London School of Economics and Political Science

Decision Theory and Counter-Factual Evaluation

H. Orri Stefánsson

A thesis submitted to the Department of Philosophy, Logic and Scientific Method
of the London School of Economics and Political Science
for the degree of Doctor of Philosophy, April 2014.
Final, revised version submitted in October 2014.

Declaration

I certify that the thesis I have presented for examination for the MPhil/PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it). The copyright of this thesis rests with the author. Quotation from it is permitted, provided that full acknowledgement is made. This thesis may not be reproduced without my prior written consent. I warrant that this authorisation does not, to the best of my belief, infringe the rights of any third party.

I declare that my thesis consists of 80,754 words.

Statement of Conjoint Work

I confirm that section 1.5 is based on a working paper that is co-authored with Richard Bradley, and that the whole of Chapter 4 is also based on a working paper co-authored with Bradley.

H. Orri Stefánsson

Abstract

The value of actual outcomes or states of affairs often depends on what could have been. Such dependencies create well-known "paradoxes" for decision theory, the best-known perhaps being the so-called *Allais Paradox*. The primary aim of this PhD thesis is to enrich decision theory such that it includes counterfactual prospects in the domains of desirability (or utility) functions, and show that, as a result, the paradoxes in question disappear.

Before discussing the way in which counterfactual propositions influence the desirability of actual outcomes, I discuss the way in which the truth of one factual proposition influences the desirability of another. This examination leads me to reject the *Invariance* assumption, which states that the desirability of a proposition is independent of whether it is true. The assumption plays an important role in David Lewis' famous arguments against the so-called *Desire-as-Belief* thesis (DAB). The unsoundness of Lewis' argument does of course not make DAB true. In fact, I provide novel arguments against different versions of DAB, without assuming Invariance.

To justify the assumptions I make when extending decision theory to counterfactual prospects, I discuss several issues concerning the logic, metaphysics and epistemology of counterfactuals. For instance, I defend a version of the so-called *Ramsey test*, and show that Richard Bradley's recent *Multidimensional Possible World Semantics for Conditionals* is both more plausible and permissive than Bradley's original formulation of it suggested.

I use the multidimensional semantics to extend Richard Jeffrey's decision theory to counterfactuals, and show that his desirability measure, extended to counterfactuals, can represent the various different ways in which counterfactuals influence the desirability of factual propositions. And I explain why the most common alternatives to Jeffrey's theory cannot be similarly extended.

I conclude the thesis by using Jeffrey's extended decision theory to construct an ethical theory I call *Modal Consequentialism*, and argue that it better satisfies certain entrenched moral intuitions than Non-Modal Consequentialism (such as classical utilitarianism and welfare economics).

Contents

Acknowledgements

Several people have helped me improve this thesis. I should first mention my primary supervisor, Richard Bradley. My thesis would have been considerably poorer if it hadn't been for his very detailed and constructive comments on all of its parts, and the many hours we spent discussing decision theory and the logic of conditionals. I would also like to thank my secondary supervisor, Katie Steele, whose comments and criticism proved especially useful at times when I was uncritically adopting all of Richard's philosophical views, and my two examiners, Jim Joyce and Wlodek Rabinowicz, with whom I had incredibly enjoyable and very educational exchanges about the topics of this dissertation.

I have benefited from discussing the topics of this thesis with several other teachers I had during my PhD years. David Makinson provided very useful comments on the chapters on the logic and semantics of conditionals. I had fruitful discussions about the value of lotteries with Alex Voorhoeve, and the philosophy of probability and chance with Roman Frigg. Exchanges with Christian List were influential on my discussion of Invariance and the Desire-as-Belief thesis. And a term spent visiting Alan Hájek at the Australian National University was extremely educational.

I am grateful for a four year scholarship from the LSE's Department of Philosophy, Logic and Scientific Method's that enabled me to write this thesis.

My fellow PhD students, both those I studied with at the LSE and those I met at various graduate conferences, taught me almost as much as my formal teachers. These people are too numerous to mention each of them individually. However, I should offer special thanks to Benjamin Ferguson, who proofread this whole thesis, and corrected a great number of grammatical mistakes and typos. Both the audience and commentators at various conferences where I have presented the ideas of this thesis have helped me to improve both its content and presentation. I would in particular like to mention Branden Fitelson, Campell Brown, Robbie Williams, Julian Jonker and Lara Buchak.

I am grateful to my parents, for encouraging me to pursue what I am passionate about, even if it means an insecure and badly paid career in academic philosophy. I am also grateful to them, and to my brothers, for patiently listening to my philosophical ramblings over the years. Most importantly, I would like to thank Ösp, for her love and support, and for creating the perfect environment for writing a thesis.

Introduction: Counterfactual Desirability and Practical Rationality

0.1 Desirability and Counterfactuals

The desirability of what actually occurs is often influenced by what *could have been*. Suppose you have been offered two jobs, one very exciting but with a substantial risk of unemployment, the other less exciting but more secure. If you choose the more risky option, and as a result become unemployed, you might find that the fact that you *could have* chosen the risk-free alternative makes being unemployed even worse. Dependencies of this kind between what is and what could have been generate well-known paradoxes for the traditional theory of rational choice, as for instance formulated by John von Neumann and Oskar Morgenstern [von Neumann and Morgenstern, 1944], Leonard Savage [Savage, 1972] and Richard Jeffrey [Jeffrey, 1983].

The above example is a simplified version of Maurice Allais' infamous paradox [Allais, 1953], [Allais, 1979], which has troubled decision theorists for decades. The paradox is generated by offering people a pair of choices between different lotteries, each of which consists in tickets being randomly drawn. First people are offered a choice between a 'lottery' that is *certain* to result in the decision maker receiving a particular prize, say £2400, and a lottery that could result in the decision-maker receiving nothing, but could also result in the decision maker receiving more than £2400. The situation can be represented as having to choose between lotteries L_1 and L_2 below, where, for instance, L_1 results in the decision maker receiving a prize of £2500 if one of tickets number 2 to 34 is drawn:

	1	2 – 34	35 – 100
L_1	£0	£2500	£2400
L_2	£2400	£2400	£2400

Having made a choice between L_1 and L_2, people are asked to make another choice which we can represent as a choice between lotteries L_3 and L_4:

	1	2 – 34	35 – 100
L_3	£0	£2500	£0
L_4	£2400	£2400	£0

Although the results are not conclusive, it seems that people tend to choose, and *strictly* prefer, L_2 over L_1 and L_3 over L_4. (See [Kahneman and Tversky, 1979] for discussion of an early experiment of the Allais Paradox.) One common way to rationalise this preference,[1]

[1] I should emphasise that the reasoning I am about to suggest is by no means the only reasoning that might generate the preference in question. Some for instance think that what generates Allais' preference is not regret aversion, but rather *risk* aversion. Personally, I think that the reason I myself have the Allais preference has more

which I will call 'Allais' preference', is that when choosing between L_1 and L_2, the chance of receiving the higher prize as a result of choosing L_1 is not worth the risk of ending up with nothing, since receiving nothing when you *could have had* £2400 for sure is bound to cause considerable regret (see e.g. [Loomes and Sugden, 1982] and [Broome, 1991]). When it comes to choosing between L_3 and L_4, however, the desire to avoid regret does not play as strong role, since decision makers reason that if they choose L_3 and end up with nothing then they would, in all likelihood, have received nothing even if they had chosen the less risky option L_4.[2]

As Wlodek Rabinowicz has pointed out to me, examples like Allais' are particularly interesting if we assume that the agent in question knows that she will not remember having had the opportunity to choose a different option than the one she does, and will thus not experience the regret, but nevertheless makes choices that are directed at avoiding this type of counterfactual dependencies. In such cases, one could argue that the agent perceives these dependencies as being bad independently of the feeling of regret, and that the regret she would feel if she were to remember the opportunities she had *reflect* rather than *cause* the badness of the situation. And that is roughly how I will, in what follows, interpret the role of rational regret in this examples; i.e., as a reflection rather than cause of disvalue.

The discussed value dependency between counterfactual and actual outcomes does not always give rise to regret. Suppose a hospital has a single kidney but two equally needing and deserving patients, Ann and Bob. Furthermore, assume that Ann and Bob's situation is symmetric in all respects that are relevant for deciding who should receive the kidney. How should we decide who gets the kidney? Most people have a strong intuition that in situations like these we should toss a coin, or hold some other lottery that gives each patient an equal chance of receiving the kidney. And they find that such a lottery is *strictly* better than giving the kidney to either Ann or Bob without holding a lottery. In other words, if *ANN* (*BOB*) represents a situation (or outcome) where Ann (Bob) has received the kidney, and H and T partition the space of possibilities into two equiprobable and random events (e.g. a coin coming heads/tails up), most people are indifferent between F_1 and F_2, and also between A and B, but strictly prefer the two 'fair' lotteries to the two biased alternatives:

	H	T
F_1	*ANN*	*BOB*
A	*ANN*	*ANN*
B	*BOB*	*BOB*
F_2	*BOB*	*ANN*

I will call this judgement *Diamond's preference*, after Peter Diamond who in [Diamond, 1967]

to do with regret aversion than risk aversion. But others might have the same preference due to risk aversion. Those who think decision theory should capture, in some very minimal sense, the reasoning behind people's preferences, and moreover think that risk aversion is a more common explanation for Allais-type preferences than regret aversion, might prefer Lara Buchak's solution to the Allais paradox to mine (see [Buchak, 2013], and [Stefánsson, 2014e] for a discussion of Buchak's book). For the remainder of this thesis, I will however assume that the reasoning I suggest explains the Allais' preference (alternatively, my discussion of Allais' preference can be seen as being limited to those who have this preference for the reason I suggest).

[2]Notice, however, that according to the *Minimax Regret Rule* (MRR) we should choose L_4 over L_3 (in contradiction with Allais' preference). The rule tells us to look at each state s_i and determine, for each alternative A, the potential for regret in that state, as measured by the difference between the prize that A gives in s_i and the highest prize you could receive in s_i (i.e. the outcome of the alternative that is most favourable in state s_i). When reasoning in the way I am suggesting, however, agents do not try to figure out the potential for regret state-by-state. Rather, people seem to reason that even if they choose the slightly more risky option in the second choice situation and get ticket 1, they will 'forgive themselves' for having taken the slight extra risk since they know that both choices were quite risky anyway. (Reasoning of this latter kind is based on aversion to what Luc Bovens and Wlodek Rabinowicz call *probability action* regret [Bovens and Rabinowicz, ms].)

first pointed out that this intuitively reasonable preference causes problems for expected utility theory (EU theory). Here is a possible consequentialist justification for this preference (which I defend in chapter 6 and [Stefánsson, 2014b]). Suppose we find ourselves in a situation S where Ann has received the kidney as a result of a lottery. Then unlike a situation where Ann is given the kidney without any lottery being held, it is true in S that although Bob didn't receive a kidney, he at least *had a chance* of receiving it. And that means, as I understand it, that things could, in some meaningful sense, have turned out differently, and if they had, then Bob would have received the kidney. So an outcome, or a situation,[3] where Bob is dead as a result of not having received the kidney is somehow made (morally) better by the truth of this counterfactual.

So in both of the above examples it is the case that the desirability of what actually occurs is affected by what could have been. These are the two examples that will be the focus of this thesis. But I should mention that there are multiple other examples where the truth of a counterfactual makes a desirabilistic difference to an actual outcome or states of affairs. Richard Arneson [Arneson, 1990] and John C. Harsanyi [Harsanyi, 1977] have for instance both argued that the moral value of a public policy depends not on how well it satisfies people's actual preferences, but the preferences that people *would have* in certain ideal circumstances. Similarly, some moral and political philosophers argue that whether or not someone is being exploited in a transaction depends not only on what she actually receives from the transaction, but rather on whether what she receives matches up with what she *would have* received from the transaction in a morally acceptable world, where, for instance, her rights have not been violated (see for instance [Ferguson, 2013]).

Both Allais' and Diamond's preferences cause trouble for orthodox decision theory (i.e., expected utility theory). As I explain more formally in chapter 4, there is no pair of utility and probability functions relative to which either preference can be represented as maximising expected utility. It is standardly assumed, at least amongst decision theorists and economists, that a necessary requirement for a preference to be (practically) rational is that it be possible to represent it as maximising expected utility. Hence, according to the standard picture, both preferences must be irrational.

Contrary to what the standard picture suggests, I think there is nothing irrational about Allais' and Diamond's preferences (and will in next section briefly explain why). However, I do agree with the standard view that a necessary requirement for a preference to be rational is that it be possible to represent it as maximising the expectation of *some* value function.[4] The main aim of this thesis is to develop a new decision theory that differs from standard EU theory in that allows us to express desirabilistic dependencies between actual and counterfactual outcomes, and, as a result, makes it possible to represent both Allais' and Diamond's preferences as maximising the expected value of a particular value function. The new decision theory, which I discuss in chapter 4 and develop jointly with Richard Bradley in [Bradley and Stefánsson, 2015], is an extension of Richard Jeffrey's [Jeffrey, 1983] decision theory to *counterfactual* prospects. (As I explain in chapter 4, extending expected utility theory, such as Leonard Savage's [Savage, 1972], to counterfactual prospects has

[3] Throughout this thesis I will talk about 'outcomes', 'situations' and 'consequences' more or less interchangeably, and take them to mean a result of some alternative being realised (or action being performed, choice being made, etc.)

[4] A note on terminology: when I speak of a 'value function', I mean just any mathematical function that represents how good something is (usually according to some particular agent). An expected utility function and a Jeffrey-desirability function are two specific value functions, with some important formal differences (as will become clear in chapter 4).

very counterintuitive consequences.) So the value function that can represent both Allais' and Diamond's preferences, is a Jeffrey-desirability function defined on a set of factual and counterfactual propositions.

Before discussing the new decision theory, we must clarify certain issues regarding how to update desirability functions, what epistemic and logical principles we think counterfactuals satisfy, and what formal theory of counterfactuals to use to introduce counterfactuals to Jeffrey's theory. Chapters 1, 2 and 3 are devoted to these issues.

0.2 A Humean View on Practical Rationality

In this thesis I assume what I call a *Humean* view on (practical) rationality. The view includes both a negative and positive thesis. The negative thesis is summed up in the following often-quoted passage of David Hume's *Treatise of Human Nature*:

> 'Tis not contrary to reason to prefer the destruction of the whole world to the scratching of my finger. 'Tis not contrary to reason for me to chuse my total ruin, to prevent the least uneasiness of an Indian or person wholly unknown to me. 'Tis as little contrary to reason to prefer even my own acknowledg'd lesser good to my greater, and have a more ardent affection for the former than the latter. In short, a passion must be accompany'd with some false judgement, in order to its being unreasonable; and even then 'tis not the passion, properly speaking, which is unreasonable, but the judgement. ([Hume, 1740]: 267)

When Hume says that something is 'not contrary to reason,' he is, as I understand him, simply saying that this something is not *irrational*. In a more modern terminology, Hume's infamous claim is that the contents of our desires are neither rational nor irrational. A desire, and thus a preference,[5] may be *mistaken* when it is based on a false belief. But even in such cases, it is, strictly speaking, the *belief*, rather than the *desire*, that is at fault. What we might however call *primitive* desires, i.e. those desires that are not based on any beliefs, cannot be mistaken.

The positive thesis of Humeanism says that given any two alternatives, a practically rational person always prefers the alternative she expect to better satisfy her desires. (Perhaps we should call this *Moderate* Humeanism, to distinguish it from the more extreme view according to which there is no such thing as *practical* irrationality.) This in turn implies that the preferences of a practically rational person satisfy certain *axioms*, which we can understand as guaranteeing that the person's attitudes are internally consistent. Some of these axioms are relatively uncontroversial, such as *Transitivity*, which requires that if a person prefers A to B and B to C, then she must prefer A to C. Other axioms are much more controversial, such as *Completeness*, which requires that for any two alternatives A and B, a rational person either prefers A to B, or B to A, or is indifferent between the two. In other words, any two alternatives must be comparable with respect to the agent's preferences. The axiom that will be of particular interest to this thesis is *Separability*, which comes in various strengths and forms, but is perhaps best-known as Savage's *Sure Thing Principle*. No preference can

[5] A note on terminology: I take a person's preferences to be her overall, comparative judgement of the alternatives that she is or could be faced with; which, if the person is rational, orders the alternatives from (in her view) 'best' to 'worst'. And I assume that a person's preferences are determined by her desires and beliefs in the standard way. For instance, although my strongest desire at the moment might be that I go to a sunny beach, I might nevertheless prefer staying in to going to the beach, since I believe that it will rain and find it more desirable to stay in than going to the beach when it is raining.

be represented as maximising expected utility unless it satisfies Separability. In chapter 4, I formally state Separability and explain how it clashes with Allais- and Diamond-style preferences. I will argue that the reasonableness of these two preferences suggests that Separability is not a genuine requirement of rationality. And in fact, like Jeffrey's original theory, the new decision theory I discuss does not require Separability (in the strong form required by EU theory.)

