fault in the money pump lies with the choice principles. However, apart from Anand [5], few have followed him in tolerating even cyclical preferences. Another early contribution on this subject is Schick, [91] who rejects the assumption that transactions are "value-wise independent" of each other. Schick intuits that transitivity is a requirement of preferences but not (for the reasons we gave above) indifferences. Maher [71] critiques the money pump but endorses the transitivity of indifference as well as preference.

give the agent a two-stage choice, first between b and c, then between whatever he picked and a. Schwartz claims that the prudent agent will resist the temptation to trade for c for b in the first round, and wait to get a. After all, if money is being spent to upgrade, then the purchase of b will just be money down the drain. This isn't as dramatic as the possibility of losing endless amounts of money, but it is still a sure and unnecessary loss.

I don't think Schwartz's money sink argument works any better than the money pump argument about Mr. S. Here the problem is not that we cannot make sense of the agent's preferences, but that we have mis-specified what they are. After all, if the agent really does prefer b to c, then isn't the short term enjoyment worth it, even while the opportunity to upgrade to a looms on the horizon? For if it isn't worth it, then it seems we have misdescribed what the agent's preferences really are.

Both the money pump argument against cyclical preferences, and Schwartz's reply, make use of the following principle.

Upgrade Requirement: If you have y and prefer x to it, you should trade.

This seems to be an innocuous principle, which we should be reluctant to give up in favor of Schwartz's claim that all that rationality requires is that preference be asymmetric (our P1). Nevertheless, I think Schwartz's diagnosis of the problem with the money pump is correct, when it is applied to money pumps not involving cyclical preferences. Those money pumps rely on implausible principles for choice between options where the agent does not have a preference.[16]

[16]It is sometimes argued that one should not draw a distinction between these cases. For instance, the cyclical preferences Mr. S over pairs are determined by an incommensurability in ranking these two criteria of salary and prestige. On the basis of this kind of example, Schumm [92] argues that fans of incommensurability of value should "have the courage of their convictions" and endorse the rationality of cyclical preferences. But just because incommensurable values can lead to cyclical preferences does not mean that they always will do so. (Indeed, the argument that they do lead to cyclical preferences relies on their being revealed by the choice rule which we criticize below.) Nor must say the thing in the case where cyclical preferences do arise from incommensurability as in the case where they do not.

1.5.6 Money Pumps Without Cyclical Preferences

According to Davidson *et al*, to avoid cyclical choice it is not sufficient to simply avoid having cyclical preferences, given the rule to choose an option which is not disprefered. As they explain, the only thing which will avoid cyclical choice is transitivity and negative transitivity of preferences (P4). These arguments do not suffer from quite the same problem as the money pump against cyclical preferences, because in these cases it is more intelligible that the agent actually has preferences which he is failing to optimize.

Consider somebody who has transitive preferences over amounts of sugar in his coffee, but displays intransitivities in his lack of preference between differences of individual grains. Whatever he is willing to pay to upgrade from one lump of sugar to another, we can buy back his sugar one grain at a time (at some amount less than his buying price for a lump, divided by the number of grains in a lump). After all, since he has no preference between differences of a grain in themselves, sweetening the deal with a bit of money should move him in our preferred direction. Eventually, having sufficiently depleted his supply, we can sell a lump of sugar back to him, at a profit for us.

The same sort of reasoning applies to the case intermediate between the above two, of intransitive preferences. Given Savage's initial car preferences, he would have been willing to pay a premium to trade a car with a radio installed after purchase for the same model without a radio, and then for one with it pre-installed; at that point he would have been willing to swap back for the first car.

On the face of it, these arguments make use of more than the Upgrade Requirement, since they invoke a principle of choice for situations in which one does not have a preference. Nevertheless, this appearance is deceiving, according to one use made of money in the money pump arguments.

> *Deal-breaker Principle:* If you don't have a preference between x and y, then since you prefer more money to less, you prefer x plus money to y unadorned by money.

By forcing upgrades for slight bonuses, the money pump argument purports to show the irrationality of preferences using only the Upgrade Requirement.[17]

Unfortunately, it is exactly in the cases where the Swapping Principle is most problematic that this infusion of money is unlikely to help. For instance, in a case discussed by Raz, [86] somebody with no preference between two very different career paths is not going to be swayed in favor of one by an arbitrarily small bonus. Indeed, this is the central feature of incommensurability.

In any case, use of the Deal-breaker Principle suffers from the same problem that afflicted the use of money against cyclical preferences. It changes the topic from the internal irrationality of intransitive (negatively intransitive) preferences, to the irrationality of such preferences combined with external preferences over money.

1.5.7 Swapping Requirements versus Status-Quo Maintenance

So let us confine ourselves to money-pump style arguments that do not make use of any trades into money, or other commodities external to the allegedly irrational preferences being critiqued. Here is one candidate for the principle at work in these cyclical choice arguments.

> *Swapping Requirement*: If you don't have a preference between x and y, you should be willing to trade whichever you have for the other.

The argument for why acyclical preferences should be transitive makes use of a Upgrade, Upgrade, Swap series of transactions; the argument for why transitive preferences should be negatively transitive makes use of a series of repeated Swaps followed by an Upgrade.

The Swapping Requirement is much more questionable than the Upgrade Requirement. Its opponents often focus on cases where lack of preference is not a case of genuine indifference, such as incommensurabilities of value. However, it isn't especially compelling even in the case of perfectly commensurable and transitive preferences.

[17]For the argument to get off the ground, the money commodity needs to be sufficiently divisible in order that our net profit isn't eliminated by the amount we pay the agent to trade for options he has no preference between.

Fans of the Swapping Requirement defend it on the grounds of revealed preferences: if somebody refuses to swap x for y, that shows they really prefer the first to the second. But this this defense rests on a crude behaviorism, as there is an analysis of preferences as dispositions which avoids the Swapping Requirement.

Preference is an all-things-considered judgment: a disposition to choose in all circumstances. But when we consider lack of preference, we have to be careful about the scope of the negation. To not have a preference between a and b is to not have a disposition to choose one over the other in all circumstances. It isn't a disposition to choose either in all circumstances. Due to past history, or the opening up of future choices, a might *become* uniquely choosable over b. Nevertheless, this isn't a change in preference, strictly speaking, because in the past b was choosable too. But then we should not treat as a sufficient condition for rational choice that an option is undominated. For although neither a nor b may be dominated by an all-things-considered preference, the particular circumstances may conspire so as to rule one out.

Indeed, there is a perfectly uniform rule which agents can use to avoid cyclical trades over intransitivities: status-quo maintenance.[18] This will allow for an agent to act upon P2 or P3 preferences without lapsing into cyclical choice. Indeed, status-quo maintenance isn't merely a strategy to avoid cyclical choice, as it makes sense even for somebody with perfectly rational P4 preferences. It is often said that if you are indifferent between a and b, then you ought to be willing to trade between them, and if you are unwilling to trade a for b that shows you prefer it. But this doesn't really seem right, since not yet having either of a or b you could be disposed to take either, yet whichever one you end up with not be disposed to trade for the other. Hence it is not mandatory to be willing to trade between options you have no preference between, although it is sometimes permissible (and is always permissible when preferences are

[18]This is advocated by Mandler [72]. Note that status-quo maintenance does not make sense as a way of coping with cyclical preferences, given the central status of the Upgrade Requirement. Some economists and political scientists claim that a desirable property of a social choice rule is to avoid status-quo maintenance, as this would be an irrational force of social conservatism. The present usage does not seem to suffer from that negative feature, since the agent is not being held back from achieving anything she regards as preferable.

negatively transitive).

1.5.8 Swapping Permission and Life-Planning

But the money pump argument can be reconstructed without Swapping Requirement, making use of only the following weaker principle.

> *Swapping Permission*: If you don't have a preference between x and y, you may trade whichever you have for the other.

By following this principle, along with Upgrade Requirement, an agent may engage in a cyclical series of choices, even if rationality doesn't require him to.

Rationality should not allow an agent to engage in such a choice cycle. But rather than blame the preferences, we can blame the rule of Swapping Permission. With it, we must jettison as a general rule the principle that an agent may rationally choose from any of his preference-undominated options. That should be replaced with the weaker rule that an agent must choose from amongst his undominated options. What the money pump argument shows is that the stronger rule is only appropriate for agents who do have P4 preferences.

This brings out the full extent of the opposition noted earlier between Plato and Aristotle. The requirement to form a unity out of preferences extends to a unity across choices in time. Otherwise, the rule to pick from undominated options leads to unacceptable results. Thus we see a precise articulation of Aristotle's complaint that Plato's principle is only suitable as a standard of planning one's whole life in advance, an inhuman standard going beyond what can be asked of the person of practical wisdom.

When Savage employed P4 as a constraint on preferences, he explained that his theory was essentially one in which the agent made a single plan for the while course of his life. [90, p. 83] A major unsolved problem of his theory was how to relate such a "grand world decision problem" to more realistic "small world" problems. This is the problem of finding a path between Plato and Aristotle. Savage put the point by saying that the truth lay between two opposed homilies: "Look before you leap", and "You

can cross that bridge when you come to it". [90, p. 16] Together with the apparent reasonableness of the counterexamples to negative transitivity, the unreasonableness of planning out one's whole life in advance seems to count against P4 as a rational constraint on preference.

But if Swapping Permission is false, then what are we to make of the undominance principle and choice functions? What seems to be left of the normative rule is this: one must choose from amongst the undominated options. But your preferences alone do not determine which ways you may do so, that is also a function of your history and anticipated future choices. Likewise, what is left of choice functions as a descriptive device is that they do not fully represent the agent's dispositions, unless he obeys at least P4.[19]

1.5.9 Moving Beyond Path Independence

Insofar as money-pump arguments use generally acceptable principles for choice, they establish nothing more than the requirement of acyclicity for choice being rationalized by a preference relation. Insofar as money-pump arguments can establish anything stronger, they settle on the classic P4 condition on preferences. But they only do so by means of a choice rule that is too strong, albeit one which is used in the classic theory of revealed choice.

The correct criterion of rational preferences—transitive but not necessarily negatively transitive—lies in between these two, and is not established by any money pump argument. Although the synchronic path independence argument purported to establish it, it was afflicted by problems, now compounded by the realization that the argument used the rule of choosing any undominated option (which is only supportable in case preferences are not only transitive, but also negatively transitive).

Nevertheless, we can still settle on transitivity as a threshold for preferences, by means of arguments considered before we turned to path independence. The platitude that a uniquely chosen option is preferred did not make full use of the principle that

[19]This was the condition they were originally designed for, with weakenings of the concept of rationality coming later in the development of economic theory.

one may choose any undominated option, since it focused on the case where that principle collapses into the rule that one must chose from amongst the undominated options. This is reinforced by the argument that an acyclic preference is uniquely extendable to a transitive one. Next we fit those arguments into a theory of the nature of preferences and preference attribution, which aims to make sense of how we may use the principles of rationality in cases that deviate from them.

1.6 Constitutive Constraints and Fragmented Minds

1.6.1 Where Reasons Come to an End

Later in his career, Davidson provided a very different answer as to why rationality requires transitivity of preferences, it is a constitutive part of our concepts of rationality and preference.

> Just as the satisfaction of the conditions for measuring length or mass may be viewed as constitutive of the range of application of the sciences that employ these measures, so the satisfaction of conditions of consistency and rational coherence may be viewed as constitutive of such concepts as those of belief, desire, intention and action. It is not easy to describe in convincing detail an experiment that would persuade us that the transitivity of the relation heavier than had failed. Though the case is not as extreme, I do not think we can clearly say what should convince us that a man at a given time (without change of mind) preferred a to b, b to c, and c to a. The reason for our difficulty is that we cannot make good sense of an attribution of preference except against a background of coherent attitudes. [23, pp. 236-7]

To say that there is a constitutive relation between preference and other mental states is one thing, but to pin down the exact nature of the constitution is another.

Anand questions Davidson's analogy with "length", and suggests that perhaps preference is rather like "ability to win in a sports league": team a might be able

beat team b, and b to beat c, yet c can beat a. In response to this, we can point out that if preferences were not transitive, then it would not connect with the platitudes about choice that we discussed earlier. A counter-argument is found in Anand's critique of the idea that choice out of triples should be governed by choice out of pairs. He claims that is to conflate a tertiary relation with a binary relation. Perhaps Anand's complaint would have some force if we were discussing the "preferred out of" relation, in which case it would amount to a challenge to justify the consistency conditions on choice functions. However, it is harder to make sense of as a critique of the standard treatment of "preferred to" as a binary relation, which is related to choice as described earlier.

With the constitutive theory comes the abandonment of the question of why somebody should be rational, which perhaps comes as a relief as the question has a paradoxical air to it. Later, Davidson writes,

> I should never have tried to pin you down to an admission that you ought to subscribe to the principles of decision theory. For I think everyone does subscribe to those principles, whether he knows it or not. This does not imply, of course, that no one ever reasons, believes, chooses, or acts contrary to those principles, but only that if someone does go against those principles, he goes against his own principles. [25, p. 195]

Arguments such as the money pump are best viewed as attempts to remind an agent that he does in fact accept the principles of transitivity.

The constitutive theory, as Davidson presents it, also helps to make sense about intuitions that deviations from the theory of rationality can still count as rational.

> The kinds and degrees of deviation from the norms of rationality that we can understand are not settled in advance. We make sense of aberrations when they are seen against a background of rationality; but the background can be constituted in various ways to make the various forms of battiness comprehensible. So it would be a mistake to put too much weight on the examples of irrationality that I have chosen, and worse to

> worry whether I have in each case drawn the line between principles constitutive of rationality and potentially intelligible flaws in just the right place. The essential point is that the more flamboyant the irrationality we ascribe to an agent, the less clear it is how to describe any of attitudes, whether deviant or not, and the more basic we take a norm to be, the less it is an empirical question whether the agent's thought and behavior are in accord with it. [25, p. 196]

The notion of degrees of deviation from an ideal of rationality helps make sense of intuitions had by many that violations of P4 are less irrational than violations of P3. In the next few sections we shall discuss ways of handling lapses from rationality.

1.6.2 Rational Choice Without Preferences?

By treating transitivity as a constitutive condition of preferences, we help make sense of the intuition that there are cases where rational choice occurs despite not having a transitive preference. What is going on in those cases is not that one has an intransitive preference. Rather, one is choosing in the absence of a genuine preference relation.

This offends against the doctrine that rationality just is maximizing or optimizing a preference relation. But it seems to fit the intuitive data. After all, a preference relation is an all-things-considered judgment. And in the cases where rational choice occurs without an underlying transitive preference, we have cases where the various factors holding between separate pairs are not treated by the agent as all fitting together into a single ranking. The agent has plural and conflicting values.

This leads to a better interpretation of what the "preference" relations are that do not satisfy transitivity. They express "better (in a way)", rather then "better (all things considered)". To choose something undominated in the weaker sense of preference is to choose something such that there is no object better than it in some way. In practice we often neglect to emphasize the difference between relative and absolute preferences, and just speak of "preference".

This way of regarding choice behavior that cannot be rationalized by a transitive preference relation fits well with the standard definitions of weak revealed preference relations. Strict preference and indifference are defined in terms of a more primitive weak preference relation, give by: $x\hat{R}y$ iff there is some S such that $x \in C(S)$ and $y \in S$.[20] Note that this approach doesn't entail any more than asymmetry of the $\hat{P}$ preference relation. Although the relation $x\hat{R}y$ is often glossed as "x is at least as good as y", this is misleading, as can be seen from the very definition of the relation. A more accurate gloss of $x\hat{R}y$ would be "there is a way in which x is at least as good as y". The glosses on the $\hat{P}$ and $\hat{I}$ relations would then be: "there is a way x is as good as y, and no way in which y is as good as x" and "there is a way in which x is as good as y, and a way in which y is as good as x".

Levi [55] advocates the view that rather than try to resolve such values into a coherent whole, one should adopt further rules of decision for picking from amongst options not ruled out as overall dominated. Jeffrey, on the other hand, claims that we should strive for more:

> In practice, we achieve coherence only for preference rankings that involve small numbers of propositions: for small fragments of what our total preference rankings would be, had we the time, intelligence, experience, sensitivity, and patience to work them out. To deliberate is to try for local coherence is in the face of local and temporary conflict; and, since the outcome is action which may be quickly done, we need seldom pause to relate different deliberations. To believe in reason is to think we do better to the extent to which we relate our deliberations to one another and to the extent that we can integrate the fragmentary preference rankings that emerge from our several deliberations into a single coherent structure. [49, pp. 533-4]

Next we will consider a model that is a plausible candidate for making sense of how the norms of decision theory provide an ideal of coherence, despite our lapses from it.

[20] See Sen [94]. $x\hat{R}y$ coincides with the stronger "base relation", defined as $x\bar{R}y$ iff $C(\{x,y\}) = \{x,y\}$, when the choice function satisfies Chernoff's condition.

1.6.3 Comparison With Deductive Rationality

A classic solution to puzzles of irrational action is to treat the person as composed of disparate elements which have failed to achieve the level cooperation that is characteristic of a fully rational agent. In Plato's hands in the *Republic*, irrationality is explained in terms of feuding between the higher and lower elements of the soul, in analogy with the different castes in his conception of society. Davidson proposes that rather than pit reason against passion, we understand irrational action as the result of a mind "partitioned" into separate, rational elements. [24, pp. 180-2] Although each part of the mind is rational in itself, the interaction between them is that of non-rational causality. In this way, a person can act against his best judgment, yet do so intentionally.

Stalnaker has proposed such a model of "compartmentalized" mental states to explain lapses from ideals of deductive consistency and closure in belief. [103, pp. 82-4] He suggests that we view the mind as composed of multiple "acceptance" states, each of which is logically perfect. (If A is accepted, than $\neg A$ is not accepted; if Θ is a set of accepted propositions and it entails A, then A is accepted.) An acceptance state represents an agent's dispositions under certain circumstances to respond to new information, and to act in accord with its desires.

An agent believes a proposition if there is some acceptance state where that proposition is accepted. But only in the (unrealized) case of a logically perfect agent does the person believe all the consequences of all of his beliefs. So a person can both believe A and believe $\neg A$, in virtue of being in a separate acceptance state for each, yet not believe their conjunction $A \wedge \neg A$.[21]

This treatment of departures from logical perfection in belief suggests a parallel theory for rational preference. Just as logical consistency and closure are ideals of rational belief, transitivity and negative transitivity are ideals of rational preference. However, an agent has a diverse variety of values. We individuate separate values

[21] Stalnaker does claim that beliefs are closed under disjunction introduction, and defends this by the "pragmatic" picture of belief: whatever actions are appropriate for the belief that A are also appropriate for the belief that $A \vee B$.

so that each obeys the ideals of rational preference. But the agent himself will only obey the full conditions of rational preference if he has integrated his separate value systems into a coherent whole.

In the example of the child choosing amongst gifts, there are two separate ranking systems, each of which we may presume is coherent. The child ranks the bicycles on the one hand (in terms of the one factor we considered, the bell option), and on the other hand ranks the ponies (this was made especially easy in that we did not consider ways in which ponies vary from each other). In the example of Mr. S. there were the criteria of salary and prestige. The child's preferences, although they did not fit together to yield negative transitivity, did not prevent a choice from being made on their basis. The cyclical preferences of Mr. S., however, did not afford the basis for a choice out of the full menu of the three professorships.

If the analogy between the norms of rationality could be maintained, it would provide support for the position that decision theory can be used to "police" one's preferences just as logic can be used to regulate one's beliefs. However, there is a difficulty in pressing the comparison between logic and decision theory. In the case of logic we have an external criterion of correctness for belief: truth. It is because no proposition can be both true and false that beliefs ought to be consistent. And one ought to believe the consequences of one's beliefs because the consequences of true propositions are also true. But in the case of preferences, we have not found such an external standard of correctness, at least not for the Humean.

This is not to say that we cannot use the partitioning/compartmentalization approach to describing lapses from perfect rationality of preferences. But it does suggest that there is less to be said about why we should have rational preferences than there is to say about why we should have rational beliefs. The neo-Humean should be able to accept this, once he abandons the attempt to provide such an extra justification by the money pump.

Although the compartmentalization of the mind into diverse "value homunculi" can account for some of the apparently rational divergences from decision theory, it cannot account for all of them. For instance, it isn't the source of the phenomenon in-

duced by perceptual indiscriminability, as in the sugar example.[22] We might describe the sugar example as a case of a different kind of fragmentation of the mind, where the agent's value system does not coordinate with his perceptual system. In any case, even if it is treated wholly differently than the phenomenon of multiple values, we still have a principled explanation of why a rational divergence from P4 occurs. But we must still account for further apparent counterexamples to decision theory that seem to call for a more radical treatment. For instance, incommensurable plural values do not seem to be lurking in the shadows in Luce and Raiffa's example of the restaurant customer. He seems to have orderly preferences of: good steak, salmon, bad steak, frog's legs.

In order to reconcile the theory with the data, we need to handle the fact that fragmentation can go beyond multiple values attached to objects described extensionally. Our preferences sometimes also involved treating objects intensionally, in ways that prevent the usual goals of analyzing decision problems in terms of other decision situations apparently involving the same objects (extensionally described).

1.6.4 Loading up the Consequences and Reflective Endorsement

When Savage realized that he had intransitive preferences amongst what he regarded as the relevant options in his choice of a car—the two models before him, as well as a hypothetical option—he changed his preferences. However, as Savage himself emphasized, care must be exercised in simply describing what the consequences of a decision problem are.[23] So rather than change his preferences, Savage might have taken the apparent intransitivity as reason to think he had misdescribed the options he faced.

Savage might have come to think that he really placed a value on the difference between whether a car had a radio installed before purchase rather than after. (This

[22]However, the intransitivities induced by basing preferences on sensations can be a source of the lack of coherence among multiple values that would otherwise fit together.

[23]See his discussion of making an omelet [90, pp. 14-5].

difference in value might be reflected in some purely factual matter which he had initially left out of his description, such as the effort to bring the car into the garage for retrofitting with a radio. However, it might simply be a brute preference.) If Savage had discovered that his preferences really were this way, it would left him with a way of resting content with them.

To do this, he would have to recognize a difference between (1) the option of a car without a radio when the other option is a radio pre-installed, and (2) a car without a radio when the other option is a radio post-installed. When the options are described this way, his dispositions express transitive preferences. He prefers no-radio car (2) to the post-installed car, he is indifferent between the later and a pre-installed radio, and that is preferred to no-radio car (1).

Of course Savage did not avail himself of this method of saving his initial preferences. Moreover, this sort of redescription maneuver has been subject to much criticism for making decision theory "vacuous". For it seems to allow that any dispositions could be maintained as rational simply be finely individuating the options. Broome presses the worry that we cannot "convict" anybody of irrationality, unless we go beyond Humean formalism and adopt substantive principle of reason, so that an agent is in error if he values things out of accord with their real value.