Given the above view on practical rationality, neither Allais' nor Diamond's preference seems to be irrational. Both seem to be preferring the alternative that can be expected to best satisfy their desires. Moreover, given the explanation I have given for the two preferences, it seems we can give plausible reasons for these preferences. People who are prone to regret that they would like to avoid, but, other things being equal, desire more money to less, seem to have a good reason for preferring lottery L_2 to L_1 but L_3 to L_4. In other words, it seems perfectly possible that someone with Allais' preference is preferring the alternatives that best satisfy her desires. Similarly, it seems that someone who is motivated by fairness has a very good reason for being indifferent between F_1 and F_2, and also between A and B, while strictly preferring the first two alternatives to the latter two. So it also seems possible that Diamond prefers the alternatives that best satisfies his desires.

If to prefer the alternatives that one expects to best satisfy ones desires simply means having preferences that can be represented as maximising expected utility, then given how Allais' and Diamond's preferences have been described, they do *not* satisfy the requirement of Humeanism, contrary to the intuition appealed to in last paragraph. But the intuition is, I think, quite strong: it seems hard to deny that given my suggested reasoning behind the two preferences, Allais and Diamond *are* preferring the alternatives that can be expected to maximally satisfy their desires. So EU theory does not perfectly capture the positive part of Humeanism. As already mentioned, one aim of this thesis is to offer a decision theory that does better than standard EU theory in this regard. And I discuss some new results that, I contend, undermines the view that EU theory is our best theory of practical rationality.

A common response to the problems Diamond's and Allais' preferences create for EU theory is that we should simply re-describing the relevant outcomes, e.g. in the way John Broome [Broome, 1991] suggests. Given how intuitive Broome's solution may be, I will briefly responding to it at this point, but will expand on this response both in chapter 4 and 6. We might for instance describe the £0 outcome of L_1 in Allais' example as '£0 + regret', and the outcomes of the fair lottery in Diamond's example as e.g. '*ANN* + fairness'. And then these two preferences no longer cause problems for EU theory (as explained in chapter 6). If descriptions of outcomes should contain everything that is important for their evaluation (as is necessary when e.g. using Savage's decision theory), then these re-descriptions should perhaps be seen as improvements.

There are various theoretical disadvantages of this approach. One technical disadvantage is the tension with the so-called *Rectangular Field Assumption*, which I discuss in chapter 6. A more intuitive problem with this approach, is that unless we have a principled way of distinguishing legitimate from non-legitimate re-descriptions of consequences, this approach threatens to make decision theory 'empty' as a theory of practical rationality, since any preference can be made consistent with the axioms of decision theory if we are clever enough in how we describe the consequences.

Perhaps the above problems can be solved. Broome and Philip Pettit, for instance, attempt to solve the second problem mentioned above – the danger of an empty decision

theory – by respectively suggesting that alternatives should be considered different (and thus described in a way that distinguishes them from each other) just in case they differ in a way that makes it rational to have a preference between them [Broome, 1991], or if they differ in properties that are desired or undesired by the agent whose preferences we are trying to represent [Pettit, 1991]. The main problem I have with this re-description strategy, and the main reason I prefer the solution offered in chapter 4, is that it does not explicitly model the desirabilistic dependencies between actual and counterfactual outcomes that, I contend, are at the heart of the preferences under discussion.

My solution moreover differs from the re-description solution in that rather than adding 'regret' and 'fairness' as primitive properties to our set of outcomes, I show that if we add counterfactual prospects to the domain of a decision theoretic value function, these properties emerge as a relationship between counterfactual and actual outcomes. Thus, I claim, my preferred solution explains why Diamond's and Allais' preferences have caused such problems for decision theorists, who have traditionally been very reluctant to admit that the value of actual outcomes can rationally depend on what merely could have been (as I further explain in chapter 6).

0.3 Summary of Chapters

In addition to this introduction, the thesis comprises six chapters. Here is a summary of each chapter, in a slightly more detail than my discussion above:

Before determining how counterfactual propositions rationally influence the desirability of actual states of affairs, we should clarify how factual propositions rationally influence such desirability. To this end I defend in chapter 1 a particular measure of *conditional desirability*, and discuss its advantages over the measure favoured for instance by David Lewis and James Joyce. The measure I defend contradicts the so-called *Invariance* assumption, according to which the desirability of a proposition is independent of whether it is true. Invariance plays an important role in Lewis' famous arguments against the so-called Desire-as-Belief thesis (DAB), an anti-Humean thesis according to which a rational agent desires a proposition to the degree that she believes (or expects) the proposition to be good. But I argue that independently of what measure of conditional desirability we prefer, we have good reasons for giving up Invariance. The unsoundness of Lewis' arguments against DAB does not, of course, make DAB true. In fact, I conclude the chapter with novel arguments against different versions of the DAB thesis, none of which assumes Invariance.[6]

When extending Jeffrey's theory to counterfactual prospects, I assume Richard Bradley's recent *Multidimensional Possible World Semantics*. Before extending Jeffrey's theory, I discuss several issues concerning the logic, semantics and probability of counterfactuals, in order to justify my use of the semantics. The discussion of these issues may seem to occupy too large parts of this thesis, given that its main topic is decision theory and practical rationality. However, as I hope will become clear, this discussion provides an important foundation for the chapters that follow. In chapter 2 I explore the formal relationships between several principles that we might want conditionals in general to satisfy. In chapter 3 I discuss the multidimensional semantics. I show that the semantics is more permissive than Bradley's original formulation suggested, and show that contrary to appearances, it is consistent with a particular Humean view on the metaphysics of modality. (Chapter 3 is largely based on

[6]The arguments against DAB are based on a joint working paper with Richard Bradley.

[Stefánsson, 2014c].)

Chapter 4 contains the main formal results of this thesis. Based on joint work in progress with Richard Bradley, I show how the multidimensional semantics allows us to extend Jeffrey's measure to counterfactuals. The resulting theory, which I will call *multidimensional* decision theory, allows us to represent both Allais' and Diamond's preferences as maximising the expected value of a decision theoretic value function, thus dissolving two 'paradoxes' that have burdened decision theorists for decades. In chapter 5, I use the result of chapter 4 to propose a general measure of the desirability of conditionals, and show how it can represent the various different ways in which conditionals influence the desirability of factual propositions.

The discussed dependency between actual and counterfactual outcomes very often occurs in moral decision-making, as we have already seen in the case of organ allocation. In chapter 6 (which is based on [Stefánsson, 2014b]) I use the the multidimensional decision theory to formulate what I call *Modal Consequentialism*. The main attraction of this theory is that it satisfies certain common moral intuitions, such as the aforementioned intuition about fair decision procedures in organ allocation and intuitions about fair distribution of risk of harms and chances for goods, that Non-Modal Consequentialism (e.g. classical utilitarianism and traditional welfare economics) violates.

0.4 Notation and terminology

Most symbols used in this thesis should be familiar. I use '$\wedge$' and '$\vee$' for conjunction and disjunction respectively, and '$\supset$' for material implication. In addition, '$\mapsto$' represents the indicative conditional connective, '$\boxright$' the counterfactual or subjunctive conditional connective (which I take to be two names for the same connective), and '$\rightarrow$' a variable that can either take $\mapsto$ or $\boxright$ as values. Italic capital letters A, B, etc., usually represent sentence variables, but non-italic capital letters, A, B, etc., denote factual propositions. I use Greek letters, α, β, etc., for propositions that could either be factual or modal. $v(w, A) = 1$ and $v(w, A) = 0$ respectively means that sentence A is true at w and false at w.

I will assume that a rational person's degrees of beliefs, aka her *credences*, can be represented by a probability function. P denotes a (subjective) probability function, P^{+}_{A} such a function on the evidential supposition that A, $P^{\Box}_{A}$ a probability function on the subjunctive supposition that A, and P_A is a variable that denotes either P^{+}_{A} or $P^{\Box}_{A}$.[7] *Ch* however denotes an *objective* probability (or chance) function. That is, unlike P, *Ch* does not represent the uncertainty of some particular agent, but some sort of objective (physical) uncertainty in the world. I will also assume that a rational persons' desires can be represented by a desirability function. *Des* denotes such a function, and $Des_{\mathrm{B}}(\mathrm{A})$ the desirability of A under the supposition that B.

[7] To avoid confusion, it might be worth pointing out right away that I interpret P^{+}_{A} as representing what the agent in question thinks, before learning A, her credence should be if she comes to believe A. Alternatively, one could interpret P^{+}_{A} as representing what the agent thinks her credence should be if A were true. However, there are various propositions such that a reasonable agent recognises that she would probably not believe them even if they were true. Hence, this alternative interpretation does not rule out the possibility that $P^{+}_{A}(A) \neq 1$ which means that $P(A \mid A) \neq P(A \wedge A)/P(A)$. Since I want to use the standard ration formula for conditional probability, I cannot endorse this alternative interpretation. But the more general and more severe problem with this interpretation, is that there is no guarantee that the truth of a proposition by itself affects an agent's beliefs.

Wlodek Rabinowicz has suggested that instead of either of the above interpretations, we think of P^{+}_{A} as representing what an agent would rationally believe "on the hypothetical assumption that A is true". And he suggests that this interpretation ensures that $P^{+}_{A}(A) = 1$ for any consistent A. I am happy with putting it that way, but as far as I can tell, when I hypothetically assume that A is true, I must, if the aforementioned equality is to hold, be imagining myself to be in a situation where I fully believe A.

Chapter 1

Conditional Desirability

1.1 Introduction

Before discussing how the truth of a counterfactual proposition can rationally influence the desirability of a factual proposition, or an actual outcome, it is useful to get clear about how the truth of a factual proposition can influence desirability. To that end, I defend in this chapter a particular measure for conditional desirability, which was first suggested by Richard Bradley. Another reason for starting by discussing measures for conditional desirability is that I will, in later chapters, assume the measure I defend here.[1]

The measure for conditional desirability I favour implies the falsity of the so-called *Invariance* principle, which states that the desirability of a proposition is independent of whether the proposition is true. The most common alternative to the conditional desirability measure I favour, however, implies the truth of Invariance. Hence, the principle can be used to determine the relative appropriateness of the two measures. Invariance has not received much critical attention in the philosophical and decision theoretic literature.[2] Therefore, this chapter is mostly devoted to discussing this principle. I will argue that we have reasons for rejecting Invariance that are independent of the plausibility of the conditional desirability measure I defend.

Invariance plays an important role in David Lewis' arguments against the so-called Desire-as-Belief thesis (DAB), an anti-Humean view according to which a rational person desires a proposition to the extent that she believes the proposition to be desirable. I conclude the chapter with arguments against two different versions DAB. Neither argument assumes Invariance. Although this last part of the chapter may seem quite disconnected from the other parts, it is important, given the Humean assumptions I make in this thesis, to establish that even if we give up Invariance, we can still provide plausible arguments against DAB.

1.2 Measuring Conditional Desirability

How should we evaluate conditional desirability in general? That is, taking any arbitrary propositions A and B, how should we calculate the desirability of A given (or under the

[1] This chapter is based on [Stefánsson, 2014a].

[2] This is only true of the subjective version of Invariance. After I submitted this thesis, Wlodek Rabinowicz pointed out to me that a *normative* Invariance principle (originally formulated by himself), according to which the normative status of an act is independent of whether it is performed, has received some attention. One of the future extensions of this thesis will be to look more closely at the normative Invariance principle, and compare the arguments I make against subjective Invariance to the discussion of the normative version of the principle.

evidential[3] supposition that) B? Although Jeffrey did not say much about conditional desirability, he did suggest in an exercise to his *The Logic of Decision* that $Des_B(A) = Des(A \wedge B)$. Several philosophers have followed Jeffrey in this regard. James Joyce makes the same assumption [Joyce, 1999] and one of Lewis' arguments against DAB makes explicit this assumption as well (see [Lewis, 1996]: 310). Here is the only remark Jeffrey makes in defence of his suggestion (with a slight change in notation):

> Suppose your beliefs and desires change from those characterized by a pair P, Des to those characterized by a new pair, P_B, Des_B, simply because you have come to fully believe a proposition B that you had not fully believed or fully disbelieved before. Under such circumstances, it is plausible to suppose that your new evaluation of each proposition A will simply be your old evaluation of $A \wedge B$. ([Jeffrey, 1983]: 90)

The reasoning behind Jeffrey's suggestion that $Des_B(A) = Des(A \wedge B)$ is presumably the following. Since you now fully believe B to be true, then whenever you learn some other proposition A, you have learnt that both A and B hold true. Hence, you should now evaluate A as you evaluated $A \wedge B$ before you learnt B. I think this idea of Jeffrey's is mistaken. The mistake in effect consists in *double counting*: B is already believed true, so evaluating A as $A \wedge B$ in effect counts the value of B twice; first, when the B was discovered to be true, then again when one evaluates the desirability of A given B.[4] Below I provide a few arguments against Jeffrey's conditional desirability measure, and in favour of the measure I prefer. In section 1.4 I will further develop some of these these arguments, but there the focus will be on the Invariance assumption that these two measures disagree about.

Let us first compare Jeffrey's suggestion with his own understanding of desirability as *News value*: the desirability of any proposition should, on this view, be thought of in terms of how much one would welcome the news of its truth. Suppose, for instance, that B is the proposition that you are going to a vacation in Thailand next month – which you find very desirable, in and of itself – and A the proposition that the weather forecast for Thailand next month is just about average for Thailand that month. Then it would seem quite odd to find the news of A given B as valuable as the news of A and B. Since you find it very desirable to vacate in Thailand, it would be very desirable to learn that you are going on vacation in Thailand next month and that the weather will be average for the region and time. But the value of learning that the forecast for Thailand is as can be expected, given that you have already learnt that you are going on vacation there, is at most slightly positive. For when evaluating the latter (conditional) piece of news, you should not count again the news that you are going on vacation in Thailand, since it is being assumed that you already know that. Instead, it seems, the value of the latter piece of news should be determined by much

[3] I will assume that when it comes to the effects of suppositions on desirability, counterfactual suppositions do not need a special treatment.

[4] This claim depends on my suggested interpretation of conditioning on A as representing an agents commitment to a particular attitude change in a situation where she believes A to be true. In footnote 7 of last chapter I give one reason why I think we should interpret conditional probabilities in this way. It would seem strange if this is how we interpret conditional probability, but think of conditional desirability as reflecting how the agent thinks she should change her desires if A were *true* (rather than if she were to believe A). In any case, this is an assumption about conditional desirability that seems to be standard in the literature, and is for instance shared by both Jeffrey and Richard Bradley, despite their disagreement on how to formulate conditional desirability (as I explain below).

The interpretation of conditional probability according to which it represents what an agent would believe "under the hypothetical assumption that A is true" also supports this claim, if we imagine ourselves to believe A when we hypothetically assume A's truth, as we must do, I argued in fn. 7 of last chapter, if the ratio formula is to be appropriate for conditional probability.

value the news of the conjunction of A and B adds to the value of B only. So Jeffrey's own interpretation of desirability does not support his conditional desirability measure.

Let us next compare the above idea of Jeffrey's with the assumption, common in most social sciences (in particular economics), that if a rational agent finds two propositions A and B equally desirable, then she should be willing to give up the same or an equally valuable good to make A true as she would be willing to give up to make B true. (The argument works independently of whether we assume that the contents of desires are propositions or more concrete goods.) In general, what an agent would be willing to give up in order to make A true, given that B is already true, should certainly *not* be identical to what she would be willing to give up in order to make A ∧ B true. To take an example, what I am willing to give up to have a piece of chocolate, given that I already have coffee, is certainly not as much as what I would be willing to give up to have the chocolate *and* coffee, since by assumption, I already have the former (which I find quite desirable on its own). Instead, it reflects how much I desire having chocolate and coffee over and above simply having coffee.

The above suggests, that given how we intuitively evaluate conditional desirability, the following measure (suggested by [Bradley, 1999]) is more appropriate than Jeffrey's:

Thesis 1 (Conditional Desirability)**.** *For any propositions A, B, the conditional desirability of A given B is given by:*

$$Des_B(A) = Des(A \wedge B) - Des(B) \tag{1.1}$$

This way of calculating conditional desirability is not only more intuitive than Jeffrey's; it can also be shown that people are vulnerable to *money pumps* unless they evaluate conditional desirability according to Bradley's formula ([Bradley, 1999]). Assume that the value of butter, given that you have bread, is the same for you as the value of bread and butter (as Jeffrey suggests); and suppose you value it as much as you value $\$x$.[5] Also assume that you would rather have bread than nothing, and say you value bread only as much as you value $\$y < \$x$.[6] As it happens, you have neither bread nor butter, whereas I have both. I first sell you the former for $\$y$. Now given that you have bread, you are willing to buy butter for $\$x$. After having sold you butter as well, I offer you $\$x$ plus some amount $\$z < \$y$ for your bread and butter, which you happily accept since bread and butter is worth $\$x$ to you.[7] But that means that you are in the same situation as before – having neither bread nor butter – except that you have lost $\$(y - z) > 0$. I could of course repeat the process, thus using you as a 'money pump'. In general, you will be vulnerable to such pumping whenever either $Des_{\mathrm{B}}(\mathrm{A}) > Des(\mathrm{A} \wedge \mathrm{B}) - Des(\mathrm{B})$ or $Des_{\mathrm{B}}(\mathrm{A}) < Des(\mathrm{A} \wedge \mathrm{B}) - Des(\mathrm{B})$. Hence, if you are instrumentally rational, and would also rather have more money than less, you will make sure that $Des_{\mathrm{B}}(\mathrm{A}) = Des(\mathrm{A} \wedge \mathrm{B}) - Des(\mathrm{B})$.

Yet another problem with the assumption that $Des_{\mathrm{B}}(\mathrm{A}) = Des(\mathrm{A} \wedge \mathrm{B})$, is that the formula implies the following symmetry: $Des_{\mathrm{B}}(\mathrm{A}) = Des(\mathrm{A} \wedge \mathrm{B}) = Des(\mathrm{B} \wedge \mathrm{A}) = Des_{\mathrm{A}}(\mathrm{B})$. In other words, the desirability of B given A should in general equal the desirability of A given B. But that is not true. Suppose I really hate going to the beach. In general I enjoy the sun, however, and although I would rather do something else than go to the beach when it is sunny, I certainly would rather go to the beach when it is sunny than when it is not. Now let A be the proposition that it is sunny, B the proposition that I go to the beach. Then A is good news conditional on B, but B is bad news conditional on A. So given Jeffrey's

[5] In other words if Br represents bread and Bu butter, then for you: $Des_{\mathrm{Br}}(\mathrm{Bu}) = Des(\mathrm{Bu} \wedge \mathrm{Br}) = Des(\$x)$.

[6] So $Des(\mathrm{Br}) = Des(\$y)$.

[7] That is, since more money is (let us assume) better than less money, $Des(\$x + \$z) > Des(\$x) = Des(\mathrm{Br} \wedge \mathrm{Bu})$.

convention that undesirable propositions get assigned a negative desirability value and desirable propositions a positive desirability value (which I will come back to below), we have in this particular case: $Des_{\mathrm{B}}(\mathrm{A}) > 0$ but $Des_{\mathrm{A}}(\mathrm{B}) < 0$. Hence, $Des_{\mathrm{B}}(\mathrm{A}) \neq Des_{\mathrm{A}}(\mathrm{B})$.

Given the formula for calculating conditional desirability that I have been endorsing, it will not in general be true that $Des_{\mathrm{B}}(\mathrm{A})$ equals $Des_{\mathrm{A}}(\mathrm{B})$. (In fact, the equality will only hold when $Des(\mathrm{A}) = Des(\mathrm{B})$.[8]) And this formula delivers the right result for the example in last paragraph: since I find the proposition that it is sunny more desirable than the proposition that I am going to the beach, i.e. $Des(\mathrm{A}) > Des(\mathrm{B})$, it is more desirable that it is sunny given that I am going to the beach than that I am going to the beach given that it is sunny, i.e. $Des_{\mathrm{B}}(\mathrm{A}) = Des(\mathrm{A} \wedge \mathrm{B}) - Des(\mathrm{B}) > Des_{\mathrm{A}}(\mathrm{B}) = Des(\mathrm{A} \wedge \mathrm{B}) - Des(\mathrm{A})$.

Before discussing Invariance and the role it plays in Lewis' argument, I should mention an objection Jim Joyce has made to my claims in this section. He points out that one way of making sense of Jeffrey's formula for conditional desirability, is to think of conditional desirability as measuring total welfare, rather than incremental changes in welfare. I agree that on that interpretation of conditional desirability, Jeffrey's formula is very natural, and the arguments in this section lose their bite. However, I take it that the fact that Jeffrey's formula strikes one as having counterintuitive implications, as the counterexamples discussed in this section show, is evidence that we tend to think of the desirability of A given B in terms of the incremental changes in welfare that A brings when B is believed true, rather than in terms of the total welfare when A and B are true.

1.3 Lewis against Desire-as-Belief

The measure of conditional desirability that I have been defending violates the Invariance principle, according to which the desirability of a proposition is independent of whether the proposition is true: if $Des_{\mathrm{A}}(\mathrm{A}) = Des(\mathrm{A} \wedge \mathrm{A}) - Des(A) = 0$, then it will not be generally true that $Des_{\mathrm{A}}(\mathrm{A}) = Des(\mathrm{A})$. The measure for conditional desirability that Jeffrey suggests however implies Invariance, since $Des(\mathrm{A} \wedge \mathrm{A}) = Des(\mathrm{A})$. I believe that we have reasons for giving up Invariance that are independent of the plausibility of Bradley's conditional desirability measure. Moreover, I believe and will argue below that A should neither be desirable nor undesirable given A. Invariance plays an important role (as an undefended assumption) in David Lewis' well-known arguments against the so-called Desire-as-Belief thesis (DAB). Before defending my claim that Invariance is false, let's look at the role it plays in Lewis' argument against DAB.

Recall that DAB says that a rational agent desires a proposition to the extent that she believes (or expects) the proposition to be desirable. Lewis took the thesis to be one way

[8] Are there no A and B such that for some rational agents, $Des(\mathrm{A}) = Des(\mathrm{B})$ but nevertheless $Des_{\mathrm{A}}(\mathrm{B}) \neq Des_{\mathrm{B}}(\mathrm{A})$? Possibly, but it is unclear that such examples should count as counterexamples in the present context. Suppose I find attaining a state of nirvana equally desirable as receiving a billion pounds. Attaining a state of nirvana, given that I have a billion pounds, is presumably also desirable. However, once I have attained a state of nirvana, worldly possessions are of no value to me. Hence, when that I have attained a state of nirvana, receiving a billion pounds is of no value to me. So we have a counterexample to the assumption that whenever $Des(\mathrm{A}) = Des(\mathrm{B})$, $Des_{\mathrm{A}}(\mathrm{B}) = Des_{\mathrm{B}}(\mathrm{A})$. The problem with this purported counterexample is that it is a mischaracterisation of conditional desirability. This might be best explained by taking an example from conditional probability. I think that conditional on being drunk, I drive very badly. However, when I am drunk, I think I drive very well. The former evaluation is what is reflected by the conditional probability I now assign to the proposition that I drive well conditional on being drunk; the latter is, for these purposes, irrelevant. Similarly, how I would evaluate a billion pounds in a state of nirvana is not relevant for evaluating the formula for conditional desirability. Rather, what we need to consider is how I *now* evaluate a billion pounds given that I have attained a state of nirvana. But when we do that, it is unclear that a counterexample like this can be made to the formula.

As far as I can tell, there do not seem to be any A and B, such that for a rational agent, $Des(\mathrm{A}) \neq Des(\mathrm{B})$ but nevertheless $Des_{\mathrm{A}}(\mathrm{B}) = Des_{\mathrm{B}}(\mathrm{A})$.

of formulating the anti-Humean theory of motivation. Humeans hold that a belief is not enough to motivate a person to act: any motivation is at least partly based on a *desire* (see e.g. [Smith, 1994]: ch. 4). Moreover, beliefs and desires are independent, according to Humeans, in the sense that a desire that A is not (even in rational agents) necessarily accompanied by any particular belief; nor does believing that B, in and of itself,[9] produce any particular desire. If DAB is true, then this Humean view must be false, since then whenever a rational agent desires that A she believes that A is desirable (and vice versa). So even the most fundamental (or 'primitive') desires of a rational person are not be independent of her beliefs.

To formally state DAB, and Lewis' argument against the thesis, let's use V as a zero-one normalised desirability function; i.e. a function from propositions[10] into the interval [0,1]. Given Jeffrey's convention of reserving 0 in the desirability measure for propositions that are neither desirable nor undesirable, his desirability function, *Des*, is clearly not zero-one normalised in this way. But we could think of V as an order preserving function that takes us from the image of *Des* – presumably from a (not necessarily proper) subset of the set of real numbers – and into the zero-one interval. Recall that P^+_{A} is an agent's revised credence function after evidentially supposing A. As I will further discuss in next chapter, it is generally accepted that the impact of learning A on a rational agent's credences and desires should in general be the same as the (hypothetical) impact of evidentially supposing that A.[11] Finally, for any proposition A, we define Å as the proposition that A is desirable.

Lewis gave a few arguments against DAB ([Lewis, 1988], [Lewis, 1996]). I will state the simplest of these arguments, to be found in section 4 of Lewis' second paper on DAB. First, we have a formal statement of DAB:[12]

Thesis 2 (Desire-as-Belief (DAB))**.** *For any A and according to any rational agent:*

$$V(A) = P(\mathring{A}) \tag{1.2}$$

Lewis claims that DAB implies what he calls 'Desire-as-Conditional-Belief':

Implication 1 (Desire-as-Conditional-Belief (DACB))**.** *For any A and according to any rational agent:*

$$V(A) = P^+_A(\mathring{A}) \tag{1.3}$$

Together DAB and DACB imply:

$$P(\mathring{\mathrm{A}}) = V(\mathrm{A}) = P^+_{\mathrm{A}}(\mathring{\mathrm{A}}) = V_{\mathrm{A}}(\mathrm{A}) \tag{1.4}$$

(The last equality holds since DAB is assumed to continue to hold after a rational agent learns that A.) In other words, A and Å are probabilistically independent:

[9] A means-end belief, such as a belief that ϕ-ing will produce result X, may on the Humean picture lead an agent desire to ϕ, but only if the agent already desires X.

[10] To state DAB, and his argument against it, Lewis needs to formalise the thesis in a framework, such as Jeffrey's, where the content of desires and beliefs are taken to be the same.

[11] Invariance thus demands that neither evidentially supposing nor learning A should affect A's desirability.

[12] I should point out that Lewis also considers and rejects a more plausible version of DAB, which allows for different degrees of goodness and states that a rational agent desires a proposition to the extent that she *expects* the proposition to be (perhaps objectively) desirable (or good) (see equation 17 in [Lewis, 1988]). For now the difference between these theses does not matter. In section 1.5.2 I provide an argument against the more general (and arguably more plausible) version of DAB.

Implication 2 (Independence (IND))**.** *For any A and according to any rational agent:*

$$P(\mathring{A}) = P^{+}_{A}(\mathring{A}) \tag{1.5}$$

Below I will show why IND is problematic. But why does Lewis claim that DAB implies DACB? Strictly speaking it does not. However, given the Invariance assumption, which Lewis takes as given, DAB *does* imply DACB. Here is a formal statement of Invariance:

Assumption 1 (Invariance (INV))**.** *For any A and according to any rational agent:*

$$V_A(A) = V(A) \tag{1.6}$$

If we already accept INV, then there is however no need to introduce DACB. For DAB and INV together imply IND.

Why is IND problematic? It is not hard to show that even if we start with a probability function for which such independence holds, it is not guaranteed that it will continue to hold after the agent in question revises her beliefs in accordance with Bayesian conditionalisation (an example is provided in next paragraph). That is, suppose that an agent's revised partial beliefs after she has learnt A, represented by the probability function P', is related to her partial beliefs *before* learning A, represented by P, by the following condition (whenever $P(A) < 0$): for any proposition B, $P'(\mathrm{B}) = P(\mathrm{B} \mid \mathrm{A}) = P(\mathrm{A} \wedge \mathrm{B})/P(\mathrm{A})$. Then we cannot be sure that this agent will satisfy IND both before and after such revision. Why is that a problem? Because it is generally assumed that Bayesian conditionalisation is how we are rationally required to revise our (partial) beliefs after learning some new proposition. (Call this norm BAYES.) Anything implied by a requirement of rationality is rationally required. Hence, if INV and DAB are both requirements of rationality then IND is rationally required. But then given certain propositions that an agent could very well learn, and thus revise her beliefs by conditionalising on, it will be impossible for her to satisfy the requirements of rationality, since she must either violate IND or BAYES.

Here is an example where IND and BAYES cannot both be satisfied. Assume that there is some proposition A such that $0 < P(\mathrm{A}), P(\mathring{\mathrm{A}}) < 1$. (If we cannot assign both A and Å credence strictly between 0 and 1, without undermining DAB, the thesis only holds in quite trivial cases, as Lewis points out.) This of course implies that $0 < P(\mathrm{A} \vee \mathring{\mathrm{A}})$. Hence, it should be possible for an agent to learn that $\mathrm{A} \vee \mathring{\mathrm{A}}$ and we should have no problems with conditionalising on this proposition, using Bayesian conditioning. But conditionalising on this disjunction turns a probability function for which IND holds into one for which IND does not hold: it leaves the conditional probability of Å given A unchanged, but increases the probability of A (since $P(\mathrm{A}), P(\mathring{\mathrm{A}}) < 1$)). Hence, when $P(\mathring{\mathrm{A}}) = P(\mathring{\mathrm{A}} \mid \mathrm{A})$ and $0 < P(\mathrm{A}), P(\mathring{\mathrm{A}}) < 1$, $P_{\mathrm{A} \vee \mathring{\mathrm{A}}}(\mathring{\mathrm{A}}) \neq P_{\mathrm{A} \vee \mathring{\mathrm{A}}}(\mathring{\mathrm{A}} \mid \mathrm{A})$. Less formally, if IND holds before an agent has learnt $\mathrm{A} \vee \mathring{\mathrm{A}}$, then she cannot still satisfy IND after having learnt this unless she violates BAYES.

Lewis' argument thus shows that an agent cannot satisfy all of BAYES, INV and DAB. But rationality presumably does not make inconsistent demands. Hence, not all of of BAYES, INV and DAB can be rationally required. Let us accept, with Lewis, that BAYES is a requirement of rationality. But then either INV or DAB must go. Lewis takes INV as given and thus rejects DAB. Below I argue that we should reject *both* INV and DAB.[13]

[13] In his first paper against DAB, Lewis defined Invariance slightly differently ([Lewis, 1988]: 327). Acquiring new evidence can change the expected value of a proposition A, Lewis claimed, just in case this new evidence

1.4 Against Invariance

Those commenting on Lewis' argument against DAB have, so far, not questioned Invariance; not even those who have tried to save some version of DAB from Lewis' criticism (see e.g. [Price, 1989], [Hájek and Pettit, 2004], [Bradley and List, 2009]). In this section I present a number of arguments against the Invariance assumption. I will start with the simplest argument, which consists in showing that INV is incompatible with Richard Jeffrey's interpretation of desirability as news value. Although Lewis based his argument against DAB on the decision theoretic framework developed by Jeffrey, he may not have endorsed Jeffrey's *interpretation* of desirability. But as I show, Jeffrey's framework is incompatible with INV on *any* interpretation of desirability. I will furthermore show that INV is incompatible with the *Efficacy value* interpretation of choice-worthiness favoured by causal decision theorists, and incompatible with the general idea that desirability is revealed through rational choices. I end this section by responding to the argument that Invariance must hold for so-called *maximally specific propositions*.

1.4.1 News value

In arguing against the DAB thesis, Lewis claimed to be basing his argument on Jeffrey's "exposition of Decision Theory" ([Lewis, 1988]: fn. 3). Jeffrey himself found it most natural to interpret his desirability function as measuring *news value*. That is, desirability, on this interpretation, is a measure of how good *news* it is that a proposition is true. To evaluate the desirability of the proposition that you are going to the beach, to take a simple example, you should ask yourself how valuable it would be to learn that you are going to the beach, which should partly depend on your credences in the different ways in which the proposition can come true; e.g. whether you will be going to the beach and it will be raining or sunny.

It seems clear that Invariance does not hold given Jeffrey's interpretation of desirability as news value. The news value of a proposition is highly dependent on whether we take it to be true; a proposition doesn't even count as news after we have learnt its truth.

To see more clearly that Invariance fails on the news value interpretation, let us look at how James Joyce justifies Jeffrey's convention of assigning a desirability value of 0 to any tautology. This convention is of course most natural if the desirability function is not bound by zero and one (as I assumed V is) since there are many propositions less desirable than a tautology. Here is Joyce's justification for the aforementioned convention (emphases added and notation slightly changed):

> There is a clear sense [...] in which $\top$ [i.e. any arbitrary tautology] cannot really be news; *learning that a tautology is true is not really learning anything*, or at least nothing informative. So, it makes sense to set $Des(\top) = 0$.
>
> This convention has a number of advantages. First, it amounts to assigning a news value of 0 to any [proposition] whose subjective probability is 1. *This reflects the fact that a proposition no longer counts as 'news' for someone who is already*

changes our evaluation of whether A is more likely to come true in a good way rather than bad. But this cannot happen if there is just one way in which A can come true (in which case we can think of A as being 'maximally specific'). In such cases, Lewis argued (ibid.), we should find that for any proposition B, $V_B(A) = V(A)$. This view of Lewis' seems very strange, given that he assumes that $V_B(A) = V(A \wedge B)$. But the view implies Invariance, as defined above but limited to propositions that can become true in one way only, for the aforementioned condition is supposed to hold also when B = A. As should become clear, my arguments against Invariance, as defined above, is also an argument against Invariance limited to propositions that can come true in one way only. I come back to this issue in section 1.4.5.