Broome considers a response that the Humean is interested in rationality as a "guide that helps people conduct their own affairs" rather than as a "criterion for condemning people from the outside". [16, p. 76] This is the sort of priority that Savage himself placed on decision theory, which he regarded as primarily a normative theory for applying in the first-person, rather than a descriptive theory for applying in the third-person. The two uses of decision theory cannot be divorced from each other. As Davidson emphasizes, the descriptive task requires a presupposition that the person is generally rational in the normative sense. And in the case of self-regulation, one must describe what the preferences are which are being normatively assessed. But it is at this point which Broome claims the Humeans enterprise of defending decision theory collapses.

According to Broome, there is no way the Humean can give an account of the

epistemology of the preferences that are involved in re-describing the options so as to save transitivity, except in a way that makes them constituted by the process of describing the options. By individuating options by reference to choice situations, we open the agent to the question of whether he has a preference between no-radio car (1) and no-radio car (2). But the agent could never be faced with a choice between those two, since the options are defined by reference to mutually exclusive choice situations. So the agent cannot assess whether he has a "practical preference" between them by considering hypothetically how he would choose between them.

So how, Broome asks, can the agent tell whether he has a "non-practical" preference between such option?[24] As he says, they do not seem to generally be accompanied by an introspectible feeling, despite everything Hume himself said about the contents of the mind. The only option left, Broome says, is that non-practical preferences are whatever would make the practical preferences satisfy the rules of transitivity. Broome protests that this prevents the non-practical preferences from providing any sort of justification for the rationality of the practical preferences.

In response to Broome, the Humean should happily grant that the non-practical preferences are whatever will make the practical ones turn out to be rational. To avoid biting any bullets here, there are two things for the Humean to say against Broome's critique. First, Broome has underestimated the extent to which the Humean picture is coherentist, and has tried to saddle his opponent with unwanted foundationalist goals. The Humean theory is that a set of preferences is rational if it obeys the appropriate principles of decision theory. It is not a theory where certain preferences (the practical ones, in this case) are to be justified by, without justifying, some other preferences (the non-practical ones).

Second, it is up to the agent whether or not to adopt a re-description of the options, so as to retain his preferences as rational. Savage himself did not adopt the re-description strategy, nor do many subjects in the experiments by Tversky and

[24] Jeffrey adopted "news values" in response to the problem of assessing preferences between causally impossible gambles required by theories of Ramsey and Savage. Broome's dilemma can be treated as a challenge to making sense of the metaphysics and epistemology of Jeffrey news values.

others; they changed their preferences in response to the perceived intransitivity. Hence the mere possibility of redescribing one's preferences does not leave agents in the position of being unable to apply decision to themselves as a genuine norm (likewise in the case of third-person descriptions).

There might be somebody who is disposed to never alter his preferences in response to apparent intransitivities. And there might be somebody who is disposed to never alter his description of the options instead. Most of us probably fall somewhere in between. The situation is again partially, but not completely, analogous to that of consistency in belief. If somebody is caught contradicting himself, he can always protest that he is being misunderstood, and what he really meant is consistent.

This sort of move doesn't go as far in the case of belief as it does in the case of preference, because there are external causal constraints on content (see, for instance, Stalnaker [103]). These external constraints limit the possibilities of Humpty Dumpty voluntarism about meaning. But for the Humean there are no such external constraints on value. (Indeed, even for most non-Humeans there will be a wide range of cases not settled by the agent's needs or telos, or objective values in the fabric of the universe.) This is something the neo-Humeans should be fine with, since although they go beyond Hume in allowing reason to regulate preferences, their position would be quite un-Humean if there was not a fair amount of subjectivity and arbitrariness left in what is allowed of preferences.

1.7 Conclusion

Standard decision theory finds constraints on preferences where Hume found none. Some of these constraints, the asymmetry and transitivity, turn out to be constitutive conditions of preference. However, this ideal of rationality is difficult to live up to, given the plurality of different desires. We can reconcile the ideal of rationality with the phenomena by treating the mind as fragmented, and applying the ideal to the fragments. Decision theory serves as a normative tool, by giving guidance in unifying the diverse preferences. Unlike logic, it is not regulated by an external norm.

The further constraint of negative transitivity of preferences, or transitivity of indifference, is an optional extra condition for us. It fits a more idealized conception of a rational agent, who is prepared for all eventualities. It need not even be met by the mind's separate value centers, and often will not be, thanks to the imprecisions in our perceptual systems that our desires piggyback on. Money-pump arguments purport to show that when a person fails to live up to these standards, the agent will be susceptible to a sure loss, and thus fail to be instrumentally rational. These arguments were found to fail, relying on implausible principles for choosing from options where the agent has no preference.

Chapter 2

Cognitivism and Two Kinds of Desire

2.1 Introduction

A striking view espoused by Plato is that everybody desires what is good, and that wrong action is the result of ignorance about what is good. Plato's views are reflected (at least in part) in the writings of modern cognitivists, who claim that value is a subject matter about which one may have true beliefs. Against them stand Hume and his legions of followers, who maintain that so-called value judgments are not beliefs at all but rather expressions of passions.

David Lewis has argued that the Humean position receives strong support from decision theory. Lewis gives a formal proof within evidential decision theory which he claims shows that beliefs and desires are distinct existences. John Collins has reinforced this argument with a proof of a similar result in a purely qualitative model which relaxes the numerical precision of standard decision theory.

Lewis aims his argument against those cognitivists who are internalists: they side with Plato's apparent view that beliefs about what is good have a motivational force. However, as I shall explain, the reach of Lewis's argument extends to cognitivists who are externalists: they hold that one may have a belief about what is good without being motivated to act on it. So I shall argue that contrary to how it has been treated

in the literature, Lewis's argument should be troubling to everybody who thinks that value judgments are a subject of belief, insofar as they accept Lewis's proposal for connecting desires and beliefs.

Not only does Lewis's argument have a broader target than has been recognized, the arsenal aimed against the target has more underlying unity than has been recognized. For although Lewis and Collins intend their theorems to be in the same spirit, it has not been clear to what extent this can be maintained due to differences in the letter of the proof techniques, apparently stemming from the differences in background assumptions. I shall show that not only do the theorems appear to be about the same subject when viewed from afar, this impression holds up under closer scrutiny. I shall do so by presenting a proof that displays the underlying unity of their results.

Having raised this general alarm for cognitivism, I shall attempt to dispel the worry that it raises. On independent grounds, causal decision theory is superior to evidential decision theory as a criterion of rational action. It turns out that causal decision theory also provides the best framework in which to precisely articulate the pre-theoretical cognitivist idea which Lewis has drawn our attention to. My proposed solution to Lewis's puzzle explicates desire as a causal utility, and beliefs about instrumental goodness as ones about causal conduciveness to goodness.

The main move that needs to be made in order to allow this as a cognitivism option is to argue, against Lewis, that causal utility is a type of desire. This is what I shall do, but I shall also argue against the claim by Byrne and Hajek that the way to do this is to deny that evidential value is a type of desire. I shall advance the position that the causal decision theorist should recognize two kinds of desire, a passive and an active kind, which converge in all but Newcomb problems.

The modest cognitivist will rest with having a way of formulating within decision theory a position which avoids the force of the triviality proof. But the more ambitious cognitivist will also claim that is the intuitively appropriate move to make, since the usual conception of desire in the internalism debate is one conducive to action rather than the passive receipt of good news; that is, it picks out causal utility rather than

evidential value. I claim to be able to place the modest cognitivist on firm ground against Lewis. The ambitious cognitivist has further to travel than I am able to guide him, but I shall sketch the beginnings of his route for him.

The treatments of goodness studied here all hang together, or fall apart, based on analyses of conditionals. Evidential value has implicit in it a notion expressed through the indicative conditional, and causal utility is articulated by means of a subjunctive conditional. The triviality theorems of Lewis and Collins about desire as belief are based on earlier triviality theorems about conditionals. Thus, deciding on the proper form of moral cognitivism draws on independent work in the philosophy of language and the philosophy of decision making. This interdisciplinary perspective carries on the rich tradition of meta-ethics, which has tried to elucidate value judgments in terms of linguistic analysis and the study of rational action.

Here is a map of the paper. Section 2 briefly motivates cognitivism, and gives a short sketch of the strategy of Lewis's triviality proof. It explains how Lewis's theorem is best construed as targeted against all forms of cognitivism, by disentangling Lewis's argument from various sub-debates within cognitivism: most notably internalism versus externalism about motivation, but also relativism versus universalism about normative contents. Section 3 presents a triviality proof displaying the unity of the theorems of Lewis and Collins. Section 4 considers the relative plausibility of various responses to the triviality result of the section preceding it. Section 5 explains how to use causal decision theory to formulate the anti-Humean position, and defends the claim that there are two kinds of desire. Finally, Section 6 briefly wraps up the discussion of how cognitivism may be coherently formulated in decision theory, and compares the plausibility of anti-Humeanism to the more standard Humean treatment of decision theory.

2.2 Beliefs About Goodness and Their Alignment With Desire

In 2.1, I briefly explain what cognitivism is and why it is worth defending. In 2.2, in order to present the real force of Lewis's argument, I explain the distinction between internalist and externalist forms of cognitivism and how they relate to the Humean opposition. In 2.3, I explain another distinction which affects the interpretation of Lewis's proof; the reader interested in getting to the heart of the argument may skip this section, referring back to it later if interested in the question of whether our formalisms should apply to all agents or just a single agent.

2.2.1 Cognitivism

Cognitivism, as I use the label, is the view that normative judgments involve beliefs, as opposed to being mere expressions of emotions. This is a rough statement of a familiar view which can be made precise in various ways, especially by clarifying what a "belief" is and what the contrasting class of "emotions" is. But even as roughly stated, cognitivism has many attractions. On the face of it, statements about what is good are of the same form as declarative sentences about straightforwardly descriptive matters. This suggests that talk about goodness expresses beliefs.

Going beyond that basic statement of cognitivism, there are many attractions to taking a realist view of normative beliefs, so they are about a genuine subject matter about which we can have true or false opinions. Questions about what is good are argued over as if there is a right or wrong answer to them. Indeed, there are attractions to viewing morality as a subject about which genuine knowledge is possible. For moral arguments are conducted in a serious spirit as if the answers to moral questions can be ascertained by rational inquiry. Although these are attractive elaborations of the basic cognitivist view, they will not play much role in what follows; and so I shall now elaborate on just how little is required to count as a cognitivist for our purposes.

One can be a cognitivist without buying into a full-fledged realism embracing moral knowledge. For instance, although Socrates claims in the *Protagoras* that virtue is a kind of knowledge, in the *Meno* he concludes that it is merely true belief without a justification. Socrates reaches the latter view by way of concluding that virtue is unteachable, and hence is a gift from the gods. Some of us might put the point by saying that moral arguments are intractable in a way that non-moral arguments are not, and hence true moral beliefs are acquirable only by the luck of clear intuition or by a proper upbringing.

Weaker forms of cognitivism may not be as attractive as a full-fledged realism which allows for a robust form moral knowledge, but it turns out they are all affected by the main argument presented by Lewis. All that it takes to be subject to the formal proof presented later is that moral judgments behave formally like ordinary beliefs, in a few simple ways that will be precisely spelled out in setting up the proof. It is important to keep in mind how large the family of cognitivist positions is, since Lewis and the commentators on his argument have failed to emphasize this. Part of the reason that this has been overlooked is that Lewis has unnecessarily presented his argument as being aimed at a limited group of cognitivists, and so we shall now examine the dichotomy underlying Lewis's choice of targets, in order to show how it distracts from the true scope of positions affected by his proof.

2.2.2 Externalism, Internalism, and Humeanism

Sincere expressions of moral judgments are typically accompanied by a desire to act on them. Cognitivist theories gravitate toward two opposed explanations of what this connection between belief and desire consists in. Externalists claim that the connection is contingent, although perhaps typical thanks to psychological and cultural forces. The internalist cognitivist claims that the connection is necessary; being motivated is constitutive of having a genuine moral belief. A popular way of defending the externalist position is by pointing out that it is possible that there could be an "amoralist": somebody who has moral beliefs but who never desires to act on them. An internalist response is that such a person does not make genuine moral judgments,

but only speaks in a "scare quotes" sense of parroting the beliefs of those who are sincerely motivated to act on them.

The internalist cognitivist position is sometimes expressed (for instance by Pettit [80]) by saying that there is a single state which is both a belief and a desire. Altham [4] calls this a "besire", which he intends to connote that such a unitary state is "bizarre". There is a variety of subtle distinctions which can be drawn, and hybrid views that can be defined, but for our purposes we can just work with the basic dichotomy of externalist and internalist cognitivisms.[1]

Against both cognitivist doctrines stands Humeanism, which is internalist but drops the cognitivist claim that moral judgments express beliefs. Despite whatever trappings of beliefs might be possessed by moral judgments, Humeanism says they are merely expressions of desires. Ignoring the externalist theories, Lewis surveys the debates over internalist cognitivism versus Humeanism, and pronounces, "All this skirmishing is inconclusive." [64, p. 45] His sympathies gravitate toward the Humean view, but his "doubts rest on intuitions that might be easy to controvert." [65, p. 60] His intention is to decide the question nearly conclusively, by giving a formal characterization of the internalist cognitivist theory, and mathematically proving that (so formulated) it is incoherent.

The basic strategy of Lewis's main argument is as follows. First, equate a rational agent's motivational states with the values described in an appropriate decision theory. Second, look for an appropriate belief about goodness, in order to satisfy cognitivism. Then, given those two explications of desire and beliefs about goodness, *demonstrate that desires and beliefs about goodness cannot coincide across a significant range of learning experiences an agent might undergo.* This will give a precise articulation and defense of the Humean claim that desire and belief can always be separated.

[1] Dreier [28] presents a sophisticated internalist theory, and gives a nice survey of many prominent positions in the field. A more thorough survey is given by Wallace [107], who makes some remarks about Lewis's argument that I am largely sympathetic to. Smith, who is a Humean about the desire-belief distinction but not about rationality, tackles the internalism-externalism dispute in his comprehensive treatment of the central meta-ethical debates [98]. In his main positive contribution to meta-ethics, Lewis gives a sort of ideal observer theory [66] which aims at reconciling internalist and externalist strains of thought.

Lewis aims his argument against internalist cognitivism. A necessary connection between desires and beliefs about goodness will be refuted by a demonstration that they do not stay connected across possible changes of rational belief. Although the force of the argument against internalism is immediate, the lesson of Lewis's proof counts heavily against externalism, too.

The easiest way to see that Lewis's argument can be applied to the externalist is to consider the opposite of the amoralist. Externalism ought to be compatible with the existence of a "saint": an individual who just happens to desire exactly what she believes to be good. But Lewis's argument against desires being accompanied by beliefs will show that the saint is impossible, even though the alignment of desire and belief is contingent. (In fact, Lewis's argument will apply to those of us of middling virtue, to the extent that we do desire what we believe to be good.) That in turn would be pretty decisive against the thesis that there are any moral beliefs at all, since surely a rational person's changes in belief should not thwart the alignment of her beliefs about what is good and her desire to act on them.

In other words, if we accept Lewis's setup, we are driven toward full-fledged Humeanism about value. Fleeing "besires" we land in the camp of emotivism. But although resolute Humeans about value embrace this conclusion, I shall argue that even they should be cautious about how Lewis's argument leads to it.

2.2.3 Absolutism versus Relativism

The challenge presented by Lewis's triviality result is about intrapersonal changes in belief. However, the precise formulation of Lewis's equations raises issues about interpersonal agreement and disagreement in belief. This arises in the first instance as a question of how to best formulate Lewis's argument; we are faced with the question as to whether they are universally quantified over all rational agents (as Lewis usually assumes), or whether we may profitably restrict their scope. But the issue also brings to the surface a secondary question about the nature of normative beliefs, which comes from within the heart of meta-ethics. I shall explain how in answering the first question, we can bypass answering the second question (just as we can bypass

resolving the internalism-externalism debate).

Each of the formulations we shall consider contains as a component a proposition about goodness. One way of coming to be dissatisfied with all of these formulas is that they cannot be fleshed out so as to do justice to two competing theses about value, which should be acceptable to a sensible anti-Humean. On the one hand, barring a truly extreme anti-Humeanism, the standards of rationality permit that value is at least to some extent contingent and variable amongst different agents. There are some issues which are simply a matter of taste. On the other hand, value must be to a certain extent objective, so there is a fact of the matter which agents may be right or wrong about in their valuations. The anti-Humean wants to explain at least some differences in value as differences in belief.

If we try hard to respect the first point, then we will end up treating the equation as holding for a single agent whose desires align with his beliefs about goodness; it interprets his personal conception of what goodness it. For other agents whose desires and beliefs are suitably aligned, we could give a similar looking equation rationalizing his state of mind, but there would no guarantee that they would express the same proposition about goodness.

If we try hard to respect the second point, then we will treat the equations as holding for all rational agents, insofar as their desires align with their beliefs about goodness. But then we end up with a puzzle about how disagreement about goodness is possible in the face of agreement about "natural" facts (by which I do not mean to commit myself to some deep metaphysical dualism, but simply to get at the intuitive idea that there are facts which are not straightforwardly about goodness). At least we have this puzzle if we hold that goodness supervenes on natural facts. In representing disagreement in goodness within our decision theoretic model as a disagreement in belief, there is indeed some temptation to deny that goodness supervenes on natural facts.

It is important that cognitivism have a story to say about what propositions are expressed by normative language, how those contents are determined, whether facts about goodness supervene on natural facts, and how agreement and disagreement

are possible. But we need not concern ourselves with these issues here, even though the response we give to Lewis's theorem may have consequences for what we end up saying about those issues. What we need to worry about right now is whether it is possible for a single agent to align his desires with his beliefs about goodness, in a way which is stable with respect to changes in belief. We need not worry whether the equation we consider is shared amongst all agents.

This is in fact related to the point made above about how the proof is largely orthogonal to the externalism-internalism debate, in that both points affect how we quantify the equations which Lewis presents. We do not need to worry, at least in the first instance, about whether the equations hold across all rational belief states (as they would in the internalist theory which Lewis has in sight). We need to worry about whether *given* an agent for whom the equation *does* hold at one time (representing alignment of desire and belief), the equation will continue to hold under changes in belief. If the equation is inherently unstable in the face of learning new information about the world, then externalism is rendered implausible (along with internalism), since even though externalism allows for beliefs and desires to come apart this should not be forced in the way presented in the triviality proof.

2.3 The Setup and Proof of the Triviality Result

In 3.1, I present the formal background for what follows. In 3.2, I explain Lewis's basic simplifying assumption about degrees of goodness, and present an equation discussed by Lewis which expresses value as a conditional belief. In 3.3, I present Lewis's target equation of desire as belief. I show how it can be understood in terms of the previous equation, and how it relates to another essentially equivalent formulation of desire as conditional belief. Using the main premise of Lewis's triviality proof, I derive from Lewis's desire as belief equation a qualitative condition. This condition is entailed by Collins's qualitative formulation of desire as belief. In 3.4, I prove the triviality of that condition, thereby demonstrating the theorems of Lewis and Collins.[2] In 3.5

[2]Collins's piece [19] appeared together with Lewis's first article [64] on the subject, in the same issue of *Mind*. Lewis continues his argument in a follow-up piece [65], as well as in the first footnote

I compare the triviality result with triviality results about conditionals, and in 3.6 I offer a diagnosis of what the basic phenomenon is that is driving the belief-revision triviality proofs about conditionals.

2.3.1 Background on the Formal Framework

The basic decision theoretic framework common to Lewis and Collins is the representation of the beliefs of an agent who believes all the logical consequences of his beliefs. We shall follow Lewis and Collins in using a semantic model, which treats a proposition as the set of possible worlds at which it is true. The logic of propositions is that of basic set theory. A belief state K is a set of possible worlds which the agent regards as open possibilities. A proposition A is believed iff it is true at every world compatible with what is believed: $K \subseteq A$. Hence, a belief state is equivalent to the conjunction (intersection) of all propositions believed. The empty set represents an inconsistent belief state, whereas the set containing all worlds in the domain of the model (the necessary proposition) represents a completely unopinionated belief state.[3]

Moving beyond this static picture to a theory of belief dynamics, what shall we say of K_A, the belief state resulting from K by learning that A? For our purposes we need only concern ourselves with the case where what is learned is consistent with what is already believed. Here the rule is simply to conjoin the new information to the old:

$$K_A = K \cap A, \text{ if } K \nsubseteq \neg A.$$

of another article [66] in meta-ethics. In this paper I shall primarily refer to the theorem of Lewis's follow-up piece [65], where the anti-Humean theory is shown not to be closed under conditionalization. As Lewis himself states, that proof is much simpler than the proof he gave in his initial article [64], where a related result was shown using probability kinematics ("Jeffrey conditioning"). Since conditionalization is a special case of probability kinematics, and Lewis [64] did not restrict himself to the other non-degenerate cases, Lewis' subsequent theorem can be regarded as establishing his initial result too. In fact, the proof given below can be adapted to non-degenerate probability kinematics; elsewhere [30] I discuss the related proof for conditionals.

[3]This is the standard framework of statisticians, such as Savage [90], where worlds are called "states" and propositions are "events". See Stalnaker [103] for a defense of this approach. Another method of modeling ideally rational belief states treats them as sets of sentences in some formal language, closed under a suitable rule of deducibility. All of the proofs presented here can be given in the sentential framework.

The condition that the observation becomes believed is simply a nod to giving experience its due. One justification for the rest of the rule is one of informational economy: beliefs should not be gratuitously given up or arrived at. Is this rule a universal requirement of rationality? Even if it is not, all that we need for our purposes is that there are a wide enough range of circumstances under which following the rule is rationally *permissible*. For the incompatibility of a permissible rule of belief change with cognitivism will be enough to cause bad trouble for cognitivism.

Bayesian theory imposes a probability measure P over the qualitative theory we have just outlined. Probability is a quantity which is scaled, non-negative, and additive over disjoint proposition: $P(A \cup \neg A) = 1$, $P(A) \geq 0$, and if $A \cap B = \emptyset$ then $P(A \cup B) = P(A) + P(B)$. So long as $P(A) > 0$, the conditional probability $P(B/A)$ is equal to the ratio of probabilities $P(A \cap B)/P(A)$. The conditionalization of P on A is $P_A(B) = P(B/A)$ for all B. Conditionalization is a refinement of our simple rule for how to change one's beliefs upon leaning information consistent with what is already believed, going beyond the qualitative rule by rescaling probabilities according to the prior ratios.

Why introduce the qualitative model first, if standard decision theory is quantitative? Because it turns out that, at least under an illustrative simplifying assumption, the purely qualitative model suffices to establish Lewis's triviality theorem as a special case of Collins's theorem. This shows that the numerical precision of standard decision theory is not ultimately responsible for the problem which Lewis is pointing to.

Although we shall prove the triviality theorem in a qualitative setting, the precision of numerical probability does have its advantages in formulating a norm for rational decision making, and and in any case we need that norm in order to state Lewis's formulation of the cognitivist thesis. So now we shall return to our brief introduction to quantitative decision theory, by addressing the topic of expected value. In Jeffrey's evidential decision theory, the value of any proposition obeys a law of additivity of the value of its cases weighted by the probability of their occurring given

the proposition:

$$V(A \cup B) = V(A)P(A/(A \cup B)) + V(B)P(B/(A \cup B).$$

For instance, consider the value of the proposition that you go to the beach today. This can come true either by its being sunny or by its being cloudy. The value of your going to the beach is determined by the value of its being sunny weighted by the likelihood of its being sunny given that you go to the beach, plus the value of its being cloudy weighted by the likelihood of that given that you go. Those cases can in turn be divided into further cases.

2.3.2 Desire as Conditional Belief

We now have all the decision theory we need to set out Lewis's argument against anti-Humeanism. In order to simplify the issues as much as possible, Lewis asks us to imagine an agent who does not distinguish multiple degrees of goodness, but treat goodness as a property which a world either has or does not have. We can collect all the world regarded as good into a proposition G. So for a rational agent who meets this simplifying assumption, the value of any proposition A compatible with his beliefs is expressed by this equation which Lewis [65] calls Desire by Necessity:

(DBN) $V(A) = P(G/A)$.

This equation follows from the formula for expected value, our definition of G, and our normalizing that proposition by 1 and its negation by 0.[4]

Lewis calls this Desire By Necessity because he regards it as an expression of the view that there are some things, described by the proposition G, which all rational agents desire. We need not share that interpretation of the equation, since we are

[4]In a more realistic model admitting more degrees of value, a multiplicity of weights would appear, making it clearer that value was not literally identical to a conditional belief, even though it is fixed by conditional belief and the appropriate weighting. For every number g, let G_g be the value-level proposition that is true of exactly those worlds where $V(W) = g$. Since these intrinsic-value propositions form a partition, $V(A) = \sum_g P([G_g]/A)g$. This weighting scheme is used in [64, p. 52] to show that the simplifying assumption of two levels of value is not needed to get the triviality result.

taking the liberty of applying the DBN label to any agent who happens to satisfy it at some stage in his life.

Also, Lewis discusses DBN as an alternative to the target of this proof, rather than a principle which should be accepted in conjunction with it. In Section 5.2, we shall consider DBN as an alternative to Lewis's target (and argue, against Lewis, that is adequate as an anti-Humean thesis so long as evidential value plays the role of desire). But as we show in the section below, DBN is entailed by the assumptions of Lewis's triviality proof, hence we may afford ourselves the luxury of using it in interpreting Lewis's target thesis.

2.3.3 Desire as Belief

Lewis prefers to formulate the cognitivist theory in terms of full belief rather than conditional belief. After all, the intuitive anti-Humean idea was that a desire is accompanied by a belief. What is the object of such a belief? Lewis never gives a very thorough explanation of it, probably because he takes the burden to be on the anti-Humean to characterize it. Lewis treats it as "the proposition X, whatever it may be, such that believing X is somehow necessarily connected" with desiring what is good. [64, p. 44] Functions are cheap, so Lewis christens the "halo function" in order to express what X is:

> To any ordinary proposition A, there corresponds another proposition: $\mathring{A}$, the proposition that it would be good that A. [64, p. 46]

And so Lewis gives us his preferred formulation of the besire theory, Desire as Belief:

> (DAB): V(A) = P(Å).

Again, as with DBN, we part ways with Lewis, who applies the DAB label to the thesis that this equation holds of all rational agents. We use the equation simply to express the equality of value and belief in goodness which holds at some stage of a rational agent's life. (This makes it anachronistic to speak of desire *as* belief, but we retain the terminology for ease of relating our discussion to Lewis's.) This minimal

reading of DAB is sufficient for formulating Lewis's triviality proof, which shows that DAB cannot hold across changes in belief.

Rather than turn immediately to Lewis's own triviality proof against this equation, it is helpful to proceed by considering what sort of proposition is determined by the halo function. Given the simplifying assumption of two levels of value, and having already familiarized ourselves with the proposition G, a natural proposal for us to make about the halo function is that it obeys this

Semantic Constraint: $\mathring{A} \cap A = G \cap A$.

This is not a definition of the halo function, since it does not specify the value of $\mathring{A} \cap \neg A$. It is important that $\mathring{A}$ have truth values at worlds where A is false rather than be a merely partial function, defined only for where A is true. Otherwise it would not suited to be an object of belief (or probability) in a context where $\neg A$ was compatible with belief. Nor would it be suited to make a claim about the objective goodness of A which can be assessed as true or false independently of A's obtaining.

Although Lewis does not explicitly acknowledge the Semantic Constraint in his formulation of his triviality proof, it is implicit in the assumption of Lewis's triviality proof: namely, that DAB is closed under conditionalization. Indeed, Lewis himself proves the lemmas needed to show this, in his discussion of issues related to his triviality proof. As Lewis notes [65, p. 62], the closure of DAB under conditionalization entails Desire as Conditional Belief:

(DACB): V(A) = P(Å/A).

The derivation works because $V(A)$ does not change when A is learned; this "Invariance assumption" is crucial to Jeffrey's conception of expected value.[5] Now, given the Semantic Constraint, the equivalence of DACB and DBN follows from the definition of conditional probability. But Lewis' triviality proof does not explicitly assume

[5]A cautionary note: the triviality theorem about to be presented might tempt us to say that the closure of DAB under conditionalization entails everything whatsoever, hence the derivation of DACB from DAB has less interest than meets the eye. But that temptation should be resisted, because to get an inconsistency we need the further assumption of non-triviality. Trivial belief states can be consistent with DAB, hence the derivation of DACB just presented is not merely the result of inconsistency.

DBN or the Semantic Constraint. However, from the closure of DACB under conditionalization, Lewis establishes that there is a proposition G such that the Semantic Constraint and DBN hold.[6] So, as Lewis emphasizes, nothing is gained by adopting DACB rather than DBN as a positive proposal.[7] More important for our present purposes, we are assured that $P(\mathring{A}) = P(G/A)$ follows from the closure of DAB under conditionalization. This is useful in setting up our proof of Lewis's triviality theorem, as we need have no concerns about using DBN to trivialize the closure of DAB under conditionalization.

We know that DBN is fine in itself, and that it is entailed by DAB, given that the latter is closed under conditionalization. But DAB is closed under conditionalization on pain of triviality; that is what Lewis's theorem shows. To see why this is so, and to finally unite Lewis's quantitative approach with Collins's qualitative approach, let us first draw a few simple consequences of the materials at hand. From DAB and DBN, we have $P(\mathring{A}) = P_A(G)$. Likewise, thanks to our two levels of goodness and scaling convention, we have $P(\neg\mathring{A}) = P_A(\neg G)$.

Since conditionalization is a probabilistic refinement of the purely qualitative belief change rule we stated, we can now proceed to dispense with probability entirely. We say that a probability model induces a qualitative model in the following way: K is the intersection of every proposition A such that $P(A) = 1$. For ease of exposition, we join Lewis in making the simplifying assumption that there are finitely many possible worlds.[8]

[6]The details are found at [65, p. 64]. After establishing a basic property of DACB's closure under conditionalization ("Initial Lemma"), Lewis applies the halo-function to the necessary proposition. He shows that the resulting proposition obeys the Semantic Constraint ("Upward Lemma") and DBN ("Downward Lemma"), hence is a way of picking out what we have called G.

[7]DACB is Price's proposal for explicating the besire theory; he also proposes a reformulated version of Collins' condition: $K \subseteq \mathring{A}$ iff $K_A \subseteq G$, $K \subseteq \neg\mathring{A}$ iff $K_A \subseteq \neg G$. Broome defends the DBN equation as something which should be held in common between Humeans and anti-Humeans. The real dispute, according to Broome, "is over what ultimately determines the goodness of a world. A Humean thinks goodness must ultimately be determined by people's desires; an anti-humean thinks this is not so." [15, p. 265] Of course there is such a metaphysical dispute to be had between Humeans and their opponents. However, we see there is more going on than that in the debate about how to formulate anti-Humeanism within decision theory. Neither Broome nor Price explain the relation between G and $\mathring{A}$.

[8]The proof given will continue to hold as literally stated if we relax the finitude assumption, and consider atomic measures (as long as we assume countable additivity). The proof ceases to hold as literally stated in the case of atomless measures, where probability is not concentrated on singleton-

Putting together the lessons of the last two paragraphs, the assumption of Lewis's proof (that DAB is closed under conditionalization) entails this condition on belief states:

($\star$) if $K \not\subseteq \neg A$ then (i) $K \subseteq \mathring{A}$ if $K_A \subseteq G$ and (ii) $K \subseteq \neg\mathring{A}$ if $K_A \subseteq \neg G$.

This is clearly also entailed by Collins's [19, p. 338] formulation of the desire as belief thesis:

(Qualitative DAB) (i) $K \subseteq \mathring{A}$ iff $K_A \subseteq G$ and (ii) $K \subseteq \neg\mathring{A}$ iff $K_A \subseteq \neg G$.

Collins interprets this as a qualitative theory of expected value for an agent who only distinguished two levels of value: an agent desires A if $K_A \subseteq G$, is averse to A if $K_A \subseteq \neg G$, and is indifferent to A otherwise. The benefit of mapping probabilities into qualitative belief states is precisely that it enabled the derivation of ($\star$) as a condition entailed by the premises of both Lewis's and Collins's proofs.

My condition ($\star$) is weaker than Collins's principle in two ways. First, it only makes use of the right to left clause of the two biconditionals labeled with little roman numerals. This is because it is the minimal condition needed for the triviality proof presented below; it is not intended as a full statement of a thesis of expected value. Second, the initial conditional of ($\star$) restricts us to the case where A is compatible with what is believed, while Collins's principle is not so restricted (whereas in Lewis's equations, the restriction is always in place tacitly, even when not explicitly stated). Collins is working within a framework of belief revision where, unlike standard conditionalization, the result of observing something contradicting one's beliefs leads to a consistent belief state. This is important for his other concerns about conditionals, but it turns out to distract from the result about desire as belief.[9]

set propositions. The problem there is that our definition of K in terms of P will fail when there is an empty intersection of propositions with probability one. However, the idea of the proof can be easily extended to the atomless case. Vann McGee has suggested two ways of reformulating ($\star$) and the theorem. The first (McGee, p.c.): "there is a Boolean algebraic formulation of the theorem, with the Boolean less-than-or-equal-to relation in place of subset, with the same proof, that has the Lewis theorem, in full generality, as an immediate corollary." The second stays closer to the presentation above, but qualifies $K \subseteq A$ as holding "almost everywhere". Either way, we retain the lesson that there is a similar problem in the qualitative and quantitative cases.

[9]Belief-contravening revision is essential for Collins's alternate "Wishful Thinking" proof against desire as belief where (ii) is dropped. Elsewhere (chapter 3 and my [30]) I discuss the related issues

2.3.4 The Triviality of Desire As Belief

DEFINITION. Let us say that a belief state is *non-trivial* iff there is some proposition A such that the agent is unsure whether it is true, and unsure whether G obtains if it is true. A non-trivial belief state K is compatible with each of $A \cap G, A \cap \neg G, \neg A$ (none of their negations are accepted).

TRIVIALITY THEOREM. There is no non-trivial belief state which, along with belief states closed under the rule $K_B = K \cap B$ if $K \cap B \neq \emptyset$, satisfies ($\star$): if $K \not\subseteq \neg A$ then (i) $K \subseteq \mathring{A}$ if $K_A \subseteq G$ and (ii) $K \subseteq \neg\mathring{A}$ if $K_A \subseteq \neg G$.

PROOF. Let K be any non-trivial belief state. From the rule for belief-change we have: $K_{\neg(A\cap\neg G)} = K \cap ((A \cap G) \cup \neg A)$ and $K_{\neg(A\cap G)} = K \cap ((A \cap \neg G) \cup \neg A)$. So we have (1): $K_{\neg(A\cap\neg G)} \cap K_{\neg(A\cap G)} \neq \emptyset$. We also have this about consecutive belief changes (2): $(K_{\neg(A\cap\neg G)})_A \subseteq G$ and $(K_{\neg(A\cap G)})_A \subseteq \neg G$. From (2) and ($\star$) we get (3): $K_{\neg(A\cap\neg G)} \subseteq \mathring{A}$ and $K_{\neg(A\cap G)} \subseteq \neg\mathring{A}$. But (3) is inconsistent with (1). QED.

COROLLARY (COLLINS'S TRIVIALITY RESULT). Qualitative DAB does not hold in a non-trivial belief state.

COROLLARY (LEWIS'S TRIVIALITY RESULT). DAB is not closed under conditionalization in a non-trivial probability assignment (that is, where positive probability is given to each of the three disjoint propositions $A \cap G, A \cap \neg G, \neg A$).

2.3.5 Comparison With Triviality for Conditionals

To get a better grip on what DAB amounts to and why it goes wrong, Lewis gives a suggestion for how to interpret the halo function. Let $\mathring{A}$ be the proposition expressed by $A \rightarrow G$ "in some appropriate sense of the conditional arrow". [64, p. 50] At first glance, there are several advantages to such an interpretation. After all, the expected

with conditionals, including the cases that correspond to having or dropping (ii). Note that there is a trivial notational variation in that Collins uses V to denote the proposition we have called G.

value of A sounds roughly like a claim about goodness obtaining if A does. And Lewis's own gloss on the halo-function as "it would be good that A" sounds roughly like the claim made by a subjunctive conditional.

Additionally, the conditional $A \rightarrow B$ has just the right logical properties to ensure the Semantic Constraint, so long as it obeys centering: $A \cap (A \rightarrow B) = A \cap B$. Moreover, helping ourselves to a standard "similarity of worlds" semantics for the conditional would be a step toward solving the mystery of what proposition the halo-function determines. Finally, given DBN, DAB would fall out as a special case (for G) of a more general constraint on models of a rational agent's mental states having nothing to do with value. Namely, the Conditional Construal of Conditional Probability

(CCCP): $P(A \rightarrow B) = P(B/A)$, if $P(A) > 0$.

To many philosophers, CCCP sounds like it captures the intuitive probability of an indicative conditional.[10] For instance, faced with a shuffled row of an Ace, King and Queen, most will assign a probability of 1/2 to the following conditional:

If the middle card isn't the Ace, its the Queen.

If the middle card weren't the Ace, it would be the Queen.

The second sentence shows that CCCP should also apply to subjunctive conditionals whose antecedents have positive probability–at least to the extent these seems to have the same acceptability conditions as indicative conditionals, a topic we will return to (in the section on "Causal Decision Theory With Subjunctive Conditionals").

However, even more philosophers (including some who would like to believe in it) are persuaded to reject CCCP thanks to Lewis' triviality result [62] showing that the

[10]I borrow this name from Hajek and Hall, who provide a thorough although primarily negative discussion of the tenability of the CCCP equation. Stalnaker [101] advocated the equation as a probabilistic interpretation of his conditional. Lewis [62] invokes conversational implicatures to attempt to reconcile the plausibility of the equation with the view that the indicative conditional is the material conditional (which certainly does not obey the equation except in special cases); in the postscript he endorses a different tactic using conventional implicatures, due to Jackson. The doctrine is most famously associated with Adams [2], who denies that the conditional has truth conditions (or straightforwardly embeds in larger sentences); his position is loosely parallel to meta-ethical emotivism.

equation is not closed under conditionalization. The lesson of the triviality result is often taken to be that indicative conditionals do not express conditional propositions, but rather express conditional beliefs (conditional probabilities, in a Bayesian model). Subjunctive conditionals, on the other hand, do express propositions.

Given this, it turns out that it would be unreasonable to interpret $\mathring{A}$ by "the dreaded 'probability conditional', a supposed connective which makes probabilities of conditionals equal the corresponding probabilities" [64, p. 50]. But there is no other way of interpreting the halo function that will evade the problem. For Lewis gives a direct proof against DAB making no explicit reference to conditionals, which is a variation on a technique of proving his earlier triviality proof about conditionals. Our own triviality proof (above) against DAB is easily adapted to show a similar result about conditionals.[11]

2.3.6 Diagnosis: Conditionals and Negation

Let us take a step back, to highlight what does the dirty work in the triviality arguments. As with the reduction of Lewis's DAB thesis to a qualitative counterpart shared with Collins, we can illustrate some the main phenomena about conditionals without reference to probability. The leading idea motivation, as well as the Achilles' heel, for the theory of non-material conditionals lies in their interaction with negation.

Bob is unsure whether the Red Sox will make it to the World Series. But as he holds on to his belief in "the curse of the Bambino", he is sure that they will not win the Series. In virtue of these beliefs, he denies the indicative conditional:

[11] One simple way to get at the result about conditionals: in the statement above of ($\star$) and the theorem following it, replace the special proposition G with arbitrary B, and $\mathring{A}$ with $A \rightarrow B$. In my paper forthcoming in *Philosophy of Science* I discuss this and related theorems in connection with several other prominent results about conditionals, including theorems of Gardenfors which Collins places at the center of his paper.

Theorems about CCCP and DAB can often be adapted to apply to the other. The well known original proof technique of Lewis' famous triviality result [62] for conditionals can be adapted to show that DAB implies that an agent is indifferent to every proposition: $P(G/A) = P(A) = P(\neg G/A)$. Going the other way, Lewis's [64] first proof against DAB can be adapted to show that there is almost no way to simultaneously change your mind by probability kinematics about a conditional and its antecedent so that CCCP holds (Lewis [63] already established that CCCP is not closed under non-degenerate probability kinematics).

If the Red Sox make it to the World Series, they will win.

This shows that the indicative conditional does not have the truth conditions of the material conditional. For the material conditional (corresponding to the example sentence) is truth functionally equivalent to:

Either the Red Sox won't make it to the World Series, or they will win it.

But Bob does not deny that disjunction, since he is unsure whether it is true; denying it would commit Bob to affirming that the Sox will make it to the Series, which he is unsure of. And given his keen linguistic and logical sense, Bob does not deny a sentence of relatively simple syntactic complexity, when it is truth conditionally equivalent to another sentence of similar syntactic complexity which he does not deny.

Furthermore, Bob affirms "If the Red Sox make it to the World Series, they will not win". This example shows that the indicative conditional behaves as if negating the conditional is equivalent to negating its consequent. (At least it does so if we accept that like speech acts involving ordinary non-conditional sentences, denying a conditional is equivalent to accepting its negation.) The material conditional does not obey that principle, since it is true whenever its antecedent is false.[12]

Although the indicative conditional behaves as if it has a logic stronger than the material conditional, there is a worry that it does not have truth conditions of any sort (not even intensional conditions going beyond the extensional conditions of truth functional logic), and hence does not have a logic in the conventional sense. Rather than state a proposition which has truth conditions, the indicative conditional expresses a *conditional belief*: a disposition to infer the consequent upon learning the antecedent. For instance, Bob is disposed to believe that the Red Sox will not win,

[12]This is adapted from Stalnaker's example favoring the Yankees in 1987. [103, Ch. 6] (My fictional character Bob does not necessarily represent the current views of Stalnaker, who resided in upstate New York instead of just outside Boston when he penned his example.) Some notes for conditionals aficionados. First, on the logic: both Stalnaker and Lewis agree that the negation may be moved from the consequent to take scope over the conditional (a full statement of conditional non-contradiction must make an exception for impossible antecedents); Stalnaker adds that negations move in from the outside to the consequent. Second, on the applicability of the logic: Stalnaker maintains that this is the correct logic of the indicative conditional as well as the subjunctive, while Lewis holds that the indicative is the material conditional.

upon learning that they have made it to the World Series.[13] A powerful argument for this theory of indicative conditionals can be made from the following story.

Allan, Zack and Jack are in the audience while Pete performs a card trick. Pete shows them an Ace, King and Queen. Then he shuffles them, and places the three cards in a row on the table. Zack gets to peek at the leftmost card and sees the King, while Jack looks at the rightmost card and sees the Queen. Zack whispers to Allan, "If the middle card isn't the Ace, its the Queen." Jack whispers to Allan, "If the middle card isn't the Ace, its not the Queen." Allan concludes, rightly, that the middle card is the Ace.

The puzzle is that both Zack and Jack are basing their reports entirely on correct observations, so if either of them says something true then so does the other. And Allan accepts what each of Zack and Jack says, and arrives at a true belief. But the conditionals appear to contradict each other, and our earlier analysis of the logic of the conditional says that they do indeed contradict each other. Something has to give.

If the indicative conditional were the material conditional, then Allan's conclusion would make sense as the acceptance of both the assertions by Zack and Jack. But we have already seen that analysis of the indicative conditional is inadequate. So, it is argued, conditionals do not express propositions at all, not even ones with (intensional) truth conditions stronger than the material conditional. Instead, Zack and Jack each express their conditional beliefs, their belief revision policies. Allan reaches his conclusion not by accepting propositions which are semantically encoded in the uttered conditionals, but rather by reasoning to the best explanation about what observations Zack and Jack would have made in order to rationally give rise to those conditional beliefs.[14]

What is more, note that the proof we gave, when adapted to conditionals, has an importantly similar structure to the example involving Allan, Zack, Jack and Pete. The main difference is that the proof does not consider the belief states of two different

[13]This "inference ticket" view of the indicative conditional can be given its own "probability logic" in the manner of Adams [2].

[14]This example is inspired by a more elaborate story by Gibbard [34].

agents following their separate observations. Rather, it considers a single agent prior to making either of two possible observations, and asks what would happen if he were to make either of the observations. In comparison to that example, we can put the point like this: before peeking, Zack's beliefs left open that he might have ended up with either the observation he turns out to make, or the observation which it turns out Jack makes. The fact that the interpersonal and intrapersonal cases parallel each other reinforces the conviction that we are dealing with a unified phenomenon in discussing triviality.

Eventually we will consider a way out of this example (Section 4.4). But let us accept for now the thesis that indicative conditionals do not have truth conditions, but instead express conditional beliefs. And so conditional beliefs are not reducible to a belief in a conditional.

Regardless of this position about indicative conditionals, we have good reason to accept that subjunctive conditionals express propositions, and that these are stronger than the material conditional. The most famous cases which make this point are counterfactual conditionals, conditionals in the subjunctive mood whose antecedents are believed to be false. "If this match were struck it would have lit" makes a true statement about a normal, unstruck match. But this cannot be analyzed as a material conditional, since that would also confer truth upon "If this match were struck it would not have lit", making a mockery of our folk science of matches.