> *certain that it is true*. Second, it follows from [Jeffrey's desirability measure] that $Des(X) > Des(\neg X)$ implies $Des(X) > Des(\top) > Des(\neg X)$ as long as $1 > P(X) > 0$. This says that an agent who regards X as 'good news' will rank X above $\top$ and $\top$ above $\neg X$; no news is always worse than 'good news' and better than 'bad news.' Setting $Des(\top) = 0$ gives 'good news' a positive value and 'bad news' a negative value, *thereby making it reasonable to think of* $\top$ *as a description of the current status quo*... ([Joyce, 1999]: 122-123)

Relative to the desirability measure Des_A, all of Joyce's justifications for assigning any tautology zero desirability also justifies assigning the proposition A zero desirability. When A is already believed (or supposed)[14] true, "learning that [A] is true is not really learning anything". Moreover, if A is already believed (or supposed) true, then A represents the status quo just as well as any tautology. Finally, given that A is already believed (or supposed) true, A should should get a higher value than bad news and a lower value than good news. For 'no news' is worse than 'good news' but better than 'bad news'. In other words, whenever A is good news relative to the desirability function *Des*, such that $Des(A) > \top$, we should have $Des_A(A) < Des(A)$; but whenever $Des(A) < \top$, we should have $Des_A(A) > Des(A)$. Hence, it is not generally true that the desirability of a proposition, thus understood, is independent of whether it is believed to be true. As we have seen, the measure for conditional desirability that I favour satisfies these inequalities, and also guarantees that $Des_A(A) = 0$.

Jim Joyce has made the following objection to the view (implied by his earlier remark) that $Des_A(A)$ should always be assigned a desirability value of 0. The view entails that we constantly rescale the desirability measure; i.e., constantly change it such that at any particular time, the desirability of the status quo at that time is 0. But this, Joyce points out, results in a loss of information. Suppose the status quo at time t is not very good for agent i, but in the time between t and a later time t^+, she wins the lottery. This presumably makes the status quo at t^+ better for agent i. However, the information that i's situation has improved since time t is lost when we rescale her desirability function at time t^+.

It would certainly be an undesirable consequence of my view if one could not compare the desirability of the status quo at different times. The right response to this problem, I would suggest, is that we need to find a way of modelling agents, withing Jeffrey's framework, as 'unlearning' what they have already come to fully believe, so as to compare a new status quo with an old one. That is, in the example discussed in last paragraph, we need to be able to ask agent i at time t^+ how undesirable it would be to discover that she did not, after all, win the lottery (assuming that nothing else changed between t and t^+). The problem is that once the agent in question is certain that she has won the lottery, and has thus conditioned on the corresponding proposition, there is no way back in the standard Bayesian framework. That is, once we have conditioned on proposition A – using standard Bayesian conditioning, as opposed to Jeffrey conditioning which is only appropriate when the agent is not certain that A – we cannot condition on its negation. There are standard ways to 'subtract' such a proposition A when working with non-probabilistic belief revision. But for the present purposes, we would need to be able to do the same for the type of probabilistic belief revision that could work in Jeffrey's decision theoretic framework. As far as I know, this has not been worked out yet.

[14] Although A may strictly speaking count as 'news' for someone who *hypothetically supposes* A, learning that A is true should not (given that the supposition is evidential) make any difference to the set of attitudes to which A has been hypothetically added. Hence, relative to that set, A does not count as news.

1.4.2 Jeffrey's framework

Although Lewis used Jeffrey's decision theory to argue against DAB, he might not, as already mentioned, have wanted to take on board Jeffrey's *interpretation* of what the theory measures. However, even without that interpretation, Jeffrey's theory is incompatible with Invariance.

According to this theory, the desirability of any proposition, A, is a weighted average of the different mutually exclusive and jointly exhaustive ways in which the proposition can be true, where the weight on each way A_i that proposition A can be true is given by $P(A_i \mid A)$. More formally:[15]

Jeffrey's Equation. $Des(A) = \sum_{A_i \in A} P(A_i \mid A).Des(A_i)$

where both *Des* and *P* are defined over an atomless Boolean algebra of propositions from which the impossible proposition has been removed.[16]

Jeffrey assumes that there is just one tautological proposition, $\top$, which for any proposition B can be expressed as $B \vee \neg B$ ([Jeffrey, 1983]: 76), and, recall, gets assigned a desirability value of 0. From Jeffrey's equation, it then follows that:

$$Des(\top) = Des(B \vee \neg B) = Des(B).P(B) + Des(\neg B).P(\neg B) = 0 \tag{1.8}$$

Before considering what happens in the limit case when the agent is certain that B, let's consider what happens as the agent considers B more and more probable. As $P(B)$ approaches 1, $P(\neg B)$ approaches 0, and therefore $Des(\neg B).P(\neg B)$ approaches 0; hence, since $Des(B).P(B) + Des(\neg B).P(\neg B) = 0$, $Des(B)$ must approach 0. So the desirability of a proposition is generally not invariant under changes in its probability.

Now suppose that the agent in question learns (and comes to fully believe) B. Then:

$$Des_B(\top) = Des_B(B).P_B(B) + Des_B(\neg B).P_B(\neg B) = Des_B(B) \tag{1.9}$$

When working with Jeffrey's framework, the desirability of the tautological proposition is assumed to stay at zero, irrespectively of what the agent may learn.[17] (For the following

[15] Alternatively, since Jeffrey takes a proposition to be a set of possible worlds (as I will also do in this thesis), we can state Jeffrey's equation as:

$$Des(A) = \sum_{w_i \in A} P(w_i \mid A).Des(w_i) \tag{1.7}$$

I will use this version of the equation in chapter 4. As should be evident, they are formally equivalent, given the aforementioned understanding of what a proposition is.

[16] Having the desirability and probability measure on the same set has to some philosophers seemed to be a disadvantage of Jeffrey's theory, since it suggests that rational agents can assign meaningful subjective probabilities to their own *actions* (see for instance [Spohn, 1977] and [Levi, 2000]). It is certainly true that an agent should not use such probabilities to *decide* how to act. Moreover, the technique that is standardly used to *determine* subjective probabilities, namely to offer agents bets on the truth of propositions, is in general not appropriate when determining the subjective probabilities an agent assigns to her own actions (or, more generally, to any propositions whose truth and falsity is under her direct control). But unless we *define* subjective probabilities either by the role they play in the production of actions, or in terms of the bets an agent is willing to take, neither of these arguments show that agents cannot have subjective probabilities over their own actions. I think we should not define subjective probabilities this way, since that would mean that, by definition, someone who is either unable or unwilling to act or take bets can, by definition, not assign meaningful probabilities to propositions. (See [Rabinowicz, 2002] and [Joyce, 2002] for more detailed discussions of this issue.)

[17] This convention might for the following reason seem strange. We could see the desirability of a tautology as a weighted average of the propositions that we find epistemically possible, where the weights are determined by the probability of each of these propositions. But then the desirability of a tautology should change, when the probability of other propositions change. Nevertheless, Jeffrey himself suggests the convention of assigning the zero-point in the desirability measure to any tautology (see e.g. [Jeffrey, 1983]: 82), independently of what the agent in question believes, which others have followed (e.g. [Joyce, 1999] and [Bradley, ms]). One way to motivate this convention, is that since rational agents are modelled as knowing all tautologies, a tautology can never be news to them, nor will they be willing to risk anything to make a tautology true.

argument against Invariance to work, we however only need to make the following weaker assumption: there is *some* B that is not ranked with the tautology, such that learning that B is true does not change the desirability of the tautology.) So $Des_{\mathrm{B}}(\top) = Des(\top) = 0$. Therefore, whenever an agent has learnt (and comes to fully believe) a proposition B, she desires the proposition to the same degree that she desired the tautological proposition before learning B. But that means that except for propositions that are ranked with the tautology, Invariance fails. Suppose that before agent *i* learnt the truth of proposition B she considered B more desirable than $\top$. But as she learns (and comes to fully believe) B, the desirability of B becomes equal to that of $\top$, and hence less than what it was before she learnt B. The opposite will hold if agent *i* considered B less desirable than $\top$ before learning B. It should be evident that this argument does not depend on any particular interpretation of desirability. Hence, Jeffrey's framework (with the aforementioned convention) is incompatible with Invariance.

1.4.3 Efficacy value

Causal decision theorists have criticised the news value account for suggesting suboptimal choices in situations where an act is probabilistically, but not causally, related to some outcome, such as in the infamous Newcomb paradox ([Nozick, 1969]). Although some such theorists accept that news value is an appropriate interpretation of desirability, they suggest that the value rational agents' preferences and choices maximise (which some call 'choice-worthiness') should not be thought of as desirability (as news value). Rather, when deciding what to do, rational agents maximise the *causal efficacy* that they expect feasible *acts* to have in bringing about desirable states of affairs (see e.g. [Gibbard and Harper, 1981], [Lewis, 1981] and [Joyce, 1999]). Quoting Joyce again: "The quantity $U(A)$, [which according to Joyce's causal decision theory rational preferences maximise], gauges the extent to which performing [some act] A can be expected to bring about desirable or undesirable outcomes" ([Joyce, 1999]: 161).

Suppose for now that V does not measure desirability (as news value), but rather causal efficacy value. It should then be clear that whatever formal framework is used, V does not satisfy Invariance. Contrary to what has been done so far, let A in the statement of INV now refer to an act rather than a proposition. That is, INV, i.e. $V(\mathrm{A}) = V_{\mathrm{A}}(\mathrm{A})$, says that the causal efficacy value of an act is independent of whether it has already been performed or not. But that will certainly not always be true.[18] Suppose we could perform the same act twice, as has to be the case to evaluate $V_{\mathrm{A}}(\mathrm{A})$ on this interpretation. Then Invariance says that, for instance, the causal efficacy that having dinner has in producing in me a pleasurable physiological state is independent of whether I have already had dinner.

Let us then stick closer to the framework used by Lewis in his argument against DAB, i.e. let V be defined over a set of propositions but now think of $V(\mathrm{A})$ as measuring the extent to which the truth of A will cause or bring about desirable states of affairs. In this case we get a result similar to the one obtained for the news value interpretation. The world can never be made different by making a tautology come true. Similarly, given that A is true, the world cannot be made any different by making A true. So on a causal efficacy interpretation of V, we should have $V(\top) = V_{\mathrm{A}}(\mathrm{A})$, and if $V(\mathrm{A}) > V(\top)$ then $V(\mathrm{A}) > V_{\mathrm{A}}(\mathrm{A})$, but if $V(\mathrm{A}) < V(\top)$ then $V(\mathrm{A}) < V_{\mathrm{A}}(\mathrm{A})$.

[18] In keeping with the general interpretation of conditional value functions that I endorse in this thesis (see footnote 3 of this chapter), I am interpreting V_{A} as representing the causal efficacy that the agent in question expects various acts to have after she comes to fully believe that A has already been performed. I thank Wlodek Rabinowicz for encouraging me to make that clear.

To examine the compatibility of Invariance with causal decision theory, we might however want to formulate the assumption in terms of a contrary-to-factual supposition rather than an evidential, or matter-of-factual, supposition. Many causal decision theorists suggest that when evaluating an alternative (typically an act), we should ask ourselves what would happen if the alternative *were* realised (or if we *were to* perform the act), rather than what happens if it *is* realised (or if we *do* perform the act) (see for instance [Joyce, 1999]). It is well known that these different types of suppositions often lead to different re-distributions in credence (as I will further discuss in next chapter). But when it comes to evaluating Invariance, it makes no difference whether we take the supposition to be matter-of-factual or contrary-to-factual. When I suppose that contrary to fact, A is true, I imagine myself to be in a counterfactual situation where I believe A to be true.[19] But relative to that situation, A is neither more nor less desirable or valuable (under any of the interpretations considered) than the tautology. Hence, the contrary-to-factual version of Invariance fails.

1.4.4 Willingness to give up

It is commonplace, in particular in the social sciences, to identify how strongly an agent desires a proposition with how much she would be willing to give up to make that proposition come true. (In what follows, I will interpret 'giving up' very broadly. When working we 'give up' leisure, when we take a risk we 'give up' safety, etc.) The general idea is that some sort of choice behaviour should be a guide to how much an agent desires a particular proposition. And many social scientists, in particular economists, think that the type of choice behaviour that best reveals how much a person desires a proposition A (or a particular good g, as they are more likely to put it) is how much she willingly gives up to make A come true (or to attain g).

The proponents of both the news value and the efficacy value interpretation generally want their interpretation of value to give the same answer as the willingness to give up criteria. In all normal cases,[20] A has more news value than B for a person i if and only if i is willing to give up more to make A true than than to make B true. The same is true for for efficacy value. Moreover, this idea should be acceptable to both subjectivists and objectivists about value. Subjectivists claim that rational agents try to make true those propositions they expect to maximise some person-relative value.[21] Objectivists, however, claim that rational agents try to make true those propositions they expect to maximise objective value. As long as both subjectivists and objectivists are willing to call that which rational agents maximise 'desirability', they should agree that desirability has a causal role in the production of action. The greater the desirability of a proposition, the greater causal force it has on action. And this causal force can be measured by what an agent is willing to go through, risk, or give up in order to make true the proposition in question.

Given this interpretation of desirability – and in fact, *any* interpretation that satisfies the idea that desirability is revealed through rational choice – it seems clear that Invariance

[19] If you think that contrary-to-factually supposing A involves imagining yourself to be in a situation where A *is* true, rather than a situation where you *believe* A to be true, then Invariance still fails, provided that you think that the probability of you knowing A increases when A becomes true.

[20] There might be special cases where you find A more desirable than B but you think that unlike B, the value of A would be diminished if you make an effort to make it true. Hence, you might be willing to give up more to make B true than A, even though you find the latter more desirable. I will set such cases aside.

More importantly, when good or bad outcomes depend probabilistically, but not causally, on the alternatives an agent is faced with, value as measured by willingness to give up might differ from news value, as is well known.

[21] Some take this one step further, and take value, according to an agent, to be *determined* by her choice. I will not discuss such strict behaviourism in this thesis.

must be false. How much an agent is willing to give up in order to make a proposition A come true is certainly not independent of whether she takes A to be true already. Suppose an agent considers A, whose truth she is uncertain about, to be desirable. In other words, she would be willing to give up at least something of value in order to make A true; let's call this something G. What about after having learnt (or under the evidential supposition) that A is true already? Surely, whatever she would be willing to give up to make A true after having learnt that A is true already (if anything at all), should be less valuable to her than G.

It might be hard to even understand the question of what a rational agent would be willing to give up to make true a proposition that she believes to be true already. I would argue that the difficulty in making sense of the question stems from the fact that the answer is so obviously 'nothing'. But we can instead consider the following question: how much would a rational agent be willing to give up to make a desirable proposition A true when she is quite uncertain as to the truth of A, compared with how much she would be willing to give up to make A true when she is *almost* certain that A is true already? Given Jeffrey's framework, the desirability of A is less in the second case (as we have already seen). And that seems intuitively correct to me. Suppose you desire strongly that person i becomes the next president and are considering how much money to donate to i's campaign. You believe that i might win, and that paying some sum of money would make a positive difference (and the more money, the greater the difference). You conclude that given how much you want i to win, it would be rational for you to donate \$$x$ to i's campaign. But before you donate the money, you see polls that you take to indicate that i will almost certainly win. That should make you willing to donate less than \$$x$. This seems to hold in general. So the desirability of a proposition, understood as willingness to give up to make the proposition true, is not invariant under changes in its probability.

The argument in this section can be made more formal and precise by connecting it to standard decision theoretic representation theorems. Moreover, the more precise argument answers the following potential worry.[22] A critic might say that these arguments simply show that what is already believed true, or a good that is already had, should not *distinguish between alternatives*. That is, whenever we evaluate some alternatives, a proposition that is already believed true, or a good that is already had, should be taken into account when evaluating each of these alternatives. But that, in itself, does not show that the *value* of a proposition diminishes when it becomes true, nor that the value of a good diminishes when it is already had.

In making the argument in this section more precise, I will draw on the representation theorem of John von Neumann and Oskar Morgenstern (vNM's) (from [von Neumann and Morgenstern, 1944]), except that in keeping with Lewis' discussion, desirability (which vNM call 'utility') will be a measure on propositions. Let '$\prec$' stand for (strict) preference and '$\sim$' for indifference (both defined on a set of propositions). And let $[p\mathrm{A}, (1-p)\mathrm{C}]$ represent a lottery (or uncertain prospect) that results in either A (with probability p) or C (with probability $1-p$). The axioms of vNM's representation theorem imply that if $\mathrm{C} \prec \mathrm{B} \prec \mathrm{A}$, then there exists a unique $p' \in [0,1]$ such that $\mathrm{B} \sim [p'\mathrm{A}, (1-p')\mathrm{C}]$. A crucial step in the proof of the vNM theorem is to set the the desirability of B – which I will now refer to as $U(\mathrm{B})$ and which may neither be zero-one normalised nor normalised around 0 in the way Jeffrey suggests – equal to the desirability of the lottery L with A and C as prizes such that the agent in question is indifferent between L and B. In other words,

[22] I thank Robbie Williams for raising this issue.

$U(\mathrm{B}) \doteq U([p'\mathrm{A}, (1-p')\mathrm{C}])$ for the p' such that $\mathrm{B} \sim [p'\mathrm{A}, (1-p')\mathrm{C}]$. And since $\prec$ satisfies the vNM axioms, $U([p'\mathrm{A}, (1-p')\mathrm{C}]) = p'U(\mathrm{A}) + (1-p')U(\mathrm{C})$.