Now, the subjunctive conditionals which we shall make use of are not counterfactual. Instead, we shall consider subjunctive conditionals whose antecedents are not believed to be false. More specifically, we shall consider decision making conditionals, where in the case of an open question about a decision, it is an open question whether or not the antecedent will turn out to be true. The interpretation we shall adopt is that these subjunctive conditionals have the same semantic analysis as the counterfactual ones. Also, we shall always treat subjunctive conditionals as having a causal interpretation where the antecedent states a cause and the consequent states its effect.[15]

[15]This causal reading contrasts with the "back-tracking" readings, such as "If Gore had won the

2.4 Attempts to Save Cognitivism from the Triviality Theorem

To repeat the lesson of theorem: Lewis's DAB equation and Collins's qualitative version are inherently unstable under a permissible, and arguably mandatory, rule of belief change. So if we accept either equation as an explication of how somebody desires something to the extent they believe it to be good (in the simplifying case of two levels of goodness), then the explicated doctrine is untenable. This is so whether the connection between desire and belief is necessary, as in the internalist theory Lewis focuses on; or merely contingent, as in the case of the saint as treated by the externalist. So given the explication, all forms of cognitivism about value are unacceptable.

Now we consider a variety of responses to the triviality proof, made on behalf of cognitivism. In 4.1, I reject the response that denies the plausibility of the model of decision theory. In 4.2, I endorse the response which defends the explanation of evidential value as a conditional belief about goodness. In 4.3, I reject the response which attempts to block the triviality proof by an indexical theory of goodness. In 4.4, I give mixed marks to the response which circumvents the triviality result for evidential value by an indexical theory of indicative conditionals.

The upshot of these sections if that if one simply wants to be a cognitivist without requiring that desires correspond with beliefs, a sensible route is to express them by conditional beliefs. But if one further wishes the connection with beliefs, as in a literal treatment of the "besire" view, one needs to make a new move in the game. That new move is the subject of section 5, where we turn from evidential to causal decision theory.

election, the Supreme Court would not have had a conservative bias." See Lewis [60] for discussion of causality and time order in conditionals.

2.4.1 Qualms About the Decision Theoretic Idealizations

As an initial attempt to dispel for the force of the triviality theorem, the cognitivist might be tempted to claim that that their theory is about rational persons, not about some abstract mathematical idealization given by decision theory. Hence, we should not leap from conclusions about the latter to conclusions about the former. First, they might be tempted to balk at the computational idealizations of decision theory. Second, they might point out discrepancies between expected value and the ordinary notion of desire.[16] I think that the interest of the triviality theorem withstands these scruples.

Deductive omniscience and quantitative precision are obviously strong idealizations. However, for our present purposes I shall just say this: it would be undesirable for the defense of the anti-Humean position to depend on limitations on an agent's inferential capacities. This is especially so if cognitivism is intended to buttress normative inquiry as a rational enterprise. Moreover, even if we adopted some more realistic picture of "bounded rationality", it should be possible to recapture the force of the proof, given how simple the basic qualitative version is.

Jeffrey admits that "desirability" (his terms for evidential value) does not directly correspond to any single pre-theoretical notion of "desire". Instead, it provides the best systematic explication of the decision theoretic idea, which is itself our best effort to make precise the intuitive idea of weighing options. One of the most notable quirks of desirability is that it attaches to propositions you are certain of, whereas the ordinary notion of desire is something that goes away once its object is known to have been achieved. One might wonder whether this makes Jeffrey values unsuitable for Lewis's purposes of finding a surrogate for ordinary desire. I don't think that the discrepancy is fatal for Lewis's project, since Jeffrey values will still be largely

[16]On the first point, Byrne and Hajek protest [17, p. 424-5] that our standards of rationality are less strict than what Lewis employs in his proof, so perhaps his result only applies to "hyper-idealized" Bayesian "Übermenschen" but not to "all-too-human agents". They also raise the second point, and are joined in this by Weintraub [108]. Considering departures from even a non-idealized picture of rationality, Wallace [107] claims that irrational action is relevant to evading Lewis's result, by showing that beliefs and desires need not coincide; this is on the wrong track, because the problem remains of showing how beliefs and desires can coincide in an agent who is rational.

determined by desires in the ordinary sense, and the triviality of the former spell trouble for the latter in the cases where the discrepancy is less marked. Especially, whenever one desires that some proposition be true which one is unsure of, then it will have positive news value; this condition is met by the relevant propositions in the definition of a non-trivial belief state.

2.4.2 Abandoning Belief for Conditional Belief

Lewis's main reason for being dissatisfied with DBN (and DACB) is that we have not been presented with a belief which is the state of valuing A. Such a belief would especially be worth identifying if we were trying to cash out the "besire" theory, the strong internalist doctrine that a desire just is a belief. But the anti-Humean can relax the demand that a desire be connected with a belief (whether through being identical to that state, or however else).

What is needed is that the desire be connected with some state which is cognitive, as opposed to emotive. The distinction is sometimes explained in terms of different "directions of fit": cognitive states aim to fit the world, while emotive states are such as to guide the agent to make the world fit them. And conditional beliefs seem to be characterized by the cognitive direction of fit. So desire can be identified with a conditional belief, a disposition to change belief which is determined by a ratio of probabilities.[17]

We can even do better than this. In this paper, we have been exclusively concerned with the case of conditional beliefs where the observation—the antecedent of the indicative conditional—is compatible with prior belief. And the rules of belief change we have considered for that case have the following property. The posterior belief state induced by making the observation is determined by the prior *overall* belief state in combination with the observed proposition. In the qualitative case, we simply added

[17]This sort of defense of conditional beliefs is given by Broome [15] and (more emphatically) by Price [83], who advocate the equations we have called DBN and DACB, respectively. The direction of fit metaphor is introduced by Anscombe [6] with this example. A shopper goes around the store picking up the items on a shopping list. The store detective follows around, jotting down every item taken. In the end they have the same lists, but the shopper's represents desire while the detective's represents belief.

the new information to the old by conjunction. In the quantitative case, we took the further step of rescaling the partial beliefs (probabilities) by their old ratios.[18]

As far as cognitivism goes, it is at first difficult to see why the position that desires are determined by beliefs (or conditional beliefs) should be less acceptable than the one Lewis focuses on, where a desire is to be reduced to a single belief. This is so whether the cognitivism is internalist or externalist. However, there is still this to be said for maintaining something with the form of DAB, with unconditional beliefs instead of conditional beliefs. We sometimes discuss judgments of values as if they expressed beliefs, and it would be nice to have an explanation of this fact; it would be unsatisfying to claim that it is a complete illusion. Next we consider two attempts to account for the appearance of value judgments expressing beliefs. The first does not work, but the second has plausibility.

2.4.3 The Relativity of Value

It is a tacit assumption of Lewis's triviality proof (and ours) that the halo function receives the same semantic interpretation across different probabilities. For rules of belief change properly apply to propositions, or at least to sentences whose meanings (whatever they might be) are invariant across contexts. So the proofs would rest on a fallacy of equivocation if the interpretation of the halo function—the proposition expressed, in our adopted framework—is a function varying with the probability assignment. Two things are needed to make this strategy plausible. First, an independent reason for thinking that the proposition expressed is relative to belief assignments, so that the move is not regarded as an *ad hoc* bit of wishful thinking by the philosophical theorist. Second, an assurance that the relativism really delivers the goods.

There is a proposal which has been taken up in the literature for adopting a

[18]In the case of belief-contravening revision, where what is observed contradicts prior beliefs, we cannot treat conditional beliefs as determined by overall belief states. This is the lesson of the triviality results surrounding conditionals, which relate to the triviality of Collins's qualitative DAB, which allows for belief contravening revision. So if we thought we needed a notion of expected value adapted to the belief-contravening case, then the defense of cognitivism would rest simply on the claim that such disposition have the right direction of fit.

contextualist solution, where the proposition expressed by the halo function changes with person's overall belief state, and in just the right way so as to avoid the Lewis result. The idea is that we have independent reason to think that value judgment are relative. Hence we already have the right sort of distinction at hand in order to dodge triviality.[19]

Unfortunately, even if values are relative, they are not relative in the way needed to dodge the triviality proof. The independently plausible way that value is relative is that different agents have different conceptions of which worlds G (aka $A \cap \mathring{A}$) is true at. To take a toy example: you think a world with all vanilla ice cream is good, I think the all chocolate world is good. But this kind of relativity of goodness is not what is required to maintain the DAB equation.

Notice that there is no reason to think that the propositions expressed by A or G are changing, when the agent changes his probabilities by conditionalization. For the indexicalist change needed to evade the Lewis result is that the proposition expressed by Å take on different truth values at certain $\neg A$ worlds when an agent's beliefs change.[20] But there is no reason to think that changing one's probabilities should change one's mind in this way. Hence, there is no reason to claim that the relativity of goodness leads to the sort of indexicality needed to maintain DAB in the face of the triviality result.

2.4.4 The Relativity of Indicative Conditionals

The suggestion of treating the halo functional as a conditional suggests a few ways to save desire as belief from triviality. There are independently motivated suggestions about conditionals, made in response to Lewis' original triviality theorem for

[19]This strategy is mentioned by Byrne and Hajek [17] and explored more fully by Hajek and Pettit [40]. The latter article cites Dreier [28] for further support for an indexicalist cognitivism (formulated without any consideration of DAB). If we endorsed the full details of Dreier's story then we would have an even better connection with besires. For Dreier claims that by analogy to the "problem of the essential indexical", the indexicality of value judgments explains how they obey the internalist constraint of being action-guiding. I hope to take up elsewhere the nature of the connection between motivation and indexical judgments, which I think is rather different.

[20]In the next section, we consider analyzing the halo function by the Lewis-Stalnaker conditional semantics. This allows us to put the point here this way: the agent is changing his conception of which $\neg A$ worlds are closer to which A worlds.

conditionals [62], which might preserve a form of the cognitivist theory close to the doomed DAB. The first is the proposal that CCCP might be maintained by a deeply contextualist theory of what proposition is expressed by indicative conditionals. The second is the idea of causal decision theory expressed in terms of subjunctive conditionals, whose probabilities diverge from CCCP. We consider the first strategy in this section, and the second strategy in the next section.

According to the contextualist proposal for indicative conditionals, although CCCP is maintained for uses of conditional *sentences*, there is no *proposition* whose probability generally equals the corresponding conditional probability. Lewis objects to this strategy: "But presumably our indicative conditional has a fixed interpretation, the same for speakers with different beliefs, and for one speaker before and after a change in his beliefs." [62, p. 81] But on the contrary, there is plenty of reason to think the conditional is highly context sensitive anyway. It does not have a fixed interpretation (similarity relation or selection function, in Lewis-Stalnaker semantics) across different occasions of use, even though it has a fixed logic constraining what interpretations are admissible.

Once this is granted, then Lewis's objection to the contextualist strategy becomes this: what appears to be a disagreement or change of belief about a conditional may turn out to be, in part or in whole, a verbal dispute rather than a factual one. To give the point an air of paradox: I may accept the proposition you communicate by a conditional even though I affirm the contrary conditional sentence! Furthermore, our account of communication with conditionals will not fit smoothly into a simple picture of communication as the sharing of thoughts expressed by what is said. Communication will be indirect in that in order to determine which proposition you are expressing, I will have to make assumptions about your conditional probabilities—assumptions about your beliefs which will rationalize what you say even though what you say does not directly express those beliefs. The advantage of paying these costs is that we respect the surface phenomena about probabilities of indicative conditionals while maintaining uniformity with the logic of subjunctive conditionals.[21]

[21]The points about communication, emphasized by Gibbard [34] against the propositional theory

So if on the whole it is found worth preserving CCCP by appeal to contextualism, and if it is plausible to interpret the halo-function as equivalent to an indicative conditional, then something close to DAB is saved. However, a further cost must be payed, beyond those already facing the contextualist theory of conditionals: the meta-ethical woes of relativism. The current proposal picks up those problems, and compounds them with the problem of relativity of normative vocabulary for different stages in the history of a single agent's beliefs.

2.5 Cognitivism Saved by Causal Conditionals

Finally, we turn to a more promising way in which cognitivism can be given a coherent formulation. The first—but by no mean the last—move in the strategy is to turn away from evidential to causal decision theory. The intuitive background is outlined in section 5.1, while the relevant formal details are presented in 5.2.

On the face of it, there are several reasons why this is a surprising direction to take the effort to save cognitivism from the force of the DAB triviality theorem. After all, Lewis is himself a causal decision theorist, so it is surprising that he should not notice that he already has the main ingredient to save cognitivism. Indeed, Lewis twice considers whether causal decision theory might be relevant to the meta-ethical problem, and argues against it. Finally, it would not be surprising if Lewis were right that there is no relevance, since the issue of cognitivism and the causal puzzles of Newcomb problems seem to quite different issues. Nevertheless, causal decision theory does provide the tools to establish a connection between desires and beliefs about goodness, which does not fall prey to the triviality proof.

There are three crucial moves which need to be made, and a fourth more ambitious

of indicative conditionals, are discussed by Stalnaker [103, pp. 108-111]. In assessing contextualism, we should also consider some distantly related views. One position rejects the whole propositional framework in favor of time-relative ("tensed") information states, and presents special belief change rules for these. Oddie adopts this temporalist framework in his treatment of desire as belief, see Lewis's [65, fn. 62] response. Another position sometimes advocated for the CCCP is that the way to dodge triviality is to adopt non-standard rules for belief change with respect to conditionals (as opposed to non-conditional sentences which receive standard treatment). In chapter 3 I discuss this move, and further discusses the plausibility of contextualism, especially in relation to conditionals embedded in larger sentences.

move which might be made, to secure cognitivism via causal decision theory. The first move (in 5.3) is to explain how to derive the halo function from the standard formulation of causal decision theory. The second move (in 5.4) is to explain how Newcomb problems are more pervasive than has been previously recognized, and that Lewis's triviality proof is in fact a recipe for generating Newcomb problems. The third move (in 5.5) is to argue that, against both Lewis and some of his opponents, there are not one but two notions of "desire" relevant to decision theory. At this point the modest cognitivist may rest content, having been given a way to connect desires and beliefs about goodness within decision theory. The more ambitious cognitivist will then attempt the fourth move, to argue (in 5.6) that it is causal utility and not evidential news value which explicates the sense of desire which is relevant to the classic meta-ethical debate over cognitivism.

2.5.1 From Evidential to Causal Decision Theory

When faced with a choice of actions, one considers the likelihood of their effects given how cooperative the outside world turns out to be, and weighs the pros and cons of those potential effects. The conventional Bayesian decision theory of statisticians and economists, notably Savage's [90], treats background states of the world as propositions ("events"), while treating acts as functions from them into outcomes (objects, treated along the lines of "commodities" in standard economic theory). Utilities attach to outcomes, while expected utilities attach to acts. The operational interpretation of expected utility is that it measures the desirability of acts given how the agent thinks they might influence the world to his advantage.

Jeffrey [50] proposed a theory which takes all of the objects of deliberation (background state of the world, act, outcome) to be described by propositions. The rough idea of the evidential value of a proposition is that it is how glad you would be to learn that it is true. It is a *passive* conception of the value of receiving good or bad news. However, it is part of the subject matter of propositions that some of them describe events which it is under an agent's control to affect, while others do not. In the case of propositions that are under your control, you can make the news. As

a theory of decision making, the evidential theory says that out of the propositions describing events under your control, chose the one with highest news value.

Despite its elegance, evidential theory delivers incorrect advise in cases where statistical correlation does not represent causation. Causal decision theory aims to remedy this difficulty, as is illustrated in cases known as Newcomb's Problem.[22] Here is an example showing the divergence between the theories.

The eminent statistician Ronald Fisher once considered explaining the high statistical correlation between smoking and cancer by the hypothesis that cancer is not caused by smoking, but rather by genetic factors that predispose people to both a lifetime of smoking and to cancer. This is surely false, but imagine what would happen if it were true, or simply if somebody believed it were true.

The agent deliberating about whether to quit would have reason to give in to the fatalistic argument that there is no point in quitting. Quitting would give him evidence that he will not get cancer, by giving him evidence that he does not have the gene which causes both lifetime smoking and cancer. Not quitting would give him evidence that his genetics have predestined him for cancer. But nothing he does will make a causal difference to his getting cancer or not. Although this is intuitively the rational decision for the agent who accepts Fisher's hypothesis, it is not the answer given by evidential decision theory.

Consider how the smoker (who accepts Fisher's scenario) will compute the evidential news values of whether to quit. The news value of quitting smoking will be high, since quitting will give him evidence that he will lead a long life. The news value of continuing smoking will be low, since it will give him evidence that he will lead a short life. So if the choice worthiness of actions goes by news value, then the smoker should quit. While this might strike us sound advice, this is because we do not accept Fisher's hypothesis. So the lesson is that the agent who accepts Fisher's hypothesis

[22]The original example by Newcomb involves an ingenious story about choosing a prize from an agent with dramatic powers of predicting your choice. The lesson I wish to draw from that example is made by the more realistic example below, borrowed from Jeffrey [50]. It is beyond the scope of this paper to defend causal theory from objections, such as the layers Jeffrey adds to his theory of decision making to avoid adopting the causal theory on the basis of the intuitive force of Newcomb's Problem.

should regard the newsworthiness of quitting as irrelevant to settling what it would be rational for him to do. Hence, we should abandon evidential decision theory for causal decision theory, as a criterion of rational action.

We know from our earlier discussion that we cannot reduce conditional beliefs as expressed by indicative conditionals to beliefs in conditionals as expressed by subjunctive conditionals. That distinction rears its head in this example. From our agent's perspective, each of the following indicative conditionals is highly probable:

If I quit smoking, I won't get cancer.

If I keep smoking, I will get cancer.

Whereas the following subjunctive conditionals will be less probable for the agent, indeed they will be as probable as the hypothesis that he has the unfortunate gene predisposing him to both smoking and cancer:

If I were to quit smoking, I wouldn't get cancer.

If I were to keep smoking, I would get cancer.

This difference between indicative and subjunctive conditionals illustrates the difference between evidential value and causal utility.

2.5.2 Causal Decision Theory With Subjunctive Conditionals

Causal decision theory aims to overcome the difficulty which Newcomb's Problem presents to using Jeffrey's value theory as a decision theory. Following the suggestion of Stalnaker, [100] Gibbard and Harper [35] formulated a decision theory employs a subjunctive conditional intended to express causal connections (they express forward-looking metaphysical dependencies, not back-tracking or epistemic connections). As Lewis [59] casts the theory, the causal utility of some proposition A is given by this equation:

$$U(A) = \sum_i P(A \rightarrow S_i)V(A \wedge S_i).$$

The intended interpretation of this is that A is a proposition which stands for a potential action; it is a cell in the narrowest partition of actions available to the agent. The S_i are cells in some appropriate partition of the world; what makes them appropriate is a subject we shall return to momentarily. The conjunction $A \wedge S_i$ is the outcome resulting from performing that act relative to that state. Note that by centering for the conditional, $A \wedge S_i = A \wedge (A \rightarrow S_i)$. Also note that the value of the outcome is calculated as its Jeffrey news value V, while the sum of these weighted by the probabilities of the conditionals is equal to causal U.[23]

When do V-maximizing and U-maximizing coincide for act A? The Jeffrey value of A over a partition S is

$$V(A) = \sum_i [V(A \wedge S_i)P(A \wedge S_i/A)] = \sum_i [P(S_i/A)V(A \wedge S_i)].$$

So $V(A) = U(A)$ when

$$\sum_i P(S_i/A)V(A \wedge S_i) = \sum_i P(A \rightarrow S_i)V(A \wedge S_i).$$

That is, when

$$P(S/A) = P(A \rightarrow S)$$

for all states S. By what was shown earlier, this happens when

$$P((A \rightarrow S)/A) = P(A \rightarrow S).$$

So $V(A)$ and $U(A)$ coincide when the decision making conditional is probabilistically independent of the act, with respect to each state. In Newcomb cases this does not hold: the conditional $A \rightarrow S$ which is probabilistically dependent on A. So the act A gives evidence for (or against) the existence of a background causal factor (expressed

[23]In the original Gibbard-Harper formulation, the outcome is specified in functional form as $O(A, S)$. For both V and U maximization, Gibbard and Harper express the utility of outcomes in terms of its desirability $D(O)$. In other words, they seem to agree with Lewis that Jeffrey's "desirability" plays a role in both theories, which explains why U and V sometimes agree given the right causal beliefs.

by the conditional) tending toward S.

Although it is not generally true that V and U coincide globally, they will coincide locally with respect to what Lewis calls a "dependency hypothesis". This is a conjunction of, for every A and S: exactly one of $A \rightarrow S$ or $A \rightarrow \neg S$, and exactly one of $\neg A \rightarrow S$ or $\neg A \rightarrow \neg S$. It is the dependency hypotheses taken together, and not the particular conditionals taken separately, which express the agent's causal picture of the world. With respect to any dependency hypothesis, $P(A \rightarrow S)$ is either zero or one, so $P((A \rightarrow S)/A)$ takes the same value (so long as $A \wedge (A \rightarrow S)$ has positive probability). Because the Gibbard-Harper decision making conditional obeys the Stalnaker principle that $(A \rightarrow S) \vee (A \rightarrow \neg S)$, the dependency hypotheses form a partition.[24]

Now let us return to the question of what an appropriate partition S is. According to Lewis, it is any partition such that for no $V(A \wedge S_i)$ would the value be changed by adding information from a dependency hypothesis. (Lewis calls this a "rich" partition.) Lewis singles out two such partitions as notable. One is the partition of point propositions (sets of single possible worlds). The other is the partition of value level propositions, which collect together the point propositions by the value they would be assigned if they were conditionalized upon. That is, if W_i are the point propositions such that $V(W) = g$ for some fixed g, the value level proposition $G_g = \cup_i W_i$.[25]

[24]Lewis uses dependency hypotheses to show how this theory is equivalent to other formulations of causal decision theory invoking partitions of causal factors K outside of the agent's control, which Lewis argues are equivalent to "dependency hypotheses". Lewis argues that conditional excluded middle is only acceptable for the decision making conditional by making the consequents statements about chances (a complication we set aside here). In any case, given centering and CEM, the conditional has a selection function semantics requires that for possible antecedents A, $A \rightarrow B$ is true at world w if B is true at the world $f(w)$ where A is true (which is w itself if A is true there). Gibbard and Harper do not require their logic to obey the full Stalnaker logic/semantics, which adds the principle of weakened transitivity; without that principle, the selection function semantics for the "weak Stalnaker logic" does not determine a notion of "similarity of worlds" (except that every world is closest to itself).

[25]The notation which Lewis uses in [59] for the value level propositions and their weights is $[V = v]$ and v. Here I use the notation from [65] for expressing the same idea, as it occurs in Lewis's discussion of desire as belief rather than causal decision theory.

2.5.3 From Causal Decision Theory to Desire as Belief

With the apparatus of Lewis's formulation of causal decision theory, we may now formulate a rival candidate to the DAB equation. This equation is better suited to play the role of desire as belief, not only in terms of dodging the triviality results, but in terms of being a better motivated explication of the pre-theoretical notions that DAB was offered to cash out. In this section we simply show how, given an agent who satisfies the demands of causal decision theory, one may define a halo function for that agent such that beliefs about goodness correspond to causal utilities. In the next sections, we argue for two claims that show the pay-off of this maneuver. First, this provides both a formal solution to the triviality problem. Second, it provides a substantive solution to the meta-ethical problem of finding a belief which corresponds with desires which form the springs of action.

In the toy model Lewis uses to study desire as belief, we denote the two cells in the value level partition as $\{G, \neg G\}$, and scale their weights to $\{1, 0\}$. Taking that value level partition as S in the equation for causal utility, we have

$$U(A) = P(A \rightarrow G)V(A \wedge G) + P(A \rightarrow \neg G)V(A \wedge \neg G).$$

Plugging in the weights gives us

$$U(A) = P(A \rightarrow G)(1) + P(A \rightarrow \neg G)V(0) = P(A \rightarrow G).$$

Now—following the suggestion of Lewis mentioned earlier—simply define the halo function for this simple agent as $\mathring{A} = A \rightarrow G$.[26] We should say a bit about what $\mathring{A}$ means when defined in terms of the subjunctive conditional of causal decision theory. In the special case of $A = G$, what we have is simply the expression of the intrinsic goodness of G. In the case where $A \neq G$, what is expressed is that A is instrumentally good, in the sense of causally conducing to G.

Substituting into our previous equation, this gives us our causalized version of

[26] In terms of the selection function semantics for the conditional, for possible A let $\mathring{A}$ be true at w iff G is true at $f(w, A)$.

desire as belief:

$$U(A) = P(\mathring{A}).$$

Thanks to the scaling conventions for the simple model, we can express causal utility as being literally the probability of a proposition about goodness. In a more realistic model with more than two value level propositions, we define $\mathring{A}_g = A \rightarrow G_g$, and get

$$U(A) = \sum_g P(\mathring{A}_g)g.$$

Here causal utilities are not exactly beliefs (unconditional probabilities), but they are value-weighted sums of beliefs. What needs to be shown next is that this solves the triviality problem, and that U is the appropriate measure of desire in the meta-ethics debate.[27]

2.5.4 A Newcomb Problem Around Every Corner

From the general facts about causal decision theory, applied to our toy model of two levels of goodness, we have that V and U coincide when

$$P((A \rightarrow G)/A) = P(A \rightarrow G).$$

A parallel phenomenon underlies the triviality result for Desire as Belief. Lewis [65, p. 62] shows that

(IND): $P(\mathring{A}/A) = P(\mathring{A})$

is equivalent to DAB in the presence of DACB (and DACB follows from the closure of DAB under conditionalization). Hence IND can be regarded as being responsible

[27]This tactic is one of the anti-Humean responses suggested by Byrne and Hajek [17], who cite Collins [20] as advocating this in his unpublished dissertation. Mentioning our equation in a footnote, Byrne and Hajek prefer the formulation $U(A) = P(A \rightarrow \mathring{A})$, which yields the same results by the logics of the conditional and halo-function. (Byrne and Hajek do not single out this solution as the best response among the menu of responses they propose, which include the responses I argued against in Sections 4.1 and 4.3 above. They also discount the adequacy of the conditional belief response I discuss in Section 4.2.) The tactic is endorsed by Oddie, [77] who prefers a formulation of causal decision theory bypassing the subjunctive conditionals in favor of dependency hypotheses K; objective chances and values also figure in his theory.

for the failure of DAB to be closed under conditionalization. Put more bluntly, the divergence between $P(\mathring{A})$ and $P(G/A)$ has a similar structure to a Newcomb Problem[28] This suggests that there is an important connection between desire as belief and causal decision theory.

A Newcomb problem has the following three features. First, there is the aforementioned failure of a subjunctive conditional to be independent of its antecedent. Second, it is a decision making conditional, whose antecedent describes a possible action. Third, the reason that the conditional depends on its antecedent is the agent's belief in a background common cause of his action and its outcome.

Although a proponent of causal decision theory, Lewis denies that it is relevant for solving the desire as belief problem. In his first article [64] on desire as belief he attempts to dodge the issue by stipulating that the case under consideration is not a Newcomb problem. That is, he is stipulating that the agent does not have any beliefs about their being a background common cause of A and its potential outcomes. So we are invited to imagine that for the A in question, it happens that $V(A) = U(A)$. If we grant that Lewis can indeed stipulate that the situation is not a Newcomb problem, then how could the divergence between V and U in Newcomb problems be of any relevance here?

Lewis's attempt to dodge Newcomb worries is reinforced by a more general anxiety that causality was never mentioned in Lewis's original postulation of the halo-function, nor does it play any explicit role in the triviality proof. A similar worry affects the triviality results for conditionals, and their relation to causal decision theory. Although Stalnaker received his inspiration for the causal decision theory equation from the original Lewis triviality theorem for probabilities of conditionals, the latter theorem makes no mention of causal connections. The fact underlying the Lewis triviality results is that there is no proposition X such that it is generally the

[28]The point about independence from the antecedent for conditionals is often mentioned in the causal decision theory literature, for instance by Gibbard and Harper [35]. Lewis notes the point about IND for DAB, stressing it in the proof in [65]. The connection between DAB and Newcomb's problem is emphasized by Oddie, [77] although he does not deal with the question of why Lewis cannot stipulate away Newcomb problems, or how this diagnosis can bear on the the apparently non-causal features of the halo-function.

case that $P(X) = P(B/A)$. It is not necessary for the proof of the Lewis results that X be expressed by a conditional connective or halo-function, or that it have a causal interpretation.

The answer to Lewis's argument is that although he can stipulate that the initial situation is not a Newcomb problem, he cannot *stipulate* that the situation will *remain* an entirely non-Newcomb problem, even after the agent changes his beliefs. Indeed, the triviality theorem generates a sort of Newcomb problem out of a non-Newcomb case by means of a suitable belief change. The posterior beliefs need not involve the sort of odd causal structures of a Newcomb problem, but they do feature the failure of the decision making conditional to be independent of its antecedent. So in analyzing the measure of "desire" we have reason to expect a divergence between causal U and evidential V even in cases which are set up to initially not be Newcomb problems.

Consider an agent foraging in the forest. Let A be the proposition the agent eats a mushroom, G is that he lives, and G is that he dies. He is unsure whether he will eat the mushroom or whether is poisonous. Also, he does not think that whether it is poisonous makes any causal difference to whether he eats it. So the decision making conditional "If A then it would be that G" is independent of A, hence both it and its negation (the negation of the consequent) are compatible with $\neg A$. (If we suppose he thinks the odds of poison are 50-50, then those conditionals are equally likely given either A or $\neg A$.). Now he is informed by the oracle of the forest: "You will not eat the mushroom and live". The conditional is now certain given A, but still uncertain given $\neg A$. This failure of independence gives eating the mushroom low causal utility despite having high evidential value, and so the agent decides not to eat the mushroom.(Having so decided, and ruled out $A \wedge \neg G$, the odds of $A \rightarrow G$ go back up to 50-50, after going down to 25-75 between the oracle's announcement and his decision.)

The example isn't a classic Newcomb problem, since the agent doesn't learn of any causal power affecting his decision and the outcome (not even one mediated by the oracle, since the example contains no information about the oracle's basis for his announcement). Rather it is one where the agent learns the disjunction "Either the

mushroom is poisonous or you won't eat it", and given just this apparently non-causal information decides that the causal structure favors not eating. If the agent thinks the oracle is a super-predictor then the example is a Newcomb problem, but the agent might just think that the oracle knows he is a trusting sort of fellow willing to accept his announcement.[29]

2.5.5 Choiceworthiness versus Desire

In his second article, [65] Lewis provides a different and more substantial argument against relevance of causal utility to solving the puzzle about desire as belief. Unlike the first article, where he said we may ignore the cases where V and U diverge, here he insists on focusing on Jeffrey news value in the cases where it *does* diverges from causal utility.

Consider some action which has maximal causal utility while its evidential news value is low. For instance, the example of Fisher's smoker who continues smoking even though it gives him evidence that he is genetically predisposed to cancer. Lewis says:

> Should you perform that action?—Yes; your destiny is not a consideration, since that is outside your control. Do you desire to perform it?—No; you want good news, not bad. [65, p. 56]

According to Lewis it is always V that analyzes desire, while U is a measure of choiceworthiness. In non-Newcomb cases they go together, but in Newcomb cases choiceworthiness and desire come apart.

Suppose we grant this to Lewis. What the anti-Humean should say in response is in trying to identify desires with beliefs, what they really meant to do was to identify being in a motivational state with having a belief about goodness. Having

[29]In Lewis's [62] version of this story, the oracle (Lewis) plucks the mushroom and enjoys it, pleased with his "dirty trick" (since the agent took him to mean that the mushroom was poisonous). Oddie [77] gives a story of Frederic and the pirates having a similar structure. Oddie says "It is precisely when one's choices alter the probabilities of the range of those possible settled conditions which make the action more or less desirable, that Newcomb problems arise." That isn't quite right, because Newcomb problems involve the deviant common-cause structure that is missing from this case.

not considered Newcomb problems, they had presupposed that being motivated just was having a desire. But now they have been persuaded by Lewis that one may have a strongest desire without it being for the choiceworthy action. What the anti-Humean wants to say of the rational agent in a Newcomb problem is that his motivational state (to choose the choiceworthy option) is a belief about goodness. And this can be done by $U(A) = P(\mathring{A})$. This will coincide with desire, $V(A)$, in the non-Newcomb problems that had been assumed in the original setup of the debate.

2.5.6 Rationalization and Two Kinds of Desire

But the anti-Humean need not follow Lewis that far. The problem with the position Lewis takes is that it seems to say that in Newcomb problems, rational actions are not subject to belief-desire explanations. A platitude about rational action is that it is which picks the most desired feasible option. But in the Newcomb case, the thing which Lewis calls the "desire" does not rationalize the act certified as choice-worthy. In fact, by way of the platitude it would provide a (pseudo) rationalization of the opposed act (which causal decision theory condemns as irrational). This is an odd position for Lewis to put himself in, since Lewis endorses the Humean philosophy of mind, according to which rational actions are explicable as the joint causal product of beliefs and desires which rationalize the action.

Byrne and Hajek [17, p. 422] press this point against Lewis. They say it it would make "wholly mysterious" why a proponent of causal decision theory would chose a V-minimal (and thereby less desired, according to Lewis) option in a Newcomb problem. They say that while an evidential decision theorist should take V as the analysis of desire, the causal decision theorist should instead take U as the analysis of desire. In this way, the theorist's favored decision theory will advocate choosing the most desired option out of the feasible acts.

As long as one is firmly wedded to the platitude about choosing the most desired option, this response to Lewis sounds so sensible that it is initially baffling what drove Lewis to his position. However, doubt about the Byrne and Hajek position begins to creep in when we ask about how to accommodate Lewis's intuitions about

Newcomb problems. Byrne and Hajek say of such cases they "can explain your sense of dissatisfaction without supposing that you wanted" the unchosen V-maximal option. Rather, as they would have it: "You had hoped that you were not the kind of person who would choose [the U-maximal option] over [the V-maximal option], and things did not turn out that way."

Stronger doubts arise when we ask what role V plays in Lewis's equation for computing U. If we follow the Byrne and Hajek line, the causal decision theorist should say that V is not a measure of desire. But then why should V play any role in the calculation of U? So while the evidential theorist should give no role to U as a measure of desire, it seems the causal theorist must give some role to V as a measure of desire. According to Lewis the causal theorist should give exclusive privileges to V; according to the proposal I shall make, the causal theorist can let U and V share duties.

What Lewis ought to say in response to Byrne and Hajek is that while the folk platitude is true for the most part, like most folk theories it is underspecified and the details lead it to subtly break down in exceptional cases. According to the position Lewis ought to take, desire always plays a role in rationalizing action in virtue of its role as V in the equation: $U(A) = \sum_i P(A \rightarrow S_i)V(A \wedge S_i)$. In non-Newcomb cases, $U(A) = V(A)$, and the folk speak the literal truth when they say that the rational action is the most desired one. But in Newcomb cases, the folk platitude goes astray, even though desire (V) still plays the same role it always does in calculating choiceworthiness. The blame for this divergence between desire and choiceworthiness can be placed on the belief factor in choiceworthiness; it is because $P(A \rightarrow S) \neq P(A \rightarrow S/A)$ for some state S. Since Newcomb cases are unusual, it is no surprise that this glitch in our folk theory of desire has not drawn wider attention.[30]

Although Lewis can make this move without diverging terribly far from our folk intuitions, another position is at hand which comes closer to respecting the platitude that the rational action is the most desired one. It also comes closer than the Byrne

[30]Compare Lewis on "Mad Pain and Martian Pain", or Field on the splitting of the reference of the theoretical term "mass".

and Hajek theory of desire, which does not make room for the causal decision theorist to say there is any sense in which a rational agent might desire the V-maximal option while choosing the U-maximal one. Insofar as this third-way position does not incur other costs offsetting this closer match to folk intuition, it is preferable as the decision theoretic explication of desire. But the main attraction of the theory for the cognitivist is that, like the Byrne and Hajek analysis of desire and unlike the Lewis analysis, it allows for us to cash out "besires" by equations of the form $U(A) = P(\mathring{A})$.

The theory in question says there are two kinds of desire: passive and active. Jeffrey thought all desire was of the passive sort, the receipt of news items. He instead drew the active versus passive distinction amongst the propositions in terms of which were in your power; his decision theory said to choose the most desired of the propositions under your control. Causal decision theory retains the distinction between propositions under your control and those which aren't; the As are a partition of the former. I am claiming that the causal decision theorist should further distinguish two kinds of desire. These two kinds of desire go together in non-Newcomb cases, but come apart in Newcomb problems (as choiceworthiness and desire do, in the position I offered to Lewis above). Passive desire is like wishing that something were so. In the Newcomb case, one wishes that it were the case that the V-maximal option would obtain. But the rational agent recognizes this wish to be idle, and chooses the U-maximal option instead; the latter reflects the agent's active desire.

Unlike Byrne and Hajek, we can say that V is a kind of desire; hence we can explain its role in the equation for U, and also explain the wistful feeling of regret in Newcomb problems. Unlike Lewis, we can say that U is also a kind of desire, and hence the rational action is always the most desired (in this sense of desire). And for the cognitivist, room is now open to give a formal statement of the claim that desires ($U(A)$) correspond to beliefs about goodness ($P(\mathring{A})$). And this is a desirable position for a modest cognitivist to be in, with respect to Lewis's argument that he could formally demonstrate the untenability of the anti-Humean position.

2.5.7 Causal Utility as Motivating

The cognitivist might not rest with this modest achievement of opening up room for him to articulate his views, without lapsing into triviality (as Lewis alleged he could not do). The ambitious cognitivist thinks that we can accomplish something better than this: we can argue that having distinguished two senses of desire, the one we have picked to go with beliefs about goodness is in fact the sense of desire that best fits the concept of desire in the traditional meta-ethical debate. In other words, without their having realizing it, meta-ethicists are better construed as discussing U rather than V. Since the cognitivism debate and Newcomb's puzzle appear to be disparate issues, such a conception is surprising even by the usual standards for the "paradox of analysis". (Roughly, the puzzle of how philosophers can reach illuminating analysis of concepts that are not obvious—and are sometimes even controversial—to the folk using those concepts.) Nevertheless, the ambitious cognitivist thinks the connection has been made, as follows.

In the traditional meta-ethical debate, the question is whether the motivating forces of rational action can be grounded in (or even identical to) beliefs. Even if the issue of Newcomb problems had never occurred to meta-ethicists, we should be charitable in attributing the correct theory of rationality to them; that is certainly better than attributing evidential decision decision theory to them, or no theory of rationality at all. Likewise, we should also attribute to them the correct theory of desires, namely the theory of active and passive desires, coinciding in all but the Newcomb problems. Since the meta-ethicists were talking about the desire which rationalizes action, they are best understood as having meant causal U.

This interpretation strategy is reinforced by the intuition that once the Newcomb phenomena is brought to their attention, they would deny that that they had been talking about the merely passive sense of desire which V plays in such cases. The Humean says that sincerity in so-called "value judgments" is reflected in action, not that it is reflected in wishing that the background causal states of the world were different than they are.

2.6 How un-Humean is Decision Theory?

I have defended the claim that cognitivism can be given a plausible formulation within decision theory which does not run afoul of the triviality results. One might think that something more has been shown, that cognitivism—indeed, its internalist version—is a consequence of decision theory. For within evidential decision theory we can characterize news value by means of conditional beliefs about goodness, and within causal decision theory we can capture utility by means of unconditional beliefs about goodness. But this seems to be too strong a result.

The answer is that we only helped ourselves to a very mild form of cognitivism in order to establish its consistency. Starting with the agent's own news values, we defined a value level partition of propositions for that agent. This, in turn, was used to define the relevant beliefs. We placed no further constraints on this value level partition, such as that it agreed with the values of other agents, or corresponded to any independently specifiable normative facts in the world. We did not even claim that it was the semantic value of "good" in the agent's idiolect, and so did not commit ourselves to even the self-reporting theory of normative language; as far as the decision-theory models were concerned, expressivism could be a correct semantics of normative language.[31]

The reason we used this weak form of cognitivism was to show that some form of cognitivism is tenable, despite the triviality proofs. The doctrine Lewis picked as his main target was both internalist and universalist. It held that every rational agent has desires which connect with beliefs about goodness, and that the latter are understood in terms of a proposition about goodness which is shared amongst all agents. I did not take this view to be the primary one under investigation, since Lewis's triviality proof relies on a point about belief change which counts against cognitivisms which are externalist or which relativize propositions about goodness to an agent.

To defend an interesting form of cognitivism, one has more work to do, although the exact nature of the work depends on the form of cognitivism being defended.

[31] But as Dreier [28] notes, it is a short step from expressivism to a weak cognitivist semantics.

Especially, one might try give a metaphysics of facts about goodness, and a theory of intentionality showing we are capable of having beliefs about those facts. Such a metaphysics would then be combined with the decision theoretic arguments given here, in an account of how those beliefs line up with desires. An externalist story would leave it as a contingent matter whether we do desire those things we believe to be good. An internalist story would have to explain how it is that we always desire what we believe to be good; perhaps by saying that goodness facts are response-dependent and individuated by our desires. In any case, these questions are worth further investigating, now that we see that Lewis has not blocked all of the anti-Humean answers by means of his triviality proof.

Chapter 3

Modus Ponens Revisited

3.1 Introduction

The compositional structure of language might have led one to expect that a proper analysis of simple conditionals would have been adequate to determine the analysis of iterated conditionals. But McGee has presented an interesting group of examples that shows that this is not so for indicative conditionals. The examples are particularly arresting since they appear to show that modus ponens does not hold as a generally valid rule of inference for conditionals in natural language.

Many attempts have been made to explain away the data without following McGee in resorting to a logic and semantics where modus ponens is only valid for simple conditionals. I shall consider the most promising alternative proposals, and argue that they are unsuccessful. Having argued that we should follow McGee in taking the data about iterated indicative conditionals at face value, I shall explore the consequences of this for two important issues in the study of conditionals.

The first issue is a prominent thesis about the indicative conditional, namely Ramsey's proposal that to evaluate a conditional you check whether you would accept its consequent upon accepting its antecedent. I shall argue that McGee's logic helps explain a puzzle that afflicts this view. But on the other hand it shows that there is a wider divergence than might have been thought between actually learning information and hypothetically adding it to one's belief in the Ramsey Test. Moreover, it shows

that the widely accepted converse of the Ramsey Test is unacceptable. Additionally, the Ramsey Test helps throw light on why it is that modus ponens fails for indicative conditionals, by making relative to a doxastic or suppositional state.

The second issue is the relation between indicative and subjunctive conditionals. The question of whether the indicative and subjunctive have basically the same analysis is a hotly debated topic, but it usually conducted at the level of simple conditionals. McGee has claimed that his analysis extends from indicatives to subjunctives, and this has been defended by Levi. Against them, I shall argue that the data is unclear about open subjunctives, and that it shows that counterfactual subjunctives obey modus ponens instead.

3.2 An Uncertain Inference

3.2.1 The Main Example

The background setting for our main example is a certain stage of the 1980 United States presidential election. The leading Republican candidate Ronald Reagan was the front-runner nationally. In the national polls, he was followed by the Democratic incumbent, Jimmy Carter. Trailing them both by a wide margin was another candidate, John Anderson, who was for a while running as a Republican (eventually he switched over to being an Independent).

McGee [73] asks us to consider the following argument, from the perspective of a typical 1980 poll watcher:

> (1) If a Republican wins the election, then if it's not Reagan who wins it will be Anderson.
>
> (2) A Republican will win the election.
>
> So, (3) if it's not Reagan who wins, it will be Anderson.

From viewpoint we are asked to imagine ourselves in, the premises seem acceptable, but the conclusion seems unacceptable since if Reagan doesn't win then Carter will.

3.2.2 Some Lessons of the Example

McGee concludes, in the first instance, that we have a counterexample to the rule *modus ponens* treated as a rule of reasonable belief. At least we have a counterexample in the case where the consequent of the conditional is itself a conditional. McGee gives a logical and semantic theory which vindicates this intuition.

But we might follow Sinnott-Armstrong, Moor and Fogelin (henceforth "SMF") [96] in worrying how McGee makes so much of this apparent counterexample. They ask: what is a rule of reasonable belief? And what does it have to do with modus ponens as a principle of validity, in the usual logical sense of preservation of truth? To show that modus ponens is invalid, don't we need an argument that the premises are in fact true and the conclusion is false?

Let us focus for now on the last question. McGee could argue for the truth of the premises and the falsity of the conclusion on the basis of his own semantic theory, but SMF would presumably regard that as entirely question begging. What SMF want to know is how we can get from intuitive data about the unacceptability (or acceptability) of a sentence to the claim that it is false (or true).

It would be relatively straightforward to draw conclusions about logical validity if the premises of the argument were certain. For if a rational person understands two relatively simple premises, and upon reflection he is certain of them while dissenting from drawing a conclusion from them, then this is strong evidence that those premises do not entail the conclusion. Otherwise we would have to conclude that the person is being inattentive or unreasonable in not accepting the conclusion or retracting a premise.[1]

However, the premises in the example are not something we would all find ourselves certain of, given the circumstances we are supposed to imagine ourselves in. If our beliefs agree with the statistics given in the background story, then we find it likely but not certain that a Republican will win. Nevertheless, if two sentences entail a

[1]Note that such an example will show that the principle in question is not valid, even if its not the case that the premises are all true and the conclusion false. What matters is that the person can be rational in accepting the premises but not the conclusion, despite not recognizing those truth values.

third, it is impossible that the conclusion is improbable while one premise is probable and the other premise is almost certain. So if acceptability conforms to the laws of probability, the intuitions about acceptability show that modus ponens is not a valid inference for conditionals containing conditional consequents.