Suppose for some agent i, $\mathrm{C} \prec \mathrm{B} \prec \mathrm{A}$ and that we have found the value $p' \in [0,1]$ such that for this agent, $\mathrm{B} \sim [p'\mathrm{A}, (1-p')\mathrm{C}]$. Consider now what happens if we ask the agent in question to suppose that A is already true, and again try to find the value of p such that i is indifferent between B and $[p\mathrm{A}, (1-p)\mathrm{C}]$; that is, indifferent between a B and a lottery that has probability p of leading to what is the status quo when the choice is made. (Here the assumption, shared with Lewis, that the objects of desires and beliefs are propositions becomes important.) What should we expect to happen now? Since the agent prefers B to C, we should not find *any* value such that the agent is indifferent between B and the lottery. For given that A is true already, the possible outcomes of the lottery is either C, which the agent finds worse than B, or A, which is supposed true already. Thus the lottery either leads to the status quo or something that is worse than B.[23] Hence, if we let $\prec_{\mathrm{A}}$ represent preference *under the supposition that A*, we now have that for this particular agent, $[p\mathrm{A}, (1-p)\mathrm{C}] \prec_{\mathrm{A}} \mathrm{B}$ for *any* $p \in [0,1]$. But for rational agents – e.g. those satisfying the vNM axioms – this can only hold if $\mathrm{A} \prec_{\mathrm{A}} \mathrm{B}$ and $\mathrm{C} \prec_{\mathrm{A}} \mathrm{B}$.

Let us assume that just as there is a U function representing and agent's unconditional preferences, so there is a U_{A} function representing the agent's preferences under the supposition that A. (What is needed for this to hold is explored in [Bradley, 1999].) Then although $U(\mathrm{B}) < U(\mathrm{A})$, $U_{\mathrm{A}}(\mathrm{A}) < U_{\mathrm{A}}(\mathrm{B})$. Moreover, for *any* β, γ such that $\gamma \prec \beta \prec \mathrm{A}$ and thus $U(\gamma) < U(\beta) < U(\mathrm{A})$, we can similarly show that $U_{\mathrm{A}}(\mathrm{A}) < U_{\mathrm{A}}(\gamma)$ and $U_{\mathrm{A}}(\mathrm{A}) < U_{\mathrm{A}}(\beta)$. Given the framework of von Neumann and Morgenstern, it is not generally possible to *directly* compare conditional utility with non-conditional utility, which means that the Invariance assumption is, technically speaking, meaningless. Nevertheless, the above argument shows that, given the above assumptions, A is considerably lower in the $\prec_{\mathrm{A}}$-ordering than the $\prec$-ordering. And since U represents $\prec$ and U_{A} represents $\prec_{\mathrm{A}}$, this gives an initial reason for being sceptical about the assumption that $U(\mathrm{A}) = U_{\mathrm{A}}(\mathrm{A})$.

In certain special cases it might be justifiable to assume that we can compare the conditional utility of a prospect with its non-conditional utility; for instance, if we assume that we know, through introspection, that the desirability of B, for us, is independent of A. Let us for now assume this independence between A and B and also between A and C; i.e. $U(\mathrm{B}) = U_{\mathrm{A}}(\mathrm{B})$ and $U(\mathrm{C}) = U_{\mathrm{A}}(\mathrm{C})$. It is standard when constructing a vNM representation, that two prospects are chosen, satisfying $\mathrm{C} \prec \mathrm{A}$, then it is stipulated that $U(\mathrm{C}) = 0$, $U(\mathrm{A})$ given some value greater than 0, and the utilities of other prospects constructed based on these. Let us continue to assume that for the agent we are modelling, $\mathrm{C} \prec \mathrm{B} \prec \mathrm{A}$, and that we have found the $p' \in [0,1]$ such that $\mathrm{B} \sim [p'\mathrm{A}, (1-p')\mathrm{C}]$, and let us stipulate that $U(\mathrm{C}) = 0$. In many circumstances, these are perfectly reasonable assumptions. There is some value p' between 0 and 1 such that I myself would be indifferent between \$10 (proposition B) and a lottery that gives me (with probability p') a trip to New York (proposition A) if I win but (with probability $1-p'$) a copy of *Time* magazine (proposition C) if I lose. But for me the value of \$10 is independent of whether I am going to New York or not; and the same holds for a copy of *Time* and the New York trip. After having learnt that I am going to New York, I would, however, no longer be willing to pay \$10 for the lottery $[p'\mathrm{A}, (1-p')\mathrm{C}]$. Given these assumptions, we have: $U(\mathrm{B}) = p'U(\mathrm{A})$, i.e. $U(\mathrm{A}) = U(\mathrm{B})/p'$; but $U(\mathrm{B}) = U_{\mathrm{A}}(\mathrm{B}) > p'U_{\mathrm{A}}(\mathrm{A})$, i.e. $U_{\mathrm{A}}(\mathrm{A}) < U(\mathrm{B})/p'$. Hence, $U_{\mathrm{A}}(\mathrm{A}) < U(\mathrm{A})$. (And the same of course holds for any zero-one

[23] In the latter case, A would *also* be true (unless A is inconsistent with C).

normalised, order preserving transformation of U.)

The above argument shows that we can have a perfectly reasonable preference for which Invariance fails. In this special case we can, given the above assumptions, compare $U(\mathrm{A})$ directly to $U_{\mathrm{A}}(\mathrm{A})$, and when we do so we find that these values are not equal. Hence, the worry that my argument only shows that a proposition believed true should not discriminate between alternatives is unwarranted.

To sum up, we have seen that Invariance does not hold given three common interpretations of desirability or choice-worthiness: it fails to hold on the willingness to give-up interpretation commonly used by social scientists, and on the news value interpretation used by the followers of Richard Jeffrey, and on the efficacy value interpretation of choice-worthiness favoured by causal decision theorists. Since decision theory is meant to provide a formal framework for predicting and explaining rational choice, it seems that no plausible interpretation of the values of decision theoretic value functions can avoid clashing with Invariance, since a rational agent's choice behaviour with respect to A is certainly often affected by the agent learning the truth of A. We might for some other purposes interpret desirability as a measure of some overall comparative goodness that is not stand-point relative in any way. But I hope to have shown that if the goodness of a proposition, in this sense, does not change, according to an agent, when the agent comes to believe that proposition to be true, then goodness in this sense cannot be the value that rational agents try to maximise. In other words, this cannot be the value that we refer to when we explain rational choices on the assumption that rational agents maximise expected utility, desirability, or good.

1.4.5 Maximally specific propositions

The proponent of Invariance might make the following objection to the above argument.[24] We want desirability to be defined over propositions that are *maximally specific* in all respects that are relevant to their value.[25] If A is such a proposition then for *any* proposition B, $V_{\mathrm{B}}(\mathrm{A}) = V(\mathrm{A})$, the defender of INV might argue: since A already includes everything that is important for its evaluation, conditionalising on B could not possibly make a difference to the evaluation of A. (Lewis himself makes an argument of this kind ([Lewis, 1988]: 332), albeit to justify the version of Invariance that he uses in his first paper on the Desire-as-Belief thesis.)

It might be true that for any maximally specific proposition A, and any proposition B such that $\mathrm{B} \neq \mathrm{A}$, $V_{\mathrm{B}}(\mathrm{A}) = V(\mathrm{A})$. But the same does not hold when $\mathrm{B} = \mathrm{A}$. Suppose A is a maximally specific proposition that I find more desirable than what I take to be my current situation. Presently, I am therefore willing to give up something of value to make A true, I judge that the truth of A would have desirable causal consequences and learning that A is actually true would be good. But what happens *after* I have learnt that this maximally specific proposition holds true? (Assuming for the moment that we can learn such specific propositions.) Then A neither counts as news for me nor will I take it to have any more casual efficacy than the tautology. Similarly, I would then not be willing to give up anything of value to make A true. Finally, it should be evident that the argument that Invariance is

[24] I thank Robbie Williams for raising this issue.

[25] Such propositions need not be *atomic*. (Indeed, to make this response compatible with Jeffrey's framework, we must assume that the maximally specific propositions in question are not atomic.) A proposition A that is maximally specific in all respects that is relevant to its value might be further divided into two propositions that differ in the outcome of a hypothetical and completely random coin toss C, provided that the outcome of C is irrelevant to A's value. Hence, A may be maximally specific, in the required sense, but not atomic.

incompatible with Jeffrey's framework also holds for maximally specific propositions.[26]

1.5 Is DAB then True?

If Invariance is false, as I have been arguing, then Lewis' arguments against DAB are not sound. But that does of course not mean that DAB is true. In this section I will assume that Invariance is false (and that desirability is stand-point rather than end-point relative), and show that Humeans still have good reasons for rejecting DAB. I first produce a direct counterexample to the simple DAB thesis, before giving a formal argument against a more general and plausible version of DAB.[27]

1.5.1 A Counterexample to Desire-as-Belief

Suppose we are sailing with our two good friends, Ann and Bob, when suddenly both of them fall overboard and find themselves in an equally difficult situation and threatened with drowning. In other words, if nothing is done, both will drown. In that situation, I think, we would *fully* believe the proposition that it would be good that Ann is saved. (Call this proposition Å.) Or if we can only fully believe tautologies (or perhaps only tautologies and factual statements about the past), then we at least believe Å as, or almost as, strongly as we believe any contingent proposition (or any contingent proposition that is not about the past). Nothing in the example hangs on treating Ann and Bob equally, but to simplify the discussion, let us suppose that our feelings for the two are identical in all relevant respects. Thus we also believe the proposition that it would be good to save Bob (call this proposition B̊) as (or almost as) strongly as we believe any contingent proposition. So for us, in that situation, $P(\text{Å}) = P(\text{B̊})$ is close to 1. To make the discussion that follows more precise, let us assume that $P(\text{Å}) = P(\text{B̊}) = 1 - \gamma$ (where γ is close or equal to 0).

In the situation we are imagining, we would also *desire* very strongly that Ann is saved (proposition A), and would desire equally strongly that Bob is saved (proposition B).[28] But we find it much more desirable – in fact about twice as desirable – that *both* Ann and Bob is saved than that one of them is saved. So $V(\text{A} \wedge \text{B})$ is roughly twice $V(\text{A})$. But then the Desire-as-Belief thesis dictates that the probability that it is good that both Ann and Bob are saved, $P(\text{A}\mathring{\wedge}\text{B})$, should be close to twice $P(\text{Å})$. But since $P(\text{Å})$ is close to 1, $P(\text{A}\mathring{\wedge}\text{B})$ can never be close to twice $P(\text{Å})$. Thus assuming that the requirements of rationality never prohibit what is rationally permissible, and if we take the attitudes towards Ann and Bob expressed in the above example to be rationally permissible, it seems that DAB cannot be a requirement of rationality.

To save DAB a proponent of it could argue that, contrary to appearances, the attitudes towards Ann and Bob assumed in the counterexample are in fact irrational. There are three ways she could do this. Firstly, she can deny that $P(\text{Å})$ is rationally permitted to be close to 1. Secondly, she can argue that $V(\text{A}) = V(\text{B})$ should be no greater than $(1 - \gamma)/2$. Thirdly, she can argue that $V(\text{A} \wedge \text{B})$ should not be much greater than $V(\text{A}) = V(\text{B})$. Alternatively, she could argue that some of the assumptions of the counterexample are meaningless.

[26] Rather than assuming that A includes everything that is important for its value, we might assume that the agent whose desires V represents *knows* everything that is important for determining A's value. In that case we might think of V as measuring 'informed value'. (I should thank an anonymous referee for raising this issue.) But although it might be true for an informed value function V that for any proposition B such that $\text{B} \neq \text{A}$, $V_{\text{B}}(\text{A}) = V(\text{A})$, the same does not hold when $\text{B} = \text{A}$, for all the same reasons that have been given above.

[27] What follows is based on a joint working paper with Richard Bradley.

[28] Assuming that the probability of them being saved is roughly equal. The significance of this assumption will become clear at the end.

Let's take each response in turn. Since we are assuming that saving both Ann and Bob is close to twice as desirable as saving one of them, the first response only works if we require that $P(\text{Å})$ is no greater than 0.5. So for this response to work, we must be no more certain about the proposition that it is good to save Ann (or Bob) than the proposition that a fair coin lands heads up when tossed. It is highly implausible that this is a requirement of rational belief.

In fact, we can make things much worse. Suppose now we are sailing with not just two but a number of our dear friends when suddenly all of them fall overboard. For the first response to the above counterexample to work, the credence we assign the proposition that it is good to save any one of our friends must get smaller and smaller as we increase the number of people that we imagine to have fallen overboard. But no matter how many friends we have, and how many of them we take out sailing, we would always be almost certain that it would be good to save each of them after having fallen overboard. To take an example, suppose we are sailing with six friends who all fall overboard. Then to save the DAB we cannot be more certain about the proposition that it would be good to save any particular friend than in the proposition that a dice shows side six when rolled! A conception of rationality that requires this seems very implausible.

The second route to saving the DAB thesis involves requiring that $V(\text{A}) \leq (1 - \gamma)/2$. But however we interpret desirability, it is hard to believe that rationality requires that saving Ann or Bob be confined to the bottom half of the desirability scale. In any case, this requirement coupled with the Desire-as-Belief thesis implies that $P(\text{Å}) \leq (1 - \gamma)/2$. In other words, this response requires us to to be no more certain about the proposition that it would be good to save Ann than in the proposition that a fair coin comes up heads if tossed. So this second response to our counterexample in the end comes down to the same as the first response and is no more plausible. And again, we can make this response even less plausible by increasing the number of people we are imagining to be in the water.

The third possible response would be to deny that it is permissible to judge it much more desirable to save both Ann and Bob than just one of them. Ordinary intuition (and many welfarist theories) suggests that saving both Ann and Bob would be roughly twice as desirable as saving one of them. Saving both might be *more* than twice as desirable as saving just one of them; for instance if we feel guilt for choosing to save one of them over the other other, or if choosing to save one over the other creates some sort of injustice or unfairness. Or it might be slightly *less* if Bob and Ann hate each other and would be happier if the other were dead. But if we set these complementarities aside then we are left with the core judgement upon which the example is based: that the desirability of saving Ann (or Bob) is independent of whether the other is saved or not. But if this is so, then it would seem to follow immediately from the assumption that saving Bob is equally desirable to saving Ann, that saving both is twice as desirable as saving one.

Could there be complementarities that we are rationally required to give weight to and which make assumed judgement irrational? It is hard to imagine what they could be. But even if there are such complementarities they are unlikely to make enough of a difference. The problem is that to rescue DAB from our counterexample, it would need to be demonstrated that the difference between the desirability of saving Ann and of saving both must of rational necessity be very small. Suppose, for instance, that we are 90% sure that it would be good to save Ann and 95% sure that it would be good to save both Ann and Bob, so that by DAB, $V(\text{A}) = 0.9$ and $V(\text{A} \wedge \text{B}) = 0.95$. Then it is just slightly above 5% more

desirable to save both Ann and Bob than to save one of them. That is implausible: having saved one of our friends, we would still make a great deal of effort and be willing to risk or pay quite a lot to save the other. It is hard to see why that would be irrational. We could, of course, argue about the plausibility of the exact numbers, but so long as the difference in the probability of Å and A∧B is not great, the difference in desirability between saving both friends and just one of them must be *very* small for this last response to work; much smaller than what most people would intuitively accept.

Let us consider then the final possible response to the counterexample, which works by questioning the meaningfulness of the assumption that the desirability of saving Ann and Bob is twice that of saving Ann. In the decision-theoretic framework in which DAB is stated, desirability functions are just numerical representations of preferences and only those properties of desirabilities that are analogues of properties of preferences should be considered meaningful. But the notion of 'twice as desirable as' fails this test, as is evidenced by the fact that a linear transformation of a desirability function (in particular one based on a different choice of zero point) will yield another desirability function that serves equally well to represent the underlying preference relation, but does not preserve properties such as one prospect being twice as desirable as another. So the counterexample trades on an unsustainable interpretation of desirabilities.

This objection is half-correct. It is true that a linear transformation of a desirability function does not preserve the property that we are interested in. But such transformations are ruled out by DAB which itself forces a particular choice of the zero and unit scaling points on the desirability function (namely the certainly bad and the certainly good propositions). One may well object that such a choice of scale is arbitrary, but this would be reason to object to DAB directly. Here I assume for the purposes of the argument that the scaling of desirabilities enforced by the DAB thesis *is* acceptable and examine its implications.

What then does 'twice as desirable as' mean within the scope of desirabilities as regulated by the DAB thesis? Roughly this: the agent who regards prospect X as twice as desirable as prospect Y is one who is indifferent between Y being true for certain and a lottery which makes X true with probability one-half and a certainly bad prospect true otherwise. Suppose for instance that it is certainly bad that both Ann and Bob are not saved. Then it is twice as desirable that both Ann and Bob are saved as that Ann is saved, just in case the prospect of Ann being saved is just as desirable as the prospect of either both being saved or neither, with an equal probability of each.[29]

To sum up: it seems that the attitudes towards our friends Ann and Bob expressed in the above counterexample are rationally permissible, and any attempt to save DAB in light of this example forces us to have attitudes that seem counterintuitive and which are certainly not rationally required. Hence, the example shows that this simple version of DAB must be false.

1.5.2 A Generalisation of Desire-as-Belief

The literature that followed in the wake of Lewis' first paper against Desire-as-Belief focused almost entirely on the version of DAB presented above, or minor variants of it. But it could be argued that the DAB thesis is rather implausible as an anti-Humean thesis and that it

[29] It might be objected that this definition assumes that the agent is risk neutral. But the objection is misplaced. Lewis formulated DAB within the decision theory developed by Jeffrey. And in that framework, risk attitudes are built into the desirabilities of propositions in the sense that the method for constructing a cardinal measure of desirability assumes risk neutrality with respect to desirability (though not with respect to specific goods).

is not really surprising that there are counterexamples to it. In [Lewis, 1996], Lewis in fact considered a more general (and arguably more plausible) version of the thesis which allows for different degrees of goodness, and claimed to show that it was equally susceptible to his negative results. It could be argued that the counterexample I discussed in last section shows the implausibility of assuming that desirability or goodness doesn't come in degrees, since the example does not undermine the more general version of DAB. Hence, I will in this section produce an argument against the generalised version of DAB (without assuming Invariance, as Lewis does).

To state the more general version, let $\mathring{\text{A}}_i$ be the proposition that A is good to degree g_i and assume for simplicity that the number of goodness degrees is finite. Then the generalised DAB thesis is that:

DAB Generalised. $V(\text{A}) = \sum_i g_i.P(\mathring{\text{A}}_i)$.

Again, we can provide an argument against the generalised principle that does not assume Invariance (unlike Lewis' argument). The important thing to note about the generalised thesis is that it implies that the $\mathring{\text{A}}_i$ are probabilistically independent of A (as I show below). It follows from Jeffrey's desirability measure and the fact that the $\mathring{\text{A}}_i$ partition the space of possibilities, that:

$$V(\text{A}) \quad = \quad \sum_i V(\text{A} \wedge \mathring{\text{A}}_i).P(\mathring{\text{A}}_i|\text{A})$$

According to what we might call the *Moral Principal Principle* (after the principle [Lewis, 1980] formulates), $V(\text{A} \wedge \mathring{\text{A}}_i) = g_i$. So it follows from DAB Generalised and Jeffrey's desirability measure that:

$$\sum_i g_i.P(\mathring{\text{A}}_i) = V(\text{A}) = \sum_i g_i.P(\mathring{\text{A}}_i|\text{A})$$

But this can only be the case non-accidentally, and hold for **any** proposition A, if $P(\mathring{\text{A}}_i) = P(\mathring{\text{A}}_i|\text{A})$. This independence implies once again that probabilities of goodness behave differently from desirabilities. For in contrast to the fact that A and the $\mathring{\text{A}}_i$ are probabilistically independent, A is not, as I have already argued, *desirabilistically* independent of its own truth (unless it is ranked with the tautology). In particular, the assumption that A is desirabilistically independent of its own truth (i.e. the Invariance assumption) is inconsistent with Jeffrey's desirability measure.

With this in mind, let us return to the example from last section to see how the generalised version of DAB fares. At first sight it seems to do much better than the original DAB thesis. Without having to specify the candidate degree of goodness it seems quite plausible that the probability that it would be good to some degree g_i that Ann is saved would be just the probability that it would be good to that same degree that Bob was saved. Furthermore, for very high degrees of goodness it might well be that it is much more probable that it is good to that degree that both are saved than that just one is. So the generalised DAB might seem to conform to our intuitions about the counterexample to the original DAB.

But if we let the imagined situation evolve, new problems emerge. Suppose that after a few minutes Ann manages to cling to the side of the boat such that it seems certain that she will be saved. Bob on the other hand remains in difficulty. Now since DAB Generalised implies that the $\mathring{\text{A}}_i$ are probabilistically independent of A, it remains the case, according to this thesis, that we should desire that Ann be saved to the same (expected) degree that we

desire that Bob be. But this is entirely wrong. Since it is nearly certain that Ann will be saved, it is much more desirable that Bob be saved than that Ann be. This follows, as we have seen, from Jeffrey's desirability measure. But we can also intuitively see this by noting that we would now devote much more resources to saving Bob than we would to saving Ann.

To sum up, I think the above considerations show that both the original and the more general DAB is false. However, I am not sure where that leaves the anti-Humean theory of motivation. Perhaps one could formalise such a theory that does not lead to difficulties like those discussed above.

1.6 Concluding Remarks

I hope to have shown that Bradley's conditional desirability measure is a plausible rule for rational desirability updating, and that its clash with Invariance is actually a good thing, since we have independent reasons for rejecting the latter. Humeans about motivation need not worry that giving up Invariance makes it impossible for them to argue against the anti-Humean DAB thesis, since there are good reasons for believing that DAB is false that do not depend on Invariance being true.

The remaining chapters of this thesis will be more focused on conditionals. In next chapter I discuss the relationship between various principles and norms we might want conditionals to satisfy. In chapter 3, I discuss the semantics of conditionals that I use in chapter 4 to introduce counterfactual prospects to Jeffrey's decision theory.

Chapter 2

Probability and Logic of Counterfactuals

2.1 Introduction

The main aim of this thesis, as already stated, is to explore the ways in which the desirability of actual states of affairs sometimes depends on the truth of counterfactual conditionals, and to construct a decision theory that allows for (and can represent) this dependency. In this chapter I discuss several epistemic and logical principles that we might want conditionals to satisfy. This will clear the ground for the semantics that I use and the assumptions I make in chapter 4 when introducing counterfactual propositions to Jeffrey's decision theory. The semantics I use, for instance, implies two logical principles known as *Centring* and *Conditional Excluded Middle*, and an epistemic principle called the *Ramsey test*. These principles are not, strictly speaking, necessary for the main aim of this thesis. Nevertheless, since the semantics I use implies these principles, I will devote this chapter mostly to discussing and defending them.

I will assume, in this chapter, that we intuitively accept certain logical principles and credence norms for conditionals, *independently of any formal semantics for conditionals*. That is, according to the methodology I will be following, we take as a starting point, before constructing semantical systems for conditionals, certain logical truths about conditionals – and similarly, norms about credence in conditionals; and make it a requirement on acceptable semantical systems that they either have these principles and norms as axioms, or generate them as theorems.[1] We might of course find out, by constructing semantic systems, that it is impossible to create theories that simultaneously satisfy all of the principles we want to accept. But I will assume that the starting point is to examine these principles pre-semantically.

It might be worth admitting at the outset that my discussion in this chapter is not particularly novel. I mainly discuss rationality constraints and logical principles that have been very much discussed already, and try to get clear about the relation between them and discuss some objections to them. And, as already mentioned, the main aim of this chapter is to clarify and justify assumptions I make in the rest of this thesis.

Ordinary language conditionals are generally thought to be either *indicative* or *subjunctive*. The latter kind is also often called *counterfactual* and for the purposes of this thesis, I will take subjunctive and counterfactual conditionals to be the same thing. Another methodological issue that may be worth stressing at the outset is that I do not follow surface grammar when it comes to classifying conditionals. Thus although I call the conditionals that I will be

[1] If the semantical system does not include probabilities, then we might make it a minimal requirement on the system that it is *consistent* with the credence norms we accept for conditionals.

discussing *subjunctive*, it is not the case that they are, in general, expressed in the subjunctive mood. Nor are they reserved for cases where the antecedent is believed to be false, as the name *counter*factuals might suggest. Instead, I will, roughly speaking, classify conditionals according to how they are being used (as will be explained below). Unfortunately, classifying conditionals according to these properties in some cases seems very counterintuitive, as will become apparent.

The following two conditionals, and the way in which they interact with evaluation and choice, provide the main motivation for this project.

Conditional 1. *If I had chosen the less risky alternative, I would have been guaranteed an 'acceptable' outcome.*

Conditional 2. *If the coin had come up head, Ann would have gotten the kidney.*

These counterfactuals create well known problems for standard decision theory (as will be discussed in chapter 4). The truth of conditional 1 may make a particular situation *worse* (in particular a situation where the agent who expresses the conditional gets an unacceptable outcome as a result of turning down the less risky alternative), whereas the truth of 2 makes some situations *better* (namely a situation where Ann is dead as a result of not receiving a kidney).

As a way of analysing how the above conditionals affect rational decision making and evaluations, I will in chapter 5 compare them with some other conditionals. Some of these will be part of straightforward means-end reasoning, such as the following conditional, expressed by someone who wants to catch a flight:

Conditional 3. *If I take the Gatwick Express, I will reach the airport before my gate closes.*

Others use backward reasoning:

Conditional 4. *If I see my girlfriend tonight, then she must have missed her flight.*

Finally, there is a quite problematic class of conditionals that expresses some sort of causal relationship, such as the following:

Conditional 5. *If the Government prints more money, inflation will rise.*

All of these conditionals – and, more generally, conditionals with this form – influence the desirability of some factual proposition. And that is the reason I will discuss conditionals like these, alongside 1 and 2, in chapter 5.[2] The third conditional makes the proposition that the person in question takes the Gatwick express more desirable, the fourth conditional presumably reduces the desirability of the proposition that the person expressing the conditional sees his or her girlfriend that night, and the fifth conditional reduces the desirability of the proposition that the government prints more money.

Although some of the above conditionals are expressed in the indicative mood, many logicians and philosophers would classify them with counterfactuals; and in fact, it seems we could just as well express them (without a change in meaning) in the subjunctive mood. Conditionals 3 and 5 are important for reasoning about how to act (for 'me' in 3 but for 'the Government' in 5), which according to some makes them counterfactuals. Moreover, for all of these conditionals it is true that they are not probabilistically independent of their

[2]The contents of the conditionals that I discuss in chapter 5 differ slightly from the contents of these conditionals. But the form (and desirabilistic import) of each conditional discussed there corresponds to the form (and desirabilistic import) of at least one conditional discussed here.

antecedent. Some take all conditionals for which such independence fails to be counterfactuals. However, below we will see an example of an indicative conditional for which that principle fails.

In any case, in chapter 4 when I extend Jeffrey's desirability measure to conditionals, I will treat these conditionals as counterfactuals. But as will become evident, this should not cause problems for those who see these conditionals as indicatives. When applied to indicatives, the measure I use for counterfactuals automatically simplifies to the measure that is appropriate for indicatives (as I explain in chapter 5). Hence, from the perspective of the main aim of this thesis, not much hangs on how we classify conditionals 3 to 5. However, to be able to represent Allais' and Diamond's preferences as maximising desirability, we need to treat conditionals 1 and 2 as counterfactuals.

2.2 Probability of Conditionals

2.2.1 The Ramsey test

> If two people are arguing 'If p will q?' and are both in doubt as to p, they are adding p hypothetically to their stock of knowledge and arguing on that basis about q ([Ramsey, 1929]: 155).

The above suggestion of Ramsey's has been very influential in all areas of logic and philosophy that deal with conditionals and hypothetical reasoning. The suggestion is often called the *Ramsey test*. In its most general form, the Ramsey test, as I understand it, says that when determining whether to believe A $\rightarrow$ B, one should first *suppose* the truth of A, adjust one's beliefs as little as is compatible with that supposition, and finally determine whether one would believe B in the new (hypothetical) epistemic state.

The Ramsey test has considerable intuitive appeal, and has been accepted by many (perhaps most) people who have worked on the subject. Some have moreover accepted something stronger, namely that there is an *identity* between a rational agent's credence (or subjective probability) in A $\rightarrow$ B and her credence in B upon the supposition that A. Formally:

Thesis 3 (Strong Ramsey test)**.** $P(A \rightarrow B) = P_A(B)$

To the strong version of the Ramsey test, some people add the so-called *Lockean thesis* [Foley, 1992], which says that for any proposition α (factual or modal) a rational agent *accepts* (or 'categorically' believes) α whenever her credence in α is above some (context dependent) threshold. With this addition, the Ramsey test can be stated thus:

Thesis 4 (Lockean Ramsey test)**.** *One should accept $A \rightarrow B$ just in case (according to one's credence) $P_A(B)$ is sufficiently high.*

Although it seems plausible to me that some version of the Lockean thesis is true of how people actually reason, I will not assume the thesis in the remaining chapters. The thesis is known to have some counterintuitive implications, in particular the infamous Lottery and Preface paradoxes (see [Kyburg, 1961] and [Makinson, 1965]). Given a conditional version of the Lockean thesis, paradoxes like these show that the Lockean thesis leads to violations of a principle that is known as *Agglomeration*, which states that $A \rightarrow B$ and $A \rightarrow C$ together imply $A \rightarrow (B \wedge C)$. Since the semantics that I will use in this PhD thesis, and will defend in next chapter, implies that Agglomeration is a logical truth, I will assume that Agglomeration is

also a norm of rational reasoning with conditionals. Therefore, I cannot accept the Lockean thesis as part of rational reasoning with conditionals.

The Ramsey test does not say very much – or at least nothing very precise – unless we spell out what it means to *suppose a condition*. There are different ways one can suppose something to be true (see e.g. [Joyce, 1999]). In particular, there is a difference between *evidentially* supposing that A is true, that is, supposing it as a matter of fact, and *subjunctive* supposing A is true, that is, supposing it contrary to fact. The following example illustrates the difference between these two types of suppositions. When supposing that, *as a matter of fact*, Oswald did not kill Kennedy, most people (hypothetically) conclude that someone else killed Kennedy. After all, they strongly believe that Kennedy was murdered. When supposing that, *contrary to fact*, Oswald did not kill Kennedy, most people (hypothetically) conclude that nobody killed Kennedy, since they do not believe that a group of people was conspiring to murder the president.

One of the appeals of the Ramsey test is that since it is general enough to allow for different types of suppositions, it may provide a basis for a unified account of indicative and subjunctive conditionals. The right way to evaluate the probability of the indicative conditional $A \mapsto B$, according to a common view, is to *evidentially* suppose that A and then figure out the (suppositional) probability of B. To evaluate the probability of the subjunctive (or counterfactual) conditional $A \boxdot\!\!\rightarrow B$, on the other hand, we should *subjunctively* suppose A and figure out the probability of B. There is not much agreement on how to formally characterise subjunctive suppositions. However, I section 2.2.3 I will suggest one way of doing so. It is on, in contrast, generally accepted that evidential suppositions are characterised by Bayesian conditionalisation, which, as we will see in section 2.2.2, gives us a precise formula for how to evaluate the probability of indicative conditionals.

Without going into the formal details about the difference between the two types of suppositions, let me say in general and informal terms what I take their difference to be. When we subjunctively (or counterfactually) suppose a condition true, we hold fixed all general truths we believe (except when their negation is what we are supposing). In particular, we hold fixed what we believe the laws of the nature to be, what causal relations we accept, what dispositions (both physical and social) we believe in, etc. When we evidentially suppose a condition, however, this is not so. The (hypothetical) impact that evidentially supposing that A is supposed to have on our belief state should (more or less) match up with how we would change our beliefs upon *learning* that A is true. And learning A may very well lead us to update our beliefs about the laws of nature, natural and social dispositions, and so on.

The difference between the two ways of supposing a condition can also be explained in terms of when we vary our credence in a proposition. Roughly speaking, when we evidentially suppose that A, we vary our beliefs in all and only propositions that are *evidentially* related to A. When we subjunctively suppose that A, we vary our beliefs in a proposition if and only if it is somehow made true by A.[3] Evidentially supposing that A thus answers questions like: what hypotheses do I take A to support? what is A evidence for? what does the truth of the proposition indicate? Subjunctively supposing that A, on the other hand, is relevant to questions like: what *causal* difference would the truth of A make? what difference would the truth of A make to system *S*? how would person/society X react to the truth of A?

[3]This is not true of the suppositions that are relevant for backtracking subjunctives, such as 'if the grass had been wet this morning, then it would probably have rained over night'. Backtracking subjunctives will play some role in chapter 5, but I will largely ignore them in this chapter.

The above suggestion does not provide a very clear-cut distinction. Yet, I believe the above distinction is useful for the task at hand. Some examples may both help in illustrating and justifying the distinction. Take first the famous Oswald-Kennedy example again. I believe that nobody else was planning to murder Kennedy at the time, and that Oswald not murdering Kennedy would not cause anybody else to murder him. When I subjunctively suppose that Oswald did not kill Kennedy, I hold fixed my belief that nobody else was planning to kill Kennedy at the time. Thus I believe that if Oswald hadn't killed Kennedy, nobody else would have done so. When I, on the other hand, evidentially suppose that Oswald did not kill Kennedy, I vary my belief in the hypothesis that nobody else was planning to kill Kennedy. For given that I have previously learnt that Kennedy was in fact murdered, Oswald not killing him is pretty good evidence that someone else must have. Thus I believe that if Oswald didn't kill Kennedy, someone else did. (But I still believe that Oswald not killing Kennedy didn't *cause* someone else to kill him.)

Or consider the set up of the famous Newcomb's paradox [Nozick, 1969]. When I evidentially suppose that I choose both boxes, I conclude that the predictor has left the opaque box empty, since me choosing both boxes provides good *evidence* for that hypothesis. Subjunctively supposing that I take both boxes however does not affect my credence in hypotheses about the content of the opaque box. For what I decide now stands in no causal relationship to what is in the box; and my now choosing both boxes does not in any sense make it true that the predictor left the opaque box empty prior to my decision.