In any case, before turning to a more careful study of how modus ponens fails to be valid, we can see how the example counts against the most prominent analyses of the indicative conditional as a sentential connective. For in order to do that, we need only focus on how those theories interpret the conditionals in the example, and how they seem to get those meanings wrong.

3.3 Analyses of the Indicative Conditional

3.3.1 The Material Conditional

The most popular analysis of the indicative conditional is still the material conditional, where "If A then B" is true just in case A is false or B is true. This is so despite the many paradoxes of material implication, such as that we find it implausible to infer a conditional from the falsity of its antecedent. It is exactly that paradox which drives the implausibility of accepting the conclusion in McGee's counterexample to modus ponens. For it is the likelihood that Reagan will win which makes it unlikely that he will lose, and hence likely that "Either Reagan will lose or Anderson will win". And it is the fact that Reagan did win which makes the conclusion true, when it is interpreted as a material conditional. The argument is, of course, valid according to the material conditional analysis. And the premises are both true, and highly likely given the polls.

The material conditional analysis of the indicative conditional is sometimes defended on the grounds that it would be unreasonable to assert "If A then B" if one had reason to assert something stronger, such as "not A". Many suppose on the authority of Grice [38] that this can be explained in terms of conversational implicatures. But as Grice himself recognized, this defense of the material conditional runs

into problems with negations of conditionals. He granted that somebody denying a conditional "could not, it seems, in any case be supposed to have committed himself to the conjunctive thesis" of the affirmation of the antecedent and the denial of the consequent. [38, p. 80] For instance, in our example, we find it highly likely that "Its not the case that if Reagan loses, Anderson wins", but we don't find it highly likely that "Reagan and Anderson will both lose".[2] And we cannot use the "don't assert the weaker" pragmatic rationale to explain how somebody who denies a conditional would *not* assert the conjunction of the antecedent and negation of the consequent. Grice notes that sometimes denial is not the same thing as asserting the negation of what is denied, and he even sketches a theory so that negations of conditionals always have the effect of being negations of their consequents. But it is implausible that denial *never* means negation if the conditional is a sentential connective.

3.3.2 The Stalnaker Conditional

The argument is also valid if interpreted by the Stalnaker conditional, where "if A then B" is true just in case B is true at the closest or most similar world where A is true. The notion of closeness is a formal one to be filled in by context, but a general constraint is that every world is most similar to itself. As Stalnaker [99] explains, for *indicative* conditionals the closest worlds should be compatible with what is believed, or presupposed in a conversation, if the antecedent is.

To evaluate the conditionals in our example for truth, we need to ask which world is most like the actual world (where Reagan wins). According to the background story, it is a world where Carter wins rather than Anderson. This is what makes the conclusion false at every world other than the possible world where Anderson wins, and hence explains the judgment that the conclusion is unlikely. Indeed, it is with respect to negations that the Stalnaker analysis of indicative conditionals scores its major victory over the material conditional analysis.

[2]Grice cites a more dramatic example about God, due to Bromberger (unfortunately, the original printing of [38] is riddled with typos in this section). A related God example appears in Stevenson's article of the same era.

Unfortunately, this same judgment of closeness of worlds makes the iterated conditional premise false. Note that the actual world is closest to itself, with respect to a Republican winning. So to evaluate (1), we see whether "if its not Reagan who wins it will be Anderson" is true at the actual world. But the closest world to it where Reagan does not win is on where Carter does (so Anderson loses). Hence, (1) is false. Indeed, the only worlds compatible with a poll watcher's beliefs where the major premise is true are those where Anderson wins. Hence, according to the Stalnaker semantics, the major premise is just as unlikely as it is that Anderson wins.

There is a way that Stalnaker can try to accommodate the data, by taking advantage of the fact that conditionals are highly context sensitive. Stalnaker's theory of indicative conditionals makes them depend on what is presupposed in a context. Namely, worlds compatible with the context should be treated as closer than worlds outside the context. In order to handle the data about iterated conditionals, Stalnaker can say that the outer antecedent modifies the context in which the embedded conditional is evaluated. However, if the data systematically requires such a modification, then it would seem that a more systematic modification of the semantic theory is in order. But before turning to McGee's modification of the Stalnaker semantics, let us consider one more attempt to make do with more familiar semantic theories.

3.3.3 The Hybrid Theory

So the material conditional analysis gets the meaning of the conclusion wrong, while the Stalnaker analysis gets the meaning of the conditional premise wrong. Perhaps we could draw on the strengths of each theory, to avoid the weakness of the other. Lowe [68] proposes that we treat simple conditionals as stronger than the material conditional, but that we interpret McGee's major premise as being of the form of a material conditional embedded in a stronger conditional. So the major premise is equivalent to "If a Republican wins the election, then either it will be Reagan who wins or it will be Anderson."

The intuitive fallaciousness of the inference is really just an equivocation on the interpretation of the conditional. So from the premises we are not warranted in

concluding the strong conditional "If it isn't Reagan it will will be Anderson", but only the weaker "Either it will be Reagan or Anderson". But the latter is not the meaning of a simple conditional. (And moreover the material conditional is unassertable for the Gricean reason that we are in a position to assert the more informative "Reagan will win".)

Note that Lowe does not intend his analysis to be an *as hoc* interpretation of McGee's example and others like it. Rather, it is put forward as a systematic analysis of what right-nested conditionals mean. As such, the theory doesn't really preserve modus ponens for iterated conditionals, except in the vacuous sense that there are no conditionals of the form "If A, then C if B" where both "if"s have the same interpretation.

Unfortunately, this hybrid theory does not work for exactly the reasons why the material conditional is inadequate on its own as an analysis of the indicative conditional. If Lowe's analysis were correct then we could generate "embedded paradoxes of the material conditional":

> If a Republican wins, then if it isn't Reagan or Anderson it is Carter.
>
> If a Republican wins, then if its Carter it is Anderson.

Given the background story (where Reagan the Republican wins), the embedded conditionals are true if they are material conditionals, and thus the compound conditionals are true by the standard semantics for stronger conditionals. Indeed, they are true at every world compatible with the beliefs of a poll watcher. So according to Lowe's semantics, these compound conditionals should be just as believable as the major premise of McGee's example. But intuitively these compound conditionals are unreasonable to believe.

If we are tempted to apply Lowe's story for why these would be unreasonable to assert, then we should resist the temptation for the same reason as we should ultimately abandon the Gricean theory for simple conditionals: it cannot work for negations of conditionals. This is a problem with respect to Lowe's analysis of iterated conditionals, since we can assert conditionals with negated conditionals in their

consequents. And this one seems completely acceptable:

> If a Republican wins, then its not the case that if Reagan loses it will be Carter who wins.

But since a Republican wins, according to Lowe's analysis the whole sentence is true just in case Reagan and Carter both lose. Thats not how the election went, so Lowe brands as false a conditional which ought to come out as true.

3.3.4 Conditional Probability and Adams' Thesis

So far we have been treating all the judgments about the acceptability of the conditionals as judgments about the probability of a proposition expressed by the conditional. SMF focus their dispute on whether the conclusion of the modus ponens example is in fact improbable. SMF maintain that McGee has confused the probability of the conditional "If A then B" with a judgment about the conditional probability $P(B/A)$. The latter interpretation of the indicative conditional is known as Adams' Thesis, after [2]. SMF illustrate their point with a nice example. Imagine that the election is to decided by a fair six-sided die with "R" on three sides, "C" on two, and "A" on one. The claim that a Republican wins is equivalent to "R or A". The conditional probability that Anderson wins given that Reagan doesn't is 1/3, while the probability of the material conditional, equivalent to "Either Reagan wins or Anderson does", is 2/3.

This is unsatisfactory as a defense of modus ponens, and especially as a defence of the claim that when the conditional does express a proposition it is the material one. SMF's position seems to be that we have an equivocal concept of the indicative conditional, which sometimes functions as the material conditional (as in the major premise of the modus ponens example) and sometimes as a conditional probability (as in McGee's reading of the conclusion of the argument). But this position is unstable, as we are owed an explanation of why we the material conditional analysis is correct if it diverges from from the conditional probability analysis that accounts for our intuitions about the simple conditionals in the example. Later we will return to

Frank Jackson's attempt to reconcile these two incompatible theories. But for now, let us note that the Stalnaker logic has the same logic as Adams' probabilistic logic with respect to simple conditionals. What is needed is a modification of the Stalnaker logic to handle iterated conditionals, and that is what we turn to next.

3.4 The Import-Export Equivalence

3.4.1 The Logic of Iterated Conditionals

McGee concludes from his example that modus ponens is invalid as applied to nested conditionals. Rather, what is valid is the equivalence of the iterated conditional (1) with "If a Republican wins the election and it's not Reagan, it will be Anderson."

Importation: $A > (B > C) \vdash A \wedge B > C$

Exportation: $(A \wedge B) > C \vdash A > (B > C)$.

(The connective $>$ is used with the intended interpretation of representing the natural language indicative conditional.) As McGee shows, the material conditional is the only conditional obeying Exportation and modus ponens.

In order to verify the Import-Export equivalence, McGee [73] modifies the Stalnaker semantics with an extra parameter so that sentences are evaluated for truth at a world with respect to a set of hypotheses. In the case of the null set of hypothesis, the semantics is equivalent to the standard Stalnaker semantics. Hence, the semantics agrees with the Stalnaker theory for *simple* conditionals. So $A, A > B \vdash B$ where A and B have no conditionals embedded within them. The theory diverges from the standard Stalnaker theory in the case of iterated conditionals. The right nested conditional $A > (B > C)$ is true at a world just in case the closest $A \wedge B$ world is a C worlds, the same truth conditions as its imported form $(A \wedge B) > C$.

3.4.2 Modus Ponens as a Reasonable Inference

With McGee's semantics in hand, we can see how even by his own lights his counterexamples to modus ponens as a rule for reasonable belief depend on the premises

being uncertain. To see this, let us follow Stalnaker [99] in saying that when A is compatible with a context, then an indicative conditional "If A then B" should be evaluated with respect to worlds compatible with the context according to this constraint: A worlds in the context are closer than A worlds outside the context. By Stalnaker's constraint for indicative conditionals, adapted to McGee's adaptation of Stalnaker's semantics, $A \rightarrow (B \rightarrow C)$ is only reasonably accepted under the following condition: $A \wedge B$ is compatible with S, and every $A \wedge B$ world in S is a C world.

Let S be any context in which that iterated conditional is reasonably accepted. Let S' be the context coming from S by the acceptance of A: $S' = S \cap A$. This leaves the $A \wedge B$ worlds as the only B worlds left in S', and those are all C worlds. Similarly, in any initial context where A is accepted, if $A \rightarrow (B \rightarrow C)$ becomes reasonably accepted, then $B \rightarrow C$ is accepted in the resulting context.

So using McGee's semantics, Modus Ponens is a pragmatically reasonable inference in Stalnaker's sense. If either premise is reasonably accepted and then the other becomes reasonably accepted, then the conclusion is accepted. This shows how the persuasiveness of McGee's example depends on the premises being uncertain.[3] It could help explain why Modus Ponens seems valid despite not really being so, since insofar as we try to evaluate logical principles by considering what beliefs should be drawn from other beliefs, we may typically imagine beliefs that we are certain about. We know that aside from the puzzle of iterated conditionals, it would be improper to determine logical validity by what conclusions we would accept if we accepted certain premises. For as Stalnaker [99] points out, from $\neg A \vee B$ one may infer the conditional "If A then B"; Stalnaker calls this "the direct argument". Yet we do not on that basis wish to say that the indicative conditional is not logically stronger than the material conditional, hence Stalnaker concludes that the direct argument is merely pragmatically reasonable but not logically valid.

[3]Likewise, the Stalnaker analysis of the conditional premise will seem fine in a context where it is accepted that a Republican will win; the possibility of Carter winning is needed for the counterexample. Over [79] and Gillies [36] argue that we ought to treat the example as if the unconditional premise were certain, by comparison to a reworded conclusion: "If it isn't Reagan, its Anderson." Since the anaphora in the reworded conclusion refers back to "A Republican wins", it picks up as a presupposition that the winner is a Republican. But, contra Over and Gillies, this doesn't show anything about the original argument where the conclusion presupposes no such thing.

3.4.3 Trouble with Exportation

McGee's example points to a conflict between Modus Ponens and Exportation. Let us take for granted what is common to all parties here (where $\supset$ is the material conditional),

Implication: If $\vdash \phi \supset \psi$ then $\vdash \phi > \psi$.

Making use of this and the validity of $\phi \wedge \psi > \phi$, Adams derived from Modus Ponens and Exportation the second paradox of material implication: $\psi \vdash \phi > \psi$. Gibbard and McGee then tweaked the proof to show the strengthened result that Exportation, Implication, and Modus Ponens entail,[4]

Equivalence: $\vdash (\phi \supset \psi) \equiv (\phi > \psi)$.

Gibbard concludes that Adams is right in thinking that the conditional does not express a sentential connective which is stronger than the material conditional, while McGee blames Modus Ponens.

Perhaps it is Exportation which is to blame for this theorem and for whatever is going on in McGee's counterexample to Modus Ponens. So lets take a closer look at it. Consider the following conditionals from Gibbard, [34, p. 246].

> If Andrew Jackson was President in 1836 and he died in 1835, he was president in 1836.
>
> If Andrew Jackson was President in 1836, then even if he died in 1835, he was president in 1836.

Although it sounds a little odd, the first conditional is valid as an instance of the principle used by Adams. And so the second conditional is valid if Exportation is. But that sentence sounds more problematic, despite Gibbard's intuition that it is valid.

However, McGee can offer the following apology for the oddity of Gibbard's second sentence. On the standard Lewis-Stalnaker theory, a conditional with an impossible

[4]Gibbard's version assumes Importation as well but this is not used in the proof. I give a version of the proof in the next section.

antecedent is vacuously true. This allows Implication to be valid, since the material conditional $\phi \supset \psi$ is valid when ϕ is invalid. It is has unintuitive consequences, since it makes Gibbard's first sentence valid. But by McGee's use of the Import-Export equivalence, Gibbard's second sentence is equivalent to his first sentence.

So the oddity of Gibbard's exported conditional can be written off by McGee in an a way analogous to how Lewis or Stalnaker would write off the oddity of impossible antecedents according to their theory. However, we still need to cope with further difficulties with $A{>}(B{>}A)$ where $A{\wedge}B$ isn't metaphysically impossible, but nomically impossible or even just very implausible. Kremer [53, p. 212] faults the import-export law for supporting fatalism, and he credits the uncovering of the fallacy to this quote from Hobbes:

> if I must do this rather than, I shall do this rather than than that, though I consult not at all; ... is a false proposition, and a false consequence, and no better than this: If I shall live tomorrow, I shall live tomorrow though I run myself through with the sword today. [47, pp. 254–5]

To defend the import-export law from this criticism, let us first note that we often express fatalism through simpler logical truths:

> If Hobbes is going to live, then Hobbes is going to live.

So the oddity of

> If Hobbes is going to live, then if Hobbes runs himself through with a sword, then he is going to live.

could be written off to the oddity of uttering logical truths. To reinforce this, note that it doesn't sound any better to say,

> If Hobbes is going to live, then if Hobbes runs himself through with a sword, then he is going to die.

Indeed this sentence sounds even worse than the fatalistic conditional, although according to the opponent of import-export it should be highly acceptable. The sentence doesn't get better if we say

> Even if Hobbes is going to live, then if Hobbes runs himself through with a sword, then he is going to die.

The best one can do to make sense of what this is trying to say is to put the embedded conditional in the subjunctive:

> Even if Hobbes is going to live, if Hobbes were to run himself through with a sword, then he would die.

This sentence sounds fine, but such a mixed mood conditional doesn't provide any direct evidence about the logic of purely indicative conditionals.[5]

3.4.4 Importation and the Direct Argument

Katz writes,

> Although McGee's examples are indeed arresting, ... I find it doubtful whether there is any consistent interpretation of the conditionals occurring in McGee's example ... which will make the premises true and the conclusion false. [51, pp. 404, 405-6]

The reason Katz reaches such a negative conclusion is that he thinks it analytic of the ordinary conditional that they obey not only Implication but its converse,

> Converse Implication: If $\vdash \phi > \psi$ then $\vdash \phi \supset \psi$.

And Katz proves that Implication, Converse Implication, and Importation together entail Modus Ponens.

> By the idempotency of the material conditional, $\vdash (\phi > \psi) \supset (\phi > \psi)$. So $\vdash (\phi > \psi) > (\phi > \psi)$ by Implication. Hence $\vdash ((\phi > \psi) \wedge \phi) > \psi$ by Importation. Therefore $\vdash ((\phi > \psi) \wedge \phi) \supset \psi$, by Converse Implication.

[5]Nor does it tell us anything about whether purely subjunctive conditionals obey import-export, contrary to Katz's claim [51, p. 407] about similar indicative-subjunctive conditionals due to Olin [78]. Appiah [7] claims that no iterated conditionals are purely indicative or purely subjunctive, despite surface appearances.

Since Katz thinks that Exportation ought to stand or fall with Importation, he concludes that McGee has chosen the wrong package of logical principles.

But it is not clear that Implication and its Converse need to be accepted in unison. After all, Converse Implication is a weakened version of

Modus Ponens: $\vdash (\phi > \psi) \supset (\phi \supset \psi)$.

So anybody opposed to the latter should distrust the former. Returning to Gibbard's example from the last section, by Converse Implication, it is valid that "Either Andrew Jackson was not President in 1836, or he was president in 1836 even if he died in 1835." So in any world where Jackson *was* President in 1836, it is true that "Jackson was president in 1836 even if he died in 1835." But that conditional is surely invalid unless it is the material conditional (in which case it is a roundabout way of saying that Jackson did not die in 1835). But the consistency of "If A then not A" is one of the paradoxes of material implication we are trying to avoid.

This distrust turns out to be well placed. For McGee is trying to steer a course clear of the theorem whereby Equivalence is entailed by Implication, Exportation, and Modus Ponens. Now, Equivalence is simply the conjunction of Modus Ponens and its converse,

Direct Argument: $\vdash (\phi \supset \psi) \supset (\phi > \psi)$.

But a closer look at the proofs by Gibbard or McGee shows that although Modus Ponens is used in proving the Direct Argument, it can be replaced in that half of the proof by Converse Implication.

> Exportation, Converse Implication, and Implication entail the Direct Argument. Proof: By classical logic $\vdash [(\phi \supset \psi) \wedge \phi] \supset \psi$. So $\vdash [(\phi \supset \psi) \wedge \phi] > \psi$ by Implication. Thus $\vdash (\phi \supset \psi) > (\phi > \psi)$ by Exportation. So $\vdash (\phi \supset \psi) \supset (\phi > \psi)$ by Converse Implication.

Perhaps Modus Ponens could be false of the ordinary language conditional while Converse Implication is true. But the proponent of Exportation is already committed to giving up Converse Implication, if he wishes his conditional to be strictly stronger than the material conditional.

3.4.5 The Supplemented Equivalence Theory

Finally, we turn to an attempt to reconcile the import-export equivalence with modus ponens, by considering the familiar logic that obeys them both: the material conditional. Jackson [48] proposes to reconcile the material conditional with Adams' thesis, which as he presents it is that a conditional is highly assertible just in case the corresponding conditional probability is high. Jackson proposed to explain this not by saying that the conditional probability is the probability of a conditional, but rather that assertability is given by a different equation which delivers a value that coincides with the corresponding conditional probability. This gives him a way to say that a valid inference might have highly assertible premises but a highly unsassertable conclusion.

Jackson claims that amongst the inferences this applies to is modus ponens for iterated conditionals. Moreover, Jackson claims that he can explain how exported conditionals are just as assertible as their imported versions. That is what one might expect if they are in fact logically equivalent (thanks to the material conditional), but it is actually a puzzle that needs solving since unlike ordinary probability, Jackson's equation for assertability is sensitive to syntactic structure.

Starting with simple conditionals, Jackson's "supplemented equivalence theory" can be broken into its two components. First, the "equivalence" part: the conditional has the logical form of the material conditional. Second, the "supplemented" part: the conditional's assertability is measured by two factors: how probable the (material) conditional is, minus how *robust* that probability is with respect to its antecedent. The justification for the latter claim is that Jackson thinks the conditional carries a conventional implicature that $P(A \supset B)/A$ is high, so that gaining evidence for A is prevented from diminishing the probability of $A \supset B$ and thereby making modus ponens unusable. Hence, [48, p. 32]

$$As(A > B) = P(A \supset B) - [P(A \supset B) - P(A \supset B)/A)] = P(B/A).$$

This formula for assertability vindicates the claim that the assertability of a condi-

tional goes by Adams' Thesis.[6] Jackson claims that this theory can explain not only why the traditional paradoxes of implication appear fallacious, but also how modus ponens appears to be invalid in examples such as the ones we have studied (he presents his own). In each case, the story is that the assertibility of the premises is high while the assertibility of the conclusion. There are a variety of objection which can be made to the supplemented equivalence theory in respect to simple conditionals. However, I shall focus on the difficulties besetting it for iterated conditionals.

To illustrate how the conditional seems to both obey the Import-Export principle and Adams' thesis, Jackson presents this example (due to Roy Sorenson), which nicely makes the probabilities explicit:

> We know that a fair coin was subjected to 20 independent tosses. Consider 'If this coin landed heads at least three times, then if the first 17 tosses were all tails, the last three were all heads'. This is exactly as highly assertible as 'If the coin landed heads at least three times but not at all in the first 17 tosses, then the last three tosses were all heads'; and yet 'The coin landed heads at least three times' is highly assertible, while 'If the first 17 tosses were all tails, the last three were all heads' is highly unassertable (remember the tosses are independent)." [48, p. 131]

Jackson claims how he can explain how the assertibility of $A > (B > C)$ is the same as that of $(A \wedge B) \supset C$, which is given by $P(C/A \wedge B)$. This is a genuine puzzle since a straightforward application Adams' thesis would say that it is given by $P(B > C/A)$. Indeed, it is in good part thanks to such puzzles that Adams only applies his thesis to simple conditionals. What Jackson says, attributing the idea to Gardenfors, is that those accepting his general theory "should expect" that the assertibility of $A > (B > C)$ is given by $P_A(B \supset C/B)$, which equals $P(C/A \wedge B)$. While the last probability identity is correct, it is unclear why the supplemented equivalence

[6] In his review of Jackson's book, Adams calls this "a rather dubious equation." He adds, "I should say parenthetically that while I am the Adams in question, Jackson graciously (and rightly) absolves me from responsibility for his formulations." [3, p. 433] Adams goes on to praise Jackson for giving "a different and simpler measure of assertibility than David Lewis", an opinion shared by Lewis himself who adopted Jackson's theory over his own (Postscript to [62]).

theorist is entitled to the first claim about what the assertibility goes by.