Finally, consider the following example. The Aragawa restaurant in Tokyo is apparently the most expensive restaurant in the world. Learning that someone frequently dines at such an expensive restaurant provides very good evidence for the hypothesis that the person in question is rich. Thus when I evidentially suppose, for instance, that my friend frequently dines at Aragawa, I (hypothetically) conclude that he must be very rich. When I subjunctively suppose that my friend frequently dines at this expensive restaurant, on the other hand, I hold fixed my beliefs about how much he owns and earns. Frequently dining at Aragawa does not cause people to become any richer, and in fact may causes my friend to go bankrupt. Thus when subjunctively suppose that my friend frequently dines at this fancy restaurant, I (hypothetically) conclude that he must be quite poor.

As we will later see, evidentially supposing a condition seems appropriate when evaluating an indicative conditional in accordance with the Ramsey test, whereas subjunctive suppositions seem appropriate for counterfactuals. Hence, the Ramsey test suggests different credence norms for these different types of conditionals. Before discussing these two types of conditional credence norms, I will briefly consider some objections to the Ramey test.

Objections to the Ramsey test

I will in this thesis more or less take the Ramsey test for granted and not provide much positive evidence for it. But in this section I would like to briefly reply to some common objections to the Ramsey test. The aim is not to conclusively defend the test against all possible objections, which would take us too far off topic. And I realise that what follows will probably not convinced those who are already skeptical of the RT. Nevertheless, I hope my discussion suggests that the objections that have been raised against the RT are not as

damaging as they may seem.[4]

(a) Confusing subjunctive suppositions with evidential Many objections to the Ramsey test (as well as to Adams' thesis, as we will see below) conflate evidential suppositions with subjunctive, and depend on using the wrong kind of supposition when evaluating a conditional. Here is one such example. Let us suppose we have strong credence in the following conditional:[5]

Conditional 6. *If the atoms in my desk simultaneously decay, the total amount of energy emitted would be x.*

We believe this conditional, let us suppose, because (and only because) physics tells us it is true. However, it would be very surprising, given our understanding of physics, if the antecedent of the conditional were to be realised. In fact, it would be so surprising that we would no longer trust the physics we have been taught, and would thus no longer have any reason to believe that the above conditional is true. Hence, we might be tempted to conclude, the Ramsey test cannot be true. Supposing the antecedent of the conditional should result in a similar change to our belief state as learning the antecedent. So whereas we take the conditional to be very likely to be true, we assign low credence to its consequent after supposing its antecedent.

The problem with the above objection to the Ramsey test is that it evaluates a counterfactual conditional as if it were indicative. Grammatically, the conditional in question seems indicative, as I formulated it above. But as already mentioned, we should not let surface grammar lead us astray, given that we take the logic of conditionals to determine the class to which they belong. And we could of course reformulate the conditional as: 'If the atoms in my desk *were to* simultaneously decay...'

The justification for calling the conditional in question counterfactual, is firstly that although we may not be certain that its antecedent will not be realised, we are pretty sure that it will not. But the fact that it will not does not in any way make it less informative. Hence, the conditional is certainly not what Jonathan Bennett calls *zero-intolerant*: "Indicative conditionals are mostly zero intolerant, meaning that such a conditional is useless to someone who is really sure that its antecedent is false" ([Bennett, 2003]: 45). Just like counterfactuals in general, the above conditional can, however, be very informative for those who are certain that the antecedent is false, since it tells them something about the nature and behaviour of atomic particles.

Secondly, the counterfactual in question expresses some sort of causal dependency, and such conditionals are in general classified with counterfactual (or subjunctive) conditionals, even when they are expressed in the indicative mood and their antecedent is considered true.

Finally, the conditional in question is not probabilistically independent of its antecedent, as we have just seen: the conditional probability of the conditional given its antecedent is lower than the unconditional probability of the conditional. It is widely accepted that indicative conditionals are usually independent of their antecedent. In fact, as we will see,

[4]The objections that follow can be found in various places in the philosophical literature on conditionals. The reason I focus on these objections is that they are the main reasons Alan Hájek, a notable critic of the Ramsey test, tells me he is sceptical of the test.

[5]Whether this is a realistic example or not does not matter for the present purposes. All that matters is the form of the argument that conditionals on this kind are problematic for the Ramsey test.

that must be true if Adams' thesis is to hold. Unfortunately, we will see an example below of an indicative where this does not seem to hold.

So conditional 6 is a counterfactual and we should thus evaluate its probability by supposing that *contrary-to-fact* its antecedent is true. And then we get the result we want. Recall that we are assuming that physics tells us that the total amount of energy emitted from all of the atoms in my desk decaying would be x. And when we counterfactually suppose a condition like the atoms in my desk decaying, we should not, as already mentioned, revise our credence in what physics tells us. Hence, when we suppose that, contrary to fact, all of the atoms in my desk simultaneously decay, we should have strong credence in the proposition that the total energy they emit is x. Contrary to appearance, however, we should not accept the conditional as an indicative, since we would not accept its consequent after having learnt its antecedent.

(a') McGee's counterexample Vann McGee [McGee, 2000] discusses a very interesting case that he takes to be a counterexample to the Ramsey test. Although it turns out not to be a counterexample to the Ramsey test, it does put great pressure on Adam's thesis which I discuss below.

The background to McGee's story is that a person called Murdoch has drowned and there has been an investigation into whether whether his business partner Brown murdered him. Having followed the case in the media, you find it most likely (although not certain) that Murdoch's death was an accident; and even more so if it turns out that Brown didn't murder Murdoch. In other words, you give low credence to the conditional:

Conditional 7. *If Brown didn't kill Murdoch, someone else did.*

But now a person you believe to be Sherlock Holmes, the great detective who actually has been leading the investigation, says that after looking at the evidence he is quite confident that Brown did kill Murdoch. Moreover, he is even more confident that Murdoch was murdered, for which he claims to have evidence that is independent of the evidence linking Brown to the murder. Suppose also that you believe that there is some, but (initially) very small, chance that the person you take to be Holmes is just someone pretending to be Holmes. But you are quite confident that that is not the case. Hence, having heard the verdict from the worlds greatest detective, you revise your credence in conditional 7, and now believe that it is very likely to be true.

Now the alleged problem for the Ramsey test is that if you were to *learn*, after all this, that Brown did not kill Kennedy, then you would no longer have high confidence in conditional 7. After all, you only had high confidence in the conditional since you accepted the judgement of the person you took to be Holmes. But this person claimed that Brown almost certainly killed Murdoch. Holmes, you believe, would almost certainly not say this unless it were true that Brown killed Murdoch. Hence, if Brown did not murder Murdoch, then you believe that the person you took to be Holmes isn't the great detective after all. Thus you change back to your initial judgement – or close to it – and assign low credence to conditional 7.

This example has some similarities to the atom decay example from above: in both cases, the conditional in question is not probabilistically independent of its antecedent. The difference between the two cases is that it seems very unnatural to classify conditional 7 with counterfactuals. For on a counterfactual (or subjunctive) reading, you would judge the conditional to be almost certainly false, even after learning the judgement of the person you take to be Holmes. For even if you believe then that Murdoch was almost certainly

murdered, most likely by Brown, the example gives us no reason for thinking that (you would believe that) someone else *would have* murdered Murdoch if Brown had not.

All is not lost however for the Ramsey test. The credence in conditional 7 does seem to correspond to your credence in the conditional's consequent after supposing the antecedent in a particular kind of way. Stefan Kaufmann has shown that credence in the conditional seems to satisfy a formula which is formally equivalent to the so-called *Skyrms' thesis* [Kaufmann, 2004].[6] And as we will later see, Skyrms' thesis can be seen as a special case of the Ramsey test when the supposition in question is *contrary-to-factual* (or subjunctive). So McGee's story is not a counterexample to the Ramsey test. What it does suggest, however, is that sometimes we evaluate indicative conditionals by supposing their antecedent in a way that is normally appropriate for the evaluation of counterfactuals. And it is hard to argue, as McGee's example shows, that such evaluations involve some sort of irrationality. Moreover, it suggests that Adams' thesis, which I discuss below, does not hold (neither as an empirical claim nor normative constraint) for all indicative conditionals.

As already stated, I take all conditionals to satisfy the Ramsey test. It would make life simpler if we could classify conditionals as indicative or subjunctive depending on whether the supposition used when evaluating them is matter-of-factual or counterfactual. Unfortunately, McGee's example shows that that is not feasible. The example also shows that we cannot always classify conditionals as indicative or subjunctive according to whether their evaluation goes by Adams' or Skyrms' thesis.[7]

(b) Thomason conditionals Many people have raised conditionals of the following kind as arguments against the Ramsey test:[8]

Conditional 8. *If my partner were to cheat on me I would not know about it.*

We can imagine various reasons why a person might accept the above conditional as true. However, when supposing that our partner is cheating on us, we must, it seems, assign full credence to the proposition that we know that our partner is cheating on us. (Otherwise we might end up assigning high credence to, and perhaps 'categorically believing', the sentence: my partner is cheating on me but I don't know it.) So, it may seem that the Ramsey test does not deliver the right result for conditionals like 8.

What conditionals like the one under discussion teach us, I think, is that when evaluating the probability of conditionals whose antecedent or consequent refer to our own beliefs, we may have to take up a third persons' perspective when applying the Ramsey test. When evaluating the probability of 8, for instance, we should imagine that 'my' in the antecedent refers to some person other than us – let's call this person ME – who is nevertheless in the exact same situation as we are, has the exact same beliefs about her partner, has a partner that (in all relevant respects) is exactly like ours, etc. Then we suppose that ME's partner cheats on her, and evaluate from that perspective the probability that ME knows about it; which we should assign low credence if we accept conditional 8.

[6] Kaufmann himself does however not seem to realise that the formula he comes up with as the correct way to measure our credence in McGee's conditional is formally equivalent to Skyrms' thesis.

[7] Another possible response to McGee's example is to simply deny that we should have high credence in conditional 7. Richard Bradley (in personal communication) has almost convinced me that this is how we should treat the conditional, and points out that since this conditional involves a nested supposition – first supposing that the speaker in question is Sherlock Holmes and then supposing that Brown didn't kill Murdoch – it should not come as a surprise if our intuitions about the conditional is mistaken. Although I see the force of that argument, I will set it aside for now, and take as given the intuition that the above story lends hight credence to 7.

[8] This type of example is, as far as I can tell, originally due to Richmond Thomason, but seems to have been first discussed in print by Bas van Fraassen [van Fraassen, 1980].

The above suggestion of how to apply the Ramsey test when either the antecedent or consequent of a conditional refers to our own beliefs also solves a problem identified by David Chalmers and Alan Hájek [Chalmers and Hájek, 2007]. They point out that according to some ideas about how to apply the Ramsey test, we should believe any conditional of the form 'if A then I believe A' and also any conditionals of the form 'if I believe A then A'. But believing the first means believing that one is omniscient, whereas believing the latter means believing that one is infallible. A credence norm that implies that rational agents have such high opinions of themselves seems very implausible.

Given how I suggest we apply the Ramsey test to cases like these, the test does not have any such implications. In the first example, you should suppose that A is true and on that basis evaluate the probability of the proposition that someone who is identical to yourself in all relevant respects (except that he or she has not supposed A) believes A. Only if you already believe that you are omniscient will you assign full credence to that proposition. In the second example, you suppose that someone who is exactly like yourself in all relevant respects believes that A and evaluate, from that perspective, the probability of A. Again, only if you already believe yourself to be infallible will you now assign full credence to A.

(c) Nested conditionals Nested conditionals are often cited as reasons why the Ramsey test must be false.Take the following examples:

Conditional 9. *If my partner would become mad if missed the dinner, then I would take a taxi rather than the bus.*

Conditional 10. *If John isn't very patient then he will eat before them if they are late.*

The first conditional seems to have the form $(A \rightarrow B) \rightarrow C$ whereas the latter has the form $A \rightarrow (B \rightarrow C)$. Neither conditional seems mysterious (although expressing them is somewhat awkward), and it seems we have no more difficulties in intuitively evaluating their probabilities than most 'simple' conditionals (i.e. conditionals with factual antecedent and consequents). However, people have taken conditionals like the above, in particular the first one, to be problematic for specific version of the Ramsey test, such as Adams' thesis which I discuss below.[9]

However, if (both indicative and counterfactual) conditionals express propositions, as they do given the semantics I discuss in next chapter, then nested conditionals pose no problems for the Ramsey test in its most general form. For then we can suppose a conditional just like any other proposition, which means that we can perfectly well evaluate $(A \rightarrow B) \rightarrow C$ by supposing that $A \rightarrow B$ and from that perspective evaluate C. To evaluate $A \rightarrow (B \rightarrow C)$, however, we should first suppose that A and from that perspective evaluate $B \rightarrow C$, which we do by supposing A, from the standpoint induced by the supposition that B, and from the final standpoint evaluate C.

(d) Impossibility results Finally, various formal results have been proved that many people have taken to show that the Ramsey test, and the specific versions of I discuss below, must

[9]The second (right nested) poses no problem for Adams' thesis, which says that (for indicative conditionals) $P(A \mapsto B) = P(B \mid A)$. According to AT: $P(A \mapsto (B \mapsto C)) = P_A(C \mid B) = P_A(C \wedge B)/P_A(B) = \frac{P(A \wedge C \wedge B)/P(A)}{P(A \wedge B)/P(A)} = P(C \mid A \wedge B)$. In other words, the probability that *if John isn't very patient then he will eat before them if they are late* is equal to the conditional probability of John eating before them given that he is not very patient and they are late.

The first (left nested) conditional however causes problems for Adam's thesis. Or rather, I should say that Adams' thesis (as presented in [Adams, 1975]) by itself does not determine how to evaluate the left nested conditional. However, given the multidimensional semantics, we should evaluate it as the probability of B under the evidential supposition that $A \wedge B$ (see [Bradley, 2011]).

be false. I will not discuss these results here. In section 2.2.2 I discuss one version of these results and in section 2.2.3 I discuss another version. At this point I will only say that these result need not worry us given that we adopt the semantics I discuss and defend in next chapter. (The impossibility results I discuss are based on probabilistic reasoning. Similar results have been proved based on non-probabilistic theories of rational belief revision. Unfortunately, I do not have room to discuss these results.)[10]

Deriving Adams' and Skyrms' theses from the RT

According to the so-called Adams' thesis (to be further discussed in later sections),[11] the probability of an indicative conditional equals the conditional probability of its consequent given its antecedent. More formally:

Thesis 5 (Adams' thesis (AT))**.** $P(A \mapsto B) = P(B \mid A)$

(As will become evident, we will have to change or limit the thesis slightly, to avoid well-known impossibility results, and the semantics I will use and discuss in next chapter suggests one such way to limiting the thesis.)

The so-called Skyrms' thesis[12] on the other hand (which I also discuss in more detail in later sections), holds that the probability of the counterfactual A $\Box\!\!\rightarrow$ B equals the expected conditional chance of B given A. Now if Ch_w is the *objective chance* function for world w – that is, a function that represent the objective probability distribution at that world[13] – and assuming a finite set W of possible worlds, the thesis under discussion can be formally stated as follows:[14]

Thesis 6 (Skyrms' thesis)**.** $P(A \Box\!\!\rightarrow B) = \sum_{w_i} P(W_i).Ch_{w_i}(B \mid A)$

where W_i is the proposition that world w_i is actual. So for each world w_i that an agent considers possible, she calculates the conditional chance of B given A at that world, weighs that by how probable she finds w_i to be actual, and sums the results of this calculation for all worlds.[15]

[10] Another potential problem with the Ramsey test, also suggested to me by Alan Hájek, is that the RT may make it seem strange how we could argue about the probability of a conditional. For the RT seems to suggest that the probability of a (indicative) conditional is always relative to the suppositional beliefs of the person who is considering the conditional. I am not, however, convinced that this should worry us too much. Even if there is no fact of the matter as to the probability of a particular statement, I may be able to give you various good or bad arguments for thinking that my probability assignment is reasonable. And it does not seem to be too unnatural to think that that is what is gong on when we argue about conditionals. In any case, *expressivists* have spent much effort discussing a very similar question in meta-ethics. It would take me too far from the aim of this thesis to try to evaluate or contribute to that debate.

[11] The thesis is named after Ernest W. Adams who famously suggests a version of it in [Adams, 1975]. The version of the thesis that I will be discussing, where P is taken to measure the probability of sentences' or propositions' truth, should perhaps be called 'Stalnaker's thesis', since Robert Stalnaker defended a view of this kind e.g. in [Stalnaker, 1970]. For Adams himself, however, P in AT should not be interpreted as measuring the probability of sentences' or propositions' truth, but rather something like their degree of assertibility.

[12] Named after Brian Skyrms who for instance discussed a version of the thesis in [Skyrms, 1981]. Like Adams, Skyrms himself does not take the thesis to determine the probability of a counterfactuals truth, but rather something like its assertability. Moreover, Skyrms himself speaks of propensities rather than chance.

[13] In section 2.3 I briefly discuss the notion of objective chance.