For a straightforward application of Jackson's assertibility equation to the iterated conditional says that the robustness factor is $P([A \supset (B > C)]/A) = P(B > C/A) = P_A(B > C)$. But what Jackson needs to get his conclusion from this is $P_A(B > C) = P_A(C/B) = P(C/A \wedge B)$. Unfortunately, that rests on a confusion between $P_A(B > C)$ and a conditional assertibility $As_A(B > C)$. I see no problem with conditional assertibility so long as we have the unconditional kind, but the problem is that it is not measure of the conditional probability of $B > C$; to think otherwise is to confuse Jackson's position with Adams' own equation. So if we want to have a conditional whose assertibility runs in line with Adams' Thesis, and we want Import-Export to be equivalent in assertibility, we should turn to McGee's theory.

3.5 Iterated Conditionals and the Ramsey Test

Adopting the Import-Export equivalence instead of Modus Ponens helps resolve a puzzle about the Ramsey Test for accepting conditionals, which says to accept a conditional if you would accept its consequent upon revising your beliefs to accept the antecedent. However, we shall see that this move has consequences about the nature of "hypothetically revising one's beliefs", so that it diverges in somewhat surprising ways from actually revising one's beliefs upon learning information. First we shall consider the Ramsey Test in framework studied by Gardenfors, where beliefs are characterized as linguistic items. Then we shall consider how the same issues appear in a theory such as Harper's, which focuses on the semantic contents of beliefs. After that we turn to comparisons between conditionals and modals, and how these favor the semantic perspective and suggest that the failure of modus ponens for iterated conditionals is part of a more general phenomenon.

3.5.1 Triviality for Simple Ramsey Test Conditionals

In order to better distinguish the treatment of conditional from non-conditional sentences, we borrow a dual alphabet lettering scheme which from McGee (Adams uses

the opposite convention, while Arlo-Costa uses a hierarchy of languages). Let us adopt a language where $\rightarrow$ is the material conditional, Roman letters (A, B, C) are factual sentences not containing the conditional $>$, and Greek letters (ϕ, ψ, θ) are sentences which may contain that connective. The language has a deduction relation $\vdash$ which is an extension of classical logic, and $\perp$ to represent a contradiction.

We shall use K to denote a belief state, a set of sentences satisfying

Closure: If $K \vdash \phi$ then $\phi \in K$.

K_ϕ is the result of revising K by learning ϕ, and iterated revisions are expressed as $(K_\phi)_\psi$. As Bradley notes, although the Ramsey Test if usually expressed as a biconditional, it is helpful to keep its two halves separate. Following his nomenclature, we reserve the title "RT" for the right to left direction of the standard formulation, and call its converse Conditional Driven Revision:

(RT) $\phi > \psi \in K$ if $\psi \in K_\phi$
(CDR) If $\phi > \psi \in K$ then $\psi \in K_\phi$.[7]

With the RT and its converse in hand, the logic of the conditional can be characterized by properties of the belief revision function.

A standard theory [33] for learning information consistent with prior belief is that the posterior is an Expansion of the prior belief state: the learned sentence is added and the logical consequences are drawn. For our purposes we can restrict ourselves to learning factual sentences. The Expansion rule can be broken into three parts, the Success condition that the learned sentence becomes accepted, the Preservation condition that no beliefs are given up, and the Inclusion condition that no beliefs are added which are not entailed by the prior beliefs plus the new information:

(S) $A \in K_A$

[7] As a historical aside, Bradley [14, p. 7] also claims that there is no "explicit" support for CDR in Ramsey's own writings, but this is implausible. For Ramsey says that RT holds because one is "hypothetically" adding the antecedent to one's "stock of knowledge", which he says amount to the person's fixing his conditional probability in the consequent given the antecedent. So Bradley would only be right if Ramsey thought that actually adding information to one's beliefs, as opposed to doing so merely hypothetically, was not by conditionalization. But Ramsey thought no such thing.

(P) If $\neg A \notin K$ and $\phi \in K$, then $\phi \in K_A$

(I) If $\phi \in K_A$ then $A \supset \phi \in K$.

Unlike standard conditionalization in probability theory, Gardenfors treats revision as well defined when you learn information inconsistent with prior belief. In order to generate his triviality result, he imposes a condition inspired by Stalnaker's [102] presentation of the Ramsey Test for belief contravening information, that one should make a minimal modification of one's beliefs to maintain consistency. Still restricting ourselves to factual formula as observations, we shall put the Consistency principle as :

(C) If $K_A \vdash \perp$ then $K \vdash \perp$ or $A \vdash \perp$.[8]

So revision only leads to inconsistency if either the prior state or the observation itself was inconsistent.

Now consider any belief state which is *non-trivial*, in the following sense: there are sentences A, B such that $A, \neg A \wedge B, \neg A \wedge \neg B$ are all compatible with K. So long as revision goes by Expansion at least for factual sentences, it follows that $B \in (K_{A \vee B})_{\neg A}$ and $\neg B \in (K_{A \vee \neg B})_{\neg A}$, with those states consistent. It follows by RT that $\neg A > B \in K_{A \vee B}$ and $\neg A > \neg B \in K_{A \vee \neg B}$. So if conditionals also obey Expansion, it follows that $\neg A > B, \neg A > \neg B \in K_A$. But then $B, \neg B \in (K_A)_{\neg A}$ by CDR, contradicting Consistency. That is the Gardenfors triviality theorem.[9]

3.5.2 Conditional-Factual Dualism and Iterated Conditionals

Gardenfors diagnoses his triviality theorem as pitting the Ramsey Test against Preservation, and opts to give up the former. But Bradley places the blame on conditionals

[8]This is based on [32]; as Arlo-Costa [8] notes, further problems are in the version employed by [33]. Arlo-Costa proves that adding Gardenfors' consistency condition (which Arlo-Costa calls the "success postulate") induces an S5 logic, when added to Gardenfors' [33] Ramsey Test theory of the Lewis conditional (which does not include the Inclusion postulate, thereby avoiding the triviality result below).

[9]This presentation is close to the more thorough proof in Rott [89]. I discuss issues surrounding this proof, including its relation to Lewis [62] and Gibbard [34], in my [30]. In order to avoid belief-contravening revisions and the consistency condition, Bradley helps himself to some logic of the conditional. Especially he adopts a form of conditional noncontradiction ($A > B$ and $A > \neg B$ are incompatible) but he does not restrict it with the standard qualification that antecedent is possible—a move he defends in [13]—and for this reason I do not follow his proof.

obeying Preservation (also see Rott [89] for a critique of conditionals obeying Expansion). After all, there is an intuitive sense in which the conditionals in the proof should then be unneeded and unwanted when the agent comes to believe A. Unneeded, since at point the agent has no need to plan for what happens if he learns $A \vee B$ or $A \vee \neg B$ since he is certain of both. Unwanted, since those iterated conditionals are what got him in trouble upon subsequently coming to believe $\neg A$.

If we simply deny that Expansion holds for conditionals, but allow that it is obeyed by factual sentences, a new puzzle arises for iterated conditionals. For by a variation on the above triviality proof, we have by RT for iterated conditionals that $(A \vee B) > (\neg A > B) \in K$ and $(A \vee \neg B) > (\neg A > \neg B) \in K$. Bradley says that even ignoring the Ramsey Test, the acceptance of those conditionals in a non-trivial belief state "is also so on most accounts of conditionals." [14, p. 9] But that isn't be right, for the most standard theories are those having a possible worlds semantics in the style of Stalnaker Lewis, and in such a theory the two conditionals cannot both be accepted in a single non-trivial belief state.[10]

In the Lewis semantics, a conditional is true at a world if the consequent is true at all the closest possible worlds where the antecedent is true. Consider any world w compatible with belief where A is true. Since $A \vee B$ is true at w, $(A \vee B) > (\neg A > B)$ is true at w only if its consequent is true at w. And $\neg A > B$ is true at w just in case B is true at all the worlds closest to w where $\neg A$ is true. Likewise, since $(A \vee \neg B)$ is also true at w, $(A \vee \neg B) > (\neg A > \neg B)$ is true at w only if $\neg B$ is true at all the $\neg A$ worlds closest to w. But its impossible that both B and $\neg B$ are true at all $\neg A$ worlds closest to w, since there is at least one $\neg A$ world compatible with belief.

The proof just given relies on the assumption that there is some $\neg A$ world accessible to some A world compatible with non-trivial belief. This is to avoid the iterated conditionals being vacuously true due to their inner conditionals having impossible antecedents. I don't think this is a terribly controversial assumption of the proof, even though there are questions about what accessibility relation is determined by

[10]Here Bradley is discussing theories that obey modus ponens, although in other work Bradley [13] endorses the import-export law.

the modality relevant for Ramsey Test conditionals. It seems to be a doxastic modality, and therefore weaker than S5, but it would be peculiar if it were restricted in the way needed to dodge triviality.

Although the iterated conditionals $(A \vee B) > (\neg A > B)$ and $(A \vee \neg B) > (\neg A > \neg B)$ cannot hold in a nontrivial belief state given the standard semantics for logics obeying modus ponens, there is no problem with them if the Import-Export principle is correct. For then they are equivalent to $B > B$ and $\neg B > \neg B$, respectively. So the Import-Export principle seems to provide a useful tool in salvaging the Ramsey Test from triviality, along with treating conditionals differently from factual sentences in belief revision. But as we shall see next, we must treat conditionals even more differently from factual sentences.

3.5.3 Import-Export Triviality Results for Revisions

Besides some of the principles have already introduced, we shall employ a weaker form of the Preservation condition, and we shall extend to the case of belief-contravening conditions a principle which holds in the case of belief-consistent expansion. First, the Weak Preservation condition, that no believed sentences are given up by revising by a factual formula which is already believed,

(WP) If $\phi, A \in K$ then $\phi \in K_A$.

Second, the Equivalence condition, that revision draws no distinction between provably equivalent factual formulas,

(E) If $\vdash A \leftrightarrow B$ then $K_A = K_B$.

Also, just to be totally clear about where the force of the Ramsey Test for iterated conditionals comes in, besides the earlier formulations RT and CDR, I present a version restricted to factual formula,

(RTF) $A > B \in K$ if $B \in K_A$.

First we show that CDR (the left to right direction of the Ramsey Test) for iterated conditionals is incompatible with Exportation. Assume Exportation as a constraint

on belief states: if $(A \wedge B) > C \in K$ then $A > (B > C) \in K$. (Since belief states are logically closed, this would follow from assuming Exportation as a logical principle, but we need not assume that.) Consider any belief state where $A \notin K$. Note that $A, \neg A$ are each consistent by Closure, so $K_{\neg A}$ and $(K_{\neg A})_A$ are consistent by (C). Since $A \wedge (A \rightarrow \neg A) \in K_{A \wedge (A \rightarrow \neg A)}$ by (S), and $A \wedge (A \rightarrow \neg A) \vdash \neg A$, it follows by Closure that $\neg A \in K_{A \wedge (A \rightarrow \neg A)}$. So by RTF, $A \wedge (A \rightarrow \neg A) > \neg A \in K$. By Exportation, $(A \rightarrow \neg A) > (A > \neg A) \in K$. So by CDR, $A > \neg A \in K_{A \rightarrow \neg A}$. Since $\vdash \neg A \leftrightarrow (A \rightarrow \neg A)$, by (E): $K_{\neg A} = K_{A \rightarrow \neg A}$. Hence $A > \neg A \in K_{\neg A}$. Therefore $\neg A \in (K_{\neg A})_A$ by RTF. But $A \in (K_{\neg A})_A$ by (S). Thus $(K_{\neg A})_A$ is inconsistent, contradicting the above.

Second we show that RT (the right to left direction of the Ramsey Test) for iterated conditionals is incompatible with Importation. Assume Importation for belief states: if $A > (B > C) \in K$ then $(A \wedge B) > C \in K$. Suppose $\neg A, A > B \in K$. Since $\neg A \vdash \neg(A \wedge B)$, by Closure $\neg(A \wedge B) \in K$, so $A > B \in K_{\neg(A \wedge B)}$ by (WP). Hence $\neg(A \wedge B) > (A > B) \in K$ by RT. Therefore $(\neg(A \wedge B) \wedge A) > B \in K$ by Importation. So $B \in K_{\neg(A \wedge B) \wedge A}$ by RTF. But $\neg(A \wedge B) \wedge A \in K_{\neg(A \wedge B) \wedge A}$ by (S), and $\neg(A \wedge B) \wedge A \vdash \neg B$, so $\neg B \in K_{\neg(A \wedge B) \wedge A}$ by Closure. So by (C), either $\neg(A \wedge B) \wedge A$ is inconsistent (so B is valid) or K is inconsistent. But that is absurd, since all we assumed was that $A > B$ is believed while its antecedent is rejected.[11]

3.5.4 Lessons of the Iterated Triviality Results

One move to try to dodge these results is to give up the consistency condition (C). Arlo-Costa [9, 10] advocates this, claiming it is shown to be untenable in the case of iterated revision. Although this blocks the derivations just presented, it doesn't seem to get at the core of of the problem. For applying the Ramsey Test to iterated conditionals gives the result that we should accept sentences such as the following, in a variation of McGee's example where we are certain that Reagan is going to win:

If a Republican wins, then if Reagan loses, Carter wins.

[11] I got the idea for the first theorem from the Gibbard-McGee proof that Export and MP reduce to the material conditional, and the idea for the second from Stalnaker's [101] proof that Import-Export is trivial in a setting of probabilities defined for conditioning on probability zero. The results can be found in some form in Cross and Thomason [21] and Arlo-Costa [9].

If a Republican wins and Reagan loses, then Carter wins.

These are intuitively the wrong thing to say, since Anderson would be the winner in the scenario these conditionals describe. But they are forced on us by the package of principles in question. We accept "If Reagan loses, Carter wins". By adding the already accepted "A Republican wins" to our beliefs we retain all our beliefs, according to Weak Preservation. So by the Ramsey Test for iterated conditionals, we accept the first conditional. So by Importation we accept the second.

One might think instead that neither half of the Ramsey Test (RT or CDR) applies to iterated conditionals.[12] This seems the correct thing to say about Conditional Driven Revision, which should be given up for iterated conditionals as it is so close to modus ponens. However, there is room to keep the Ramsey Test, if we give up Weak Preservation. As with the earlier treatment of the Preservation condition, we restrict its application to sentences containing no conditionals. This might seem a hard pill to swallow, since the Weak Preservation condition just seems to be such a weak condition.

Some think that we should abandon Weak Preservation for all sentences (and so give up Centering for simple conditionals). Levi [57], and following him Arlo-Costa, [9] argue by means of the following kind of example. A fair coin is tossed and lands head, but it seems odd to say "If that coin is tossed it landed heads" or "If that coin had been tossed it would have landed heads". But the oddity of these conditionals need not show that they are not true, since the antecedents are odd since they are known to be satisfied. To address the worry that the semantics fails to accommodate the indeterminism of the coin toss, we can point out that before the coin was tossed there was no fact of the matter about how it would land. So on the Stalnaker analysis both "If the coin is tossed it will land heads" and "...it will land tails" are both lacking in truth value, and on the Lewis analysis they are both false.

So we only want to give up Weak Preservation with respect to conditional sentences, so that it does not cause trouble for iterated conditionals. Friedman and

[12]Indeed, in the case of the probabilistic Ramsey Test, it is common to hold that $P(A > B) = P(B/A)$ is restricted to simple conditionals. For the triviality proof for Import-Export see Skyrms [97], for the Lewis style logics see the Weakened Transitivity Result in Hajek and Hall [39].

Halpern [31] argue that the principle gets its appeal from an uncritical examination of what it is to accept a sentence. They suggest that while re-observing something you already believe might not alter your beliefs, it can re-prioritize your belief revision policies. That is, in a framework where such policies are encoded by simple Ramsey Test conditionals, you can give up (or add) conditionals. Unfortunately, it is difficult to accept this as a story of the process of learning information.

The standard story about conditionalization is that it applies to your strongest evidence, and it is difficult to see how your strongest new evidence can be something you antecedently believe. Since you never do "learn" something you already believe, Weak Preservation is a way of saying that the revision function is defined in an innocuous way for that degenerate case. But we could instead treat the degenerate case differently, so as to get the Ramsey Test to work correctly for iterated conditionals, without this in any way affecting actual learning. (Alternately, we can drop Weak Preservation for the function characterizing the Ramsey Test, and hold that it diverges in this case from the revision function which characterizes learning.) The hypothetical supposition involved in the Ramsey Test will still agree with the result of learning information which is compatible with, but not already accepted in, prior belief. Indeed, this was the case which Ramsey concerned himself with.[13]

3.5.5 Semantics and Indexicalism

So far we have studied belief states as sets of sentences, and conditionals as elements of those sets which get treated differently than factual sentences in belief-revision. A somewhat different perspective on the Ramsey Test and triviality comes from taking as primary the information content of beliefs and sentences, as in a possible worlds semantics. We could replicate the moves above by drawing purely conditional distinctions between possible worlds, which do not otherwise differ with respect to non-conditional facts.[14] However, rather than locate the factual-conditional distinction

[13]Levi [57] advocates the view that believed sentences should first be suspended, and then re-added, in a way that may not be weakly preservative. That position is unnecessary if our sole goal is to keep Import.

[14]See Stalnaker's discussion of Tweedledee and Tweedledum in [103].

at the level of content, it is more natural to say that it is because they are indexical to belief or supposition states that indicative conditionals behave differently from factual sentences (pretending the latter are all "eternal" sentences).

Among fans of the propositional analysis of indicative conditionals, this has been the most popular means of dealing with the triviality results for simple conditionals. Since different propositions may be expressed by the same conditional sentence relative to different doxastic states, the triviality proof is blocked.[15] In this way of proceeding, the primary objects of belief and belief revision are propositions, and these obey preservation. When one believes the proposition expressed by a conditional relative to some belief state, one will continue to believe that same proposition upon learning new information (compatible with prior belief); its just that one may not be able to use the same conditional to express that belief.

Just as the sententialist response to triviality for simple conditionals had a role to play in understanding iterated conditionals, so to does the semantic response to triviality. Gibbard considers but rejects that a mid-sentence context shift might provide a way out of the proof showing that modus ponens and exportation collapse the conditional into the material one. But Harper [44] notes that such a move appears natural, once we relativize Ramsey Test conditionals to doxastic states. The semantic value of a right-embedded conditional will be evaluated relative not to the agent's actual belief state but to the belief state modified by adding the antecedent which the conditional is governed by.

The mid-sentence context shifting move will only provide a solution if at can do at least one of two things. It must render either modus ponens or exportation invalid. Harper's comments suggest that he thinks that modus ponens will fail for iterated conditionals. This wouldn't be needed to block the material analysis if the Ramsey Test conditional did obey exportation, although Harper notes that at least in some cases it does appear to obey the import-export equivalence. Especially, he thinks that $(A \wedge B) > C$ obeys import-export when $A \wedge B$ is compatible with belief. Harper

[15]Lewis attributes this move to Van Fraassen. It is endorsed enthusiastically by Harper, and with reservations by Stalnaker [103].

[44] suspects that natural language conditionals are not assertable when $A \wedge B$ is incompatible with prior belief, although our earlier discussion shows this is wrong. Harper thinks that belief revision itself, as opposed to its linguistic expression in conditionals, treats successive revisions as a conjunction of revisions.

However, if that is the only case where import-export is obeyed, then perhaps we do not need to treat it as part of the logic of conditionals. When $A \wedge B$ is compatible with prior belief, the result of adding it is the same as adding A and then B. So it might seem that we can treat import-export as a pragmatically reasonable inference, in Stalnaker's sense.[16]

But it is not clear whether the import-export law can be reduced to pragmatics, even in this version where the embedded conditional is interpreted relative to its outer antecedent. For one thing, if this is to be solely a story about conditionals which are fully accepted, then it fails to do justice to the intuitions in McGee's examples that are about conditionals which are only partially accepted.[17] For another thing, it fails to do justice to the data about belief-contravening revisions. All of these conditionals appear to obey the import-export equivalence, which suggests that it is a genuinely semantic feature of them.

In any case, whether we say that the import-export law is pragmatic or semantic, the case of belief contravening revisions needs to handle the case we studied above, where A is already accepted but B is incompatible with belief. In the sentential models of belief, we gave up Weak Preservation with respect to conditionals. This move expresses itself differently in the propositional model of the Ramsey Test. It isn't accepted information which isn't preserved upon supposing something already

[16]More exactly, we can treat importation that way under the stronger definition of reasonable inference which Stalnaker gives for contraposition: the antecedent of the conclusion is compatible with belief. (Note that the stronger definition makes superfluous Stalnaker's pragmatic conditions for disjunctions in his treatment of the direct argument.) An even stronger condition, going beyond anything countenanced by Stalnaker, is required for Exportation to be reasonable. For $A > (B > C)$ to be a reasonable conclusion of its imported version, it would require the further condition that $B \supset A$ is accepted.

[17]One might try to extend it to the probabilistic case by means of the law that $P(A/(B \wedge C)) = P((A \wedge B)/C)$ if $P(A) > 0$. Triviality would be avoided here by denying that import-export is logically valid. But Stalnaker [101] show that further triviality ensues if we extend the analysis to Popper functions.

believed. Rather, the semantic value of the conditional (similarity ordering or selection function) changes, despite there being no change in information accepted when the supposition is made. So the semantic value of a conditional is relative to a belief state, it is not solely a function of the informational content of that state.

3.5.6 Modals

Doxastic and epistemic modals make for interesting comparisons with conditionals. In the Ramsey Test literature, the following triviality problem has been noted. At first an agent is undecided whether or not A, hence he accepts both: "it might be that A" and "it might be that $\neg A$". But if the agent learns A, then he accepts "it must be that A", which would make his beliefs inconsistent if he preserved his belief "it might be that $\neg A$".

This has led Rott [89] to conclude that, as with conditionals, modals do not obey the standard laws of belief revision.[18] Rott reaches this conclusion in a sentential model of belief. But from the semantic viewpoint, it is more natural to say that claims of doxastic possibility are indexical, so despite surface appearances they do not conflict. After learning A the agent still believes that $\neg A$ used to be compatible with his beliefs, but he would not say "it might be that $\neg A$ since that expresses compatibility with his current beliefs.

The naturalness of the semantic treatment of modals as indexical suggests that a similar thing should be said about indicative conditionals. And the issue fits together directly with the puzzle of iterated conditionals, when we consider modals embedded in conditionals. Without settling on the exact semantics of "might", let us agree that if it has truth conditions then both of "might A" and "might $\neg A$" are true when said by an agent unsure of A. Likewise, both of these conditionals should be not only acceptable by the Ramsey Test but true: "if A, then it must be that A" and "if $\neg A$, then it must be that $\neg A$". But then, since either A or $\neg A$ is true, one of the conditionals entails the contradiction of one of the modals. So we have a violation of modus ponens for the logic of natural language, with modals instead of conditionals

[18]Levi [56] concludes that, as with conditionals, modals are not objects of belief.

embedded in conditionals.[19]

The way out of the paradox is to say that the unembedded modals are indexical to the current belief state, while the embedded modals are indexical to the supposition states. So they do not contradict each other. Does this resolution of the paradox show that modus ponens really is valid for simple conditionals in natural language? As we noted much earlier in our discussion of McGee's puzzle, there is a sense in which we can say that modus ponens is valid. The proposition expressed by the consequent of a conditional is entailed by the conditional plus its antecedent. But again, this is an entailment obscured by the form of language, since that proposition is not what is expressed (in the same context) by the consequent when it is a standalone sentence.

Indicative conditionals, according to the Ramsey Test theory, are dependent on doxastic and suppositional states. Doxastic modals are even more clearly dependent in this way. In both cases, we can trace the failure of modus ponens to the semantic value of the consequent of a conditional being parasitic on the suppositional state induced by supposing the antecedent. In the case of iterated conditionals, something positive arises out of the negative ashes of modus ponens: the import-export law.

3.6 Subjunctive Conditionals

3.6.1 Extending the Analysis?

So far we have exclusively looked at indicative conditionals, and argued that they do not obey modus ponens for iterated conditionals. But what subjunctive conditionals? Adams [1] argues that modus ponens is a correct rule of inference for indicative conditionals, but not for subjunctive ones. This is right, in a way, if we focus just on the cases he considered involving simple conditionals, and realize that the point is not about the logical validity of modus ponens. What he meant was that the acceptance

[19]When the doxastic modals are replaced with deontic ones, the puzzle is a version of the fatalism dilemma discussed earlier. Stalnaker [99] provides a context shifting solution to that puzzle. Kolodny and MacFarlane [52] offer different solution to the problems of embedded modals and conditionals; like Gillies [36] they relativize all content to information states. It has been noted (by Andrew Bacon in a blog thread initiated by Schulz [95]) that the puzzle will arise for the Kratzer theory of conditionals, where modality is not veridical.

of an indicative conditional never depends on the rejection of its antecedent, while the acceptance of a subjunctive conditional often does. Hence upon learning the antecedent of an indicative, one will never reject the conditional itself; so one will then infer the consequent from the conditional and its antecedent. But when one learns the antecedent of a subjunctive counterfactual—where the consequent is thought false too, as opposed to Goodman's [37] "semi-factuals"—one may give up the conditional itself. In this way, one avoids drawing an unacceptable conclusion by modus ponens.[20]

McGee [73] points out that there are counterfactuals in the subjunctive mood where Import-Export sounds right:

> If Juan hadn't married Xochitl and Sylvia hadn't run off to India, Juan and Sylvia would have become lovers.
> If Juan hadn't married Xochitl, then if Sylvia hadn't run off to India, Juan and Sylvia would have become lovers.

If this were a genuine feature of counterfactuals, and they were uniform in logic with indicative conditionals, then we would have a further objection to Lowe's attempt to save modus ponens (since the material conditional is hopeless as an analysis of the counterfactual conditional). However, the prospects for a subjunctive counterexample to Modus Ponens are not good, and both of Importation and Exportation have counterfactual counterexamples.

3.6.2 Are There Subjunctive Counterexamples to Modus Ponens?

Here is a subjunctive example from Levi [57]. A spinner is divided into equal parts labeled 1, 2, and 3. It is spun and we observe that it lands on 1. Levi claims it is reasonable to assent to "If the spinner had landed on an odd number, then if it hadn't landed on 1 it would have landed on 3". Yet we do not assent to "If it hadn't landed on 1 it would have landed on 3".

[20] As Adams notes elsewhere, counterfactuals are especially apt for modus tollens reasoning.

Given that we know that the spinner did land on an odd number, this example sounds odd since the subjunctive mood connotes counterfactuality.[21] On the other hand, if we had not yet observed where the spinner landed, then the subjunctive of the inner conditional is out of place for the reason that it connotes counterfactuality despite there not yet being any presupposition to suspend about how the spinner landed.

Perhaps the difficulties in getting Levi's example to work the way he intends are artifacts of his particular example. It seems we can do a little better with this variation on the McGee example:

> If a Republican were to win, then Anderson would win if Reagan were to lose.

As in the original indicative example, this sentence seems fine in a context where Reagan is the likely winner, even though we do not detach the consequent "If Reagan were to lose, then Anderson would win". And the iterated subjunctive cannot be given the standard Stalnaker semantics for the same reason as we explained when we studied its indicative counterpart. This example reinforces the often noted similarity between open (non-counterfactual) future-directed subjunctives and indicatives. However, the parallels are not totally clear. Consider:

> If a Republican were to win, then Reagan would win if Anderson were to lose.

This sentence sounds odd in the election scenario, although the corresponding indicative is fine ("If a Republican wins, then Reagan will win if Anderson loses"). The subjunctive seems to say that Anderson's losing is counterfactual with respect to a Republican victory. It is not clear why it should do this, since the antecedent "If a Republican were to win" presumably is not treated as counterfactual. On the other hand, it sounds odd to say these:

[21] We cannot get around this problem to changing the major premise to a subjunctive conditional embedded in an indicative one: "If the spinner landed on an odd number, then if it hadn't landed on 1 it would have landed on 3". Since we already accept the antecedent, the conditional is as uncompelling as its consequent is on its own as the conclusion of Levi's argument.

If Carter were to lose, then Reagan would win.

If Carter were to lose, then Anderson would win if Reagan were to lose.

This suggests an asymmetry in the use of the subjunctive, where in some cases but not others one may suppose the truth of something which is thought to be likely anyway. This makes it difficult to sort out what the proper analysis of open subjunctives is.

Things become more clear when we turn to the classic case of the subjunctive, the counterfactual conditional. Indeed, an example of McGee's form cannot be constructed where the major premise is thought to be counterfactual. For it the antecedent of the major premise is thought to be likely false, then we will not have high probability in the minor premise. And therefore there will be no grounds for acceptance of the consequent being supported by modus ponens. So modus ponens is not impugned by our not accepting the consequent in that scenario.

3.6.3 Counterfactual Counterexamples to Import-Export

Although it is difficult to construct subjunctive counterexamples to Modus Ponens, and impossible to construct counterfactual ones, there are counterfactual counterexamples to the Import-Export equivalence. Imagine that the spinner is observed to land on 2, so it is acceptable to say, "If it had landed on an odd number but not 1, then it would have been 3". Yet it is odd to say, "If it had landed on an odd number, then if it had not landed on 1 it would have been 3". For the inner antecedent inappropriately connotes that landing on 1 is counterfactual with respect to the counterfactual claim that the spinner landed on an odd number.

We can bring out the problem with counterfactual by means of some iterated conditionals that are easier to interpret. Especially with some cues using the word "even", we can use iterated subjunctive conditionals to make counterfactual suppositions which are then contravened by further subjunctive suppositions. Consider saying of a match which is known to be normal, dry, unstruck:

> Even if this match had lit at noon today, it would not have done so if it had been soaked in water last night.

> If this match had lit at noon today, then even if it had been soaked in water last night it would have lit at noon today.

The first sounds true, while the second sounds false. Each conditional asks us to imagine that the match had been struck and lit (which is assumed possible since the match is normal and dry); then holding fixed its having been struck, it asks whether it would have lit if it had been previously soaked in water. The second conditional (incorrectly) affirms, while the first conditional (correctly) denies, that it would have lit. They are therefore counterexamples to Importation and Exportation, respectively, in relation to the following:

> If this match had lit at noon today and had been soaked in water last night, then it would have not lit at noon today.
> If this match had lit at noon today and had been soaked in water last night, then it would have lit at noon today.

For these sentences are, respectively, logically false and true (as instances of $A \wedge B > \neg A$ and $A \wedge B > A$).[22]

The only way for McGee to get out of these counterexamples is to to say that we do not interpret the embedded sentences as conditionals, but as non-conditional. A move along these lines has been made by others with respect to left-nested conditionals such as:

> If this vase were to break if thrown against the ground, then this vase would break if thrown against the wall.

[22] Although I judge that the counterexamples work without the "even", some (such as Schulz [95]) disagree and would place the blame there. Skyrms [97, p. 176] presents parallel indicative and subjunctive conditionals which both seem to obey the import-export equivalence, but says: "Nevertheless, in the subjunctive case (though not in the indicative) I can imagine appropriate promptings and side remarks that would lead me to take some variant of the counterfactual in the Stalnaker way:

> If this sample were burning green (say it was barium) then it would still be true that had it been sodium if would have burned yellow.

The question of what cues in English lead you to take a counterfactual one way rather the other is, I think, a very complicated business."

Gibbard claims that we should regard the embedded conditionals as expressing categorical claims about dispositions. If McGee adopted such a strategy, he could say that we cannot apply Importation and Exportation to the problematic sentences about the match. These iterated counterfactual conditionals are interpreted, ad hoc, as the regular Stalnaker semantics would treat them.

3.7 Conclusion

We have seen that the data supports a distinction between indicative and subjunctive counterfactual conditionals. The latter obey the standard possible worlds semantics of Stalnaker's theory [102], while the former obey McGee [73] modified version. Indicative conditionals, but not subjunctives, obey the Import-Export law. Subjunctive counterfactuals, but not indicatives, obey modus ponens without restriction. Open subjunctives fall in a difficult to classify category, bearing strong but not total resemblance to indicatives.

With respect to the Ramsey Test, we have seen that its converse does not hold for iterated conditionals, and that we should give up the Weak Preservation law with respect to conditionals. Additionally, by treating indicatives as relative to doxastic and suppositional states, the Ramsey Test helps explain the failure of modus ponens for iterated indicatives. The phenomenon is related to similar failings of modus ponens for conditionals containing modals in their consequents.

Bibliography

[1] Ernest Adams. Subjunctive and indicative conditionals. *Foundations of Language*, 6:89–94, 1970.

[2] Ernest Adams. *The Logic of Conditionals.* Reidel, Dordrecht, 1975.

[3] Ernest Adams. Review of Frank Jackson *Conditionals*. *Philosophical Review*, 99:433–435, 1990.

[4] J. E. J. Altham. The legacy of emotivism. In Graham McDonald and Crispin Wright, editors, *Fact, Science, and Morality.* Blackwell, 1986.

[5] Paul Anand. *Foundations of Rational Choice Under Risk.* Oxford, 1993.

[6] G. E. M. Anscombe. *Intention.* Blackwell, Oxford, 1957.

[7] Anthony Appiah. *Assertion and Conditionals.* Cambridge, Cambridge, 1985.

[8] Horacio Arló Costa. Conditionals and monotonic belief revisions: The success postulate. *Studia Logica*, XLIX:557–566, 1990.

[9] Horacio Arló Costa. Belief revision conditionals: basic iterated systems. *Annals of Pure and Applied Logic*, 96:3–28, 1999.

[10] Horacio Arló Costa. Bayesian epistemology and epistemic conditionals: On the status of the export-import laws. *Journal of Philosophy*, 98:555–593, 2001.

[11] W. E. Armstrong. The determinateness of the utility function. *Economic Journal*, 49:453–467, 1939.

[12] Kenneth Arrow. *Social Choice and Individual Values.* Wiley, 2nd edition, 1963.

[13] Richard Bradley. A representation theorem for a decision theory with conditionals. *Synthese*, 117:187–229, 1998.

[14] Richard Bradley. A defence of the Ramsey test. *Mind*, 116:1–21, 2007.

[15] John Broome. Desire, belief and expectation. *Mind*, 100:265–267, 1991.

[16] John Broome. Can a Humean be moderate? In *Ethics Out of Economics.* Cambridge, Cambridge, 1999.

[17] Alex Byrne and Alan Hájek. David Hume, David Lewis, and decision theory. *Mind*, 106:411–428, 1997.

[18] Ruth Chang, editor. *Incommensurability, Incomparability, and Practical Reason.* Harvard, Cambridge, 1997.

[19] John Collins. Belief, desire, and revision. *Mind*, 97:333–342, 1988.

[20] John Collins. *Belief Revision.* PhD thesis, Princeton, 1991.

[21] Charles Cross and Richmond Thomason. Conditionals and knowledge-base update. In Peter Gärdenfors, editor, *Belief Revision.* Cambridge, Cambridge, 1992.

[22] Stephen Darwall. *Impartial Reason.* Cornell, Ithaca, 1983.

[23] Donald Davidson. Psychology as philosophy. In *Essays on Actions and Events.* Oxford, Oxford, 1980.

[24] Donald Davidson. Paradoxes of irrationality. In *Paradoxes of Rationality* [26].

[25] Donald Davidson. Incoherence and irrationality. In *Paradoxes of Rationality* [26].

[26] Donald Davidson. *Paradoxes of Rationality.* Oxford, Oxford, 2004.

[27] J. J. C. McKinsey Davidson, Donald and Patrick Suppes. Outlines of a formal theory of value, I. *Philosophy of Science*, 22:140–160, 1955.

[28] James Dreier. Internalism and speaker relativism. *Ethics*, 101:6–26, 1990.

[29] Ellery Eells and Brian Skyrms, editors. *Probability and Conditionals.* Cambridge, Cambridge, 1994.

[30] David Etlin. The problem of noncounterfactual conditionals. *Philosophy of Science*, Forthcoming 2009.

[31] Nir Friedman and Joseph Halpern. Belief revision: A critique. *Journal of Logic, Language and Information*, 8:401–420, 1999.

[32] Peter Gärdenfors. Belief revisions and the Ramsey test for conditionals. *Philosophical Review*, 95:81–93, 1986.

[33] Peter Gärdenfors. *Knowledge in Flux.* MIT, Cambridge, 1988.

[34] Allan Gibbard. Two recent theories of conditionals. In Harper et al. [46].

[35] Allan Gibbard and William Harper. Counterfactuals and two kinds of expected utility. In Harper et al. [46].

[36] Anthony Gillies. Epistemic conditionals and conditional epistemics. *Noûs*, 38:585–616, 2004.

[37] Nelson Goodman. *Fact, Fiction, Forecast.* Harvard, Cambridge, 1955.

[38] H. P. Grice. *Studies In The Ways of Words.* Harvard, Cambridge, 1989.

[39] Alan Hájek and Ned Hall. The hypothesis of the conditional construal of conditional probability. In Eells and Skyrms [29].

[40] Alan Hájek and Philip Pettit. Desire beyond belief. *Australasian Journal of Philosophy*, 82:77–92, 2004.

[41] Sven Ove Hansson and Tille Grünne-Yanoff. Preferences. *Stanford Encyclopedia of Philosophy*, `http://plato.stanford.edu/entries/preferences/`, 2006.

[42] William Harper. Ramsey test conditionals and iterated belief change. In Harper and Hooker [45], pages 117–135.

[43] William Harper. Rational belief change, Popper functions, and counterfactuals. In Harper and Hooker [45], pages 73–111.

[44] William Harper. A sketch of some recent developments in the theory of conditionals. In Harper et al. [46].

[45] William Harper and C. A. Hooker, editors. *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*, volume 1. Reidel, Dordrecht, 1976.

[46] William Harper, Robert Stalnaker, and Glenn Pearce, editors. *Ifs.* Reidel, Dordrecht, 1981.

[47] Thomas Hobbes. Of liberty and necessity. In W. Molesworth, editor, *The English Works of Thomas Hobbes*, volume 4. J. Bohn, London, 1839.

[48] Frank Jackson. *Conditionals.* Blackwell, Oxford, 1987.

[49] Richard Jeffrey. Ethics and the logic of decision. *Journal of Philosophy*, 62:528–539, 1965.

[50] Richard Jeffrey. *The Logic of Decision.* University of Chicago, 2nd edition, 1983.

[51] Bernard Katz. On a supposed counterexample to modus ponens. *Journal of Philosophy*, 96:404–415, 1999.

[52] Niko Kolodny and John MacFarlane. Ifs and oughts. Manuscript online `http://johnmacfarlane.net/`, August 11, 2008.

[53] Michael Kremer. 'If' is unambiguous. *Nous*, 21:199–217, 1987.

[54] Keith Lehrer and Carl Wagner. Intransitive indifference: The semi-order problem. *Synthese*, 65:249–56, 1985.

[55] Isaac Levi. *Hard Choices.* Cambridge, Cambridge, 1986.

[56] Isaac Levi. Iteration of conditionals and the Ramsey test. *Synthese*, 76:49–81, 1988.

[57] Isaac Levi. *For the Sake of Argument.* Cambridge, Cambridge, 1996.

[58] David Lewis. *Counterfactuals.* Harvard, Cambridge, 1973.

[59] David Lewis. Causal decision theory. In *Philosophical Papers* [61].

[60] David Lewis. Counterfactual dependence and time's arrow. In *Philosophical Papers* [61].

[61] David Lewis. *Philosophical Papers*, volume 2. Oxford, Oxford, 1986.

[62] David Lewis. Probabilities of conditionals and conditional probabilities. In *Philosophical Papers* [61].

[63] David Lewis. Probabilities of conditionals and conditional probabilities II. *Philosophical Review*, 95:581–589, 1986.

[64] David Lewis. Desire as belief. In *Papers in Ethics and Social Philosophy* [67].

[65] David Lewis. Desire as belief II. In *Papers in Ethics and Social Philosophy* [67].

[66] David Lewis. Dispositional theories of value. In *Papers in Ethics and Social Philosophy* [67].

[67] David Lewis. *Papers in Ethics and Social Philosophy.* Cambridge, Cambridge, 1999.

[68] E. J. Lowe. Not a counterexample to modus ponens. *Analysis*, 47:44–47, 1987.

[69] R. Duncan Luce. Semiorders and a theory of utility discrimination. *Econometrica*, 24:178–91, 1956.

[70] R. Duncan Luce and Howard Raiffa. *Games and Decisions.* Wiley, New York, 1957.

[71] Patrick Maher. *Betting on Theories.* Cambridge, Cambridge, 1993.

[72] Michael Mandler. A difficult choice in preference theory: Rationality implies completeness or transitivity but not both. In Elijah Millgram, editor, *Varieties of Practical Reasoning.* MIT, Cambridge, 2001.

[73] Vann McGee. A counterexample to modus ponens. *Journal of Philosophy*, 82:462–471, 1985.

[74] Vann McGee. Conditional probabilities and compounds of conditionals. *Philosophical Review*, 98:485–541, 1989.

[75] Yew-Kwang Ng. Sub-semiorder: A model of multidimensional choice with preference intransitivity. *Journal of Mathematical Psychology*, 16:51–59, 1977.

[76] Graham Oddie. Harmony, purity, truth. *Mind*, 412:451–472, 1994.

[77] Graham Oddie. Hume, the BAD paradox, and value realism. *PHILO*, 4:102–22, 2001.

[78] Doris Olin. Newcomb's problem: Further investigations. *American Philosophical Quarterly*, 13:129–133, 1976.

[79] D. E. Over. Assumptions and the supposed counterexamples to modus ponens. *Analysis*, 47:142–146, 1987.

[80] Philip Pettit. Humeans, anti-humeans, and motivation. *Mind*, 96:530–533, 1987.

[81] Charles Plott. Path independence, rationality, and social choice. *Econometrica*, 41:1075–1091, 1973.

[82] Karl Popper. *The Logic of Scientific Discovery.* Basic Books, New York, 1959.

[83] Huw Price. Defending desire-as-belief. *Mind*, 98:119–127, 1989.

[84] Howard Raiffa. *Decision Analysis.* Addison-Wesley, Reading, MA, 1968.

[85] F. P. Ramsey. General propositions and causality. In *Philosophical Papers.* Cambridge, Cambridge, 1990.

[86] Joseph Raz. *The Morality of Freedom.* Oxford, Oxford, 1986.

[87] Martin Rechenauer. On the non-equivalence of weak and strict preference. *Mathematical Social Sciences*, 2008.

[88] Lionel Robbins. *An Essay on the Nature and Significance of Economic Science.* Macmillan, London, 2nd edition, 1935.

[89] Hans Rott. Conditionals and theory change: Revisions, expansions, and additions. *Synthese*, 81:91–113, 1989.

[90] Leonard Savage. *The Foundations of Statistics.* Wiley, New York, 1956.

[91] Frederic Schick. Dutch books and money pumps. *Journal of Philosophy*, 83:112–119, 1986.

[92] George Schumm. Transitivity, preference and indifference. *Philosophical Studies*, 52:435–437, 1987.

[93] Thomas Schwartz. Rationality and the myth of the maximum. *Noûs*, 6:97–117, 1972.

[94] Amartya Sen. *Choice, Welfare, and Measurement.* MIT, 1982.

[95] Moritz Shulz. Do counterfactuals violate modus ponens?, 2008. Blog entry at `http://eppe.wordpress.com/2008/08/09/do-counterfactuals-violate-modus-ponens/`.

[96] Walter Sinnott-Armstrong, James Moor, and Robert Fogelin. A defence of modus ponens. *Journal of Philosophy*, 83:296–300, 1986.

[97] Brian Skyrms. *Causal Necessity.* Yale, New Haven, 1980.

[98] Michael Smith. *The Moral Problem.* Blackwell, Oxford, 1993.

[99] Robert Stalnaker. Indicative conditionals. In Harper et al. [46].

[100] Robert Stalnaker. Letter to David Lewis. In Harper et al. [46].

[101] Robert Stalnaker. Probability and conditionals. In Harper et al. [46].

[102] Robert Stalnaker. A theory of conditionals. In Harper et al. [46].

[103] Robert Stalnaker. *Inquiry.* MIT, Cambridge, 1984.

[104] Kotaro Suzumura. *Rational Choice, Collective Decisions, and Social Welfare.* Cambridge, Cambridge, 1983.

[105] Amos Tversky. Intransitivity of preferences. In Eldar Shafir, editor, *Preference, Belief and Similarity: Selected Writings of Amos Tversky.* MIT, 2004.

[106] Edna Ullman-Margalit and Sidney Morgenbesser. Picking and choosing. *Social Research*, 44:757–85, 1977.

[107] R. Jay Wallace. How to argue about practical reason. *Mind*, 99:355–385, 1990.

[108] Ruth Weintraub. Desire as belief, Lewis notwithstanding. *Analysis*, 67:116–122, 2007.

[109] David Wiggins. *Needs, Values, Truth.* Oxford, Oxford, 3rd edition, 1998.