[14] Assuming a finite set of chance functions.

[15] In Skyrms' own formulation, the 'weights' (subjective probabilities) are on chance (or propensity) *functions*, rather than on worlds. But since I am interpreting each Ch_{w_i} function as correctly measuring the chances in world w_i, the probability that we are in world w_i and the probability that chance function Ch_{w_i} correctly describes chances at the actual world, comes (at least formally) down to the same thing.

In William Harper's ([Harper, 1981]: 24) formulation of Skyrms' thesis the weights are on *chance hypothesis*. (Skyrms' himself later formulates his thesis in similar fashion ([Skyrms, 1994]).) In other words, the thesis on this version is:

With the Ramsey test (and one additional assumption), it is possible to derive these two thesis from a version of David Lewis' famous *Principal Principle* [Lewis, 1980] which informally states that if an agent knows the objective chance of A, but does not have information about A that does not come through information about the chance of A, then her credence in A should match the chance of A. According to both a generalised and suppositional version of this principle, which we might the *Generalised Principal Suppositional Principle*, an agent should set her credence in B under the supposition that A according to her expectation of the conditional objective chance of B given A (see [Bradley, 2012] and [Williams, 2012]). Formally, the thesis we need is this:

Thesis 8 (GPSP). $P_A(B) = \sum_{W_i} Ch_{w_i}(B \mid A).P_A(W_i)$

Recall that the version of the RT that I have been endorsing states that $P_{\mathrm{A}}(\mathrm{B}) = \mathrm{A} \rightarrow \mathrm{B}$ (where '$\rightarrow$' represents either the indicative or subjunctive conditional connective). To get AT from GPSP and RT, we assume that when the supposition is evidential, $P_{\mathrm{A}}(W_i)$ equals $P(W_i \mid A)$. Then for evidential suppositions, we have $P^+_A(A) = \sum_i Ch_{w_i}(B \mid A).P(W_i \mid A) = P(B \mid A)$. By combining this with RT, we get Adams' thesis: $P(A \mapsto B) = P^+_{\mathrm{A}}(B) = P(B \mid A)$.

How do we justify the assumption that for evidential suppositions, $P_{\mathrm{A}}(W_i) = P(W_i \mid A)$? It follows from our assumption that evidential suppositions are characterised by conditional probabilities: if $P_{\mathrm{A}}(\cdot)$ represents an evidential supposition that A, then $P_{\mathrm{A}}(W_i) = P^+_A(W_i) = P(W_i \mid A)$. But let me give an example to intuitively motivate the assumption. Imagine that you have a cup in front of you and you don't know whether it is made of plastic or glass. Evidentially supposing that it is made of glass should affect what you (hypothetically) believe about chances in the actual world. In particular, it should affect your views about the chances of the cup breaking if dropped. More generally, evidentially supposing a sentence whose truth has implications for the causal structure of (some part of) our world should affect our beliefs about chances at the actual world. And the same no doubt holds true when evidentially supposing sentences that do not exactly tell us something about the causal structure of the world, but nevertheless tell us something about what the chances are.

From GPSP and RT we can also derive Skyrms' thesis, which, recall says that for any A, B, and any rational credence function P: $P(A \boxdot\!\!\rightarrow B) = \sum_{w_i} Ch_{w_i}(B \mid A).P(W_i)$. To get ST, we assume that when the supposition is contrary-to-factual, $P_{\mathrm{A}}(W_i) = P(W_i)$. To see why that assumption is justified, recall that supposing that A is true, contrary to fact, should not, *by itself*, change our views about what the *actual* world is like. (What we *infer* from such a supposition can however have drastic impacts on our beliefs about the actual world.) Instead, we are trying to figure out what counteractual worlds where A is true are like. Hence, unlike when we evidentially suppose that A, we should not, when we suppose A contrary to fact, update our credences concerning the objective chances of the actual world. For instance, supposing that, contrary to fact, the cup in front of me is made of glass, should not change my views about the chances of the cup breaking *in the actual world*.

Thus the GPSP and natural assumptions about evidential vs. subjunctive suppositions deliver both Adams' and Skyrms' theses for indicative and subjunctive conditionals respec-

Thesis 7. $P(A \boxdot\!\!\rightarrow B) = \sum_{H_i \in \mathcal{H}} P(H_i).Ch_i(B \mid A)$

where H_i is the hypothesis that Ch_i is the chance function that correctly describes chances in our world and $\mathcal{H}$ is the set of chance hypothesis the agent considers. Again, there is no real difference between this formulation and mine, if we assume that there is a one-to-one mapping between (conditional) chance functions and worlds, since the subjective probability that an agent attaches to the hypothesis that her world is correctly described by Ch_i should be the same as her subjective probability that the world described by Ch_i is actual.

tively. But, of course, the argument may not convince those who were sceptical of Adams' and Skyrms' theses; for instance since it relies on the Ramsey test.

2.2.2 Indicative conditionals

Adams' Thesis and Independence

Let us now briefly check whether the conditionals that I am most interested in for the purpose of this PhD thesis satisfy Adams' thesis. Recall that according to Adam's thesis, the probability of an (indicative) conditional equals the conditional probability of its consequent given its antecedent: $P(\mathrm{A} \mapsto \mathrm{B}) = P(\mathrm{B} \mid \mathrm{A})$. Above we saw how the Ramsey test potentially implies AT, given a version of the Principal Principle. We can however derive AT from RT without the last principle, if we make certain additional assumptions.

Recall that there are different ways one can suppose something to be true. How to formally characterise counterfactual suppositions is a vexed issue, which I will come back to below. But it is generally agreed that evidential suppositions are characterised by Bayesian conditioning. In other words, rational degree of belief in B upon the evidential supposition that A is given by:[16]

$$P^{+}_{\mathrm{A}}(\mathrm{B}) = P(\mathrm{B} \mid \mathrm{A}) = \frac{P(\mathrm{A} \wedge \mathrm{B})}{P(\mathrm{A})} \tag{2.1}$$

It is widely, but not universally, accepted that when evaluating conditionals according to the Ramsey test, the supposition appropriate for indicative conditionals is evidential. Combining this assumption with the Strong Ramsey test gives us Adams' thesis: $P(\mathrm{A} \mapsto \mathrm{B}) = P^{+}_{\mathrm{A}}(\mathrm{B}) = P(\mathrm{B} \mid \mathrm{A})$.

Adams' thesis has been used by Adams himself to formulate a logic that invalidates inference patterns that are classically valid (i.e. valid for the material conditional), but intuitively invalid for indicative conditionals [Adams, 1975]. But perhaps more importantly, the principle itself has great intuitive appeal and seems to accord with how we do – and should – evaluate the probability of conditionals. It seems that one should accept the (prototypical) indicative conditional

Conditional 11. *If Oswald didn't shoot Kennedy, then someone else did,*

exactly to the extent that one would accept that someone else shot Kennedy on the evidential supposition that Oswald didn't shoot him. The corresponding counterfactual conditional, on the other hand, shows that Adams' thesis does not hold in general for counterfactuals. Most people attach high conditional probability to someone else having shot Kennedy given that Oswald did not. However, most people assign low probability to following counterfactual:

Conditional 12. *If Oswald hadn't shot Kennedy, then someone else would have.*

So Adams' thesis is not generally true for counterfactuals. And neither is it true for all conditionals that are of interest to this thesis. Reflecting on conditional 5 (the money-printing-and-inflation example) illustrates this. This is an 'interventional conditional' [Bradley, ms] where (let us assume) the agent uttering the conditional is herself not capable of making the

[16] Alan Hájek [Hájek, 2003] and Branden Fitelson and Hájek [Fitelson and Hájek, ms] argue against the standard definition of conditional probability. Instead, they suggest we use a Popper definition of conditional probability; so take conditional probabilities to be primitive and analyse unconditional probabilities in terms of these. If they are right, then we can still accept that the evidential supposition is characterised by conditional probability; but then we should not use the ratio definition of conditional probability. Although the arguments against the ratio formula are convincing, I will ignore them for now. So keeping with the tradition, I will be using the ratio formula for conditional probability.

intervention in question (i.e. she is not part of the Government). Assuming that the speaker believes the government to be rational, inflation averse and knowledgeable about how to avoid inflation, it seems clear that Adams' thesis does not hold for this conditional. Let M represent the proposition that the Government prints money, and I the proposition that inflation goes up. Then while $P(\mathrm{M} \rightarrow \mathrm{I})$ is by assumption high, $P(\mathrm{I} \mid \mathrm{M})$ may very well (and rationally) be low (since $P^{+}_{\mathrm{M}}(\mathrm{M} \rightarrow \mathrm{I})$ is low). While the speaker believes, before learning that money has been printed, that if the Government prints money inflation will rise, she also believes that the government is rational and inflation averse. Hence, it may seem most natural for the agent in question to make room for the supposition that M by changing her belief in the conditional, rather than taking M it as (hypothetical) evidence that inflation will rise.

Arguments similar to the one above can be made to show that conditionals 1, 3, and 4 do not satisfy AT either. Those who take AT to hold true for all indicative conditionals will take this to justify classifying these conditionals as subjunctives. Unfortunately, McGee's example seems to show that AT does not hold for all indicatives. The conditional he discusses clearly seems to be indicative: the conditional actually seems true, on the most natural reading, but is false if we read it as a counterfactual. But rational evaluation of McGee's conditional does not seem to satisfy AT, as we have seen. Nevertheless, I will take it for granted that Adams' thesis holds for most indicative conditionals. (Later I discuss how to avoid Lewis' infamous triviality results). Hence, I suggest this provides initial reasons for classifying conditionals 1, 3, and 4 as indicatives (but recall that for the present purposes not much hangs on the classification).

A condition related to Adams' thesis is probabilistic independence between an indicative conditional and its antecedent:

Thesis 9 (Independence)**.** $P(A \mapsto B \mid A) = P(A \mapsto B)$

The informal argument that was meant to show that Adams' thesis does not hold for conditional 5 also shows that Independence fails in this case. We saw that when $P(M)$ increases, $P(M \rightarrow I)$ decreases (where $\rightarrow$ represents the conditional operator in conditional 5, whatever kind it is). And in line with symmetry of probabilistic dependence, as $P(M \rightarrow I)$ increases $P(M)$ decreases. So M and $M \rightarrow I$ are not probabilistically independent of one another. By similar reasoning, the same holds for conditionals 1, 3 and 4, if we assume, for instance, that the agents that these conditionals are about are knowledgable, rational and have the preferences that seem most natural to assume they have.

Independence is often taken to be a characteristic of indicative conditionals. In fact, as I will show below, Adams' thesis implies Independence, and Independence implies Adam's thesis given Modus Ponens (MP) and the Conditional Excluded Middle (CEM):

Thesis 10 (MP)**.** $\{A \wedge A \rightarrow B\} \vdash \{A \wedge B\}$

Thesis 11 (CEM)**.** $\emptyset \vdash (A \rightarrow B) \vee (A \rightarrow \neg B)$

I leave the proof that given MP and CEM, AT and Independence are equivalent to the appendix to this chapter (where this is proved as Theorem 11), since the proof relies on some results that I will prove in next section

Modus Ponens (MP) will be discussed in detail in next section. But for now, it suffices to say that when it comes to at least *simple* indicative conditionals – that is, indicative conditionals with factual antecedents and consequents – MP has much evidence in its

favour; and it is hard to see what role such conditionals would play in reasoning if MP were not true of at least simple indicative conditionals. In section 2.3.1 I will prove that Adam's thesis implies a condition called *Weak Centring* which is logically equivalent to MP. As I prove in 2.3.2, a probabilistic version of CEM is also implied by Adams' thesis (given a natural additional assumption).

Since AT holds for most indicatives, and AT implies Independence, most indicatives should satisfy Independence. But McGee's example illustrates, as already discussed, that not all indicatives satisfy Independence.

So some indicative conditionals, it seems, do not satisfy Independence. Moreover, some counterfactual conditionals *do* satisfy independence.[17] Most authors on conditionals would agree that the following conditional is a counterfactual conditional (even when its antecedent is true):

Conditional 13. *If a sugar cube is put in water, it will dissolve.*[18]

Nevertheless, this conditional satisfies Independence: the supposition that the cube has been put in water should neither increase nor decrease the probability of the conditional. So the fact that a conditional satisfies Independence does not mean that it is not a counterfactual.

Similarly, conditional 2 (the coin-flip-and-kidney example), does satisfy Independence, despite my wanting to classify it as a counterfactual. (The coin is, let us assume, being tossed to make a choice procedure fair, and it wouldn't, for instance, be fair to decide that Ann gets the kidney if it lands tails up, but then change the decision if one learns that the coin doesn't land tails up.) But that does not, in and of itself, mean that this is an indicative conditional. In fact, conditional 2 fails to satisfy a *desirability condition*, to be discussed in chapter 4, that all indicative conditionals should satisfy. Just to give a preview, the condition is that the truth of an indicative conditional makes no difference to a situation where its antecedent is false. Given a common conception of fairness, conditional 2 does not satisfy this desirability condition. In a situation where the coin comes up tails and Ann receives the kidney, it makes a difference whether or not it is true that had the coin come up heads Bob would have gotten the kidney.

Zero-tolerance

As a final reason for classifying the conditionals that are of interest to this thesis as counterfactuals, let me point out that they all fall into this category according to Jonathan Bennett's [Bennett, 2003] 'zero-tolerance rule'. The rule is supposed to conclusively tell whether an ordinary language conditional is indicative or subjunctive/counterfactual (assuming that any ordinary language conditional belongs to either of these classes): indicative conditionals are always what he calls 'zero-intolerant' whereas subjunctive conditionals are always 'zero-tolerant'. Being zero-intolerant, an indicative conditional is never of any use, Bennett claims, in a context where we know the antecedent to be false (i.e. where we attach *zero* probability to the antecedent). Subjunctive conditionals, on the other hand, are often very useful in such contexts – and, in fact, are often especially intended for such contexts (hence *counter*factuals.) The claim is not that an indicative conditional is *false* whenever its antecedent is false – if it were, then it would obviously not be completely probabilistically

[17] The same is true of Adams' thesis: while some subjunctives obviously do not satisfy the thesis, others seem to do so. For instance, the causal conditional 13 satisfies AT.

[18] It sounds of course more natural to simply say *D*: 'Sugar cubes dissolve in water'. But I think most would agree that the logical form of *D* is the same as that of conditional 8.

independent of its antecedent – but rather that an indicative conditional is 'useless' (i.e. completely uninformative) in such contexts, Bennett says (p. 45).[19]

Reflections on the two Oswald-Kennedy conditionals seem to support this view. The indicative Oswald-Kennedy conditional stated that if Oswald didn't kill Kennedy, then someone else did. We know what role this conditional plays if we have at least the slightest doubt as to whether or not Oswald in fact killed Kennedy. But what is the use of this conditional, *as an indicative*, if we *learn* that Oswald killed Kennedy? In such context, what does the conditional express? A popular view is that in such contexts, the conditional expresses nothing (see e.g. [Edgington, 1995] and [Bradley, 2002]). The corresponding counterfactual, which states that if Oswald hadn't killed Kennedy then someone else would have, however clearly expresses a proposition (assuming that counterfactuals ever express propositions) even if we know that in fact Oswald did kill Kennedy.

I will not here evaluate Bennett's zero-tolerance (nor Edgington's and Bradley's (old) view). But given this rule, the conditionals that motivate this thesis (i.e. conditionals 1 and 2), in the form that will most concern me, are clearly counterfactuals. For I will particularly be interested in the way they affect our evaluation of actual outcomes when we know their antecedent to be false. Since these conditionals have this effect when their antecedent is false, they are clearly not completely uninformative in such contexts.

Triviality results

In the late eighties David Lewis produced a pair of triviality results [Lewis, 1976] that many people have taken to be devastating for Adams' thesis. In this subsection I will discuss both the first of Lewis's original triviality result for Adams' thesis and Richard Bradley's argument that triviality results can be constructed for logically weaker theses than Adams'. Several other triviality results have been produced that I will not discuss,[20], since they all depend on an assumption about *how* the set of true conditionals is determined by to the set of true factual propositions that I will later reject.

Lewis's triviality results can be summarised by the following theorem:

Theorem 1 (Lewis' First Triviality Result)**.** *If Adams' thesis holds for all probability measures on a language $\mathcal{L}$, then $\mathcal{L}$ is trivial in the sense of not containing three sentences that have positive probability but are pairwise logically incompatible.*

Proof.
Assume the following:

1. $P(C \mid A) = \frac{P(A \wedge C)}{A}$ if $P(A) > 0$ [definition]
2. $P(A \mapsto C) = P(C \mid A)$ holds for *any* indicative where $P(A) > 0$ [AT]

Then the rest follows:

3. $P(A \mapsto C \mid B) = P(C \mid A \wedge B)$ when $P(A \wedge B) > 0$ [from 2]
4. $P(A \mapsto C \mid C) = P(C \mid A \wedge C) = 1$ when $P(A \wedge C) > 0$ [from 1, 3 and PC][21]
5. $P(A \mapsto C \mid \neg C) = P(C \mid A \wedge \neg C) = 0$ when $P(A \wedge C) > 0$ [from 1, 3 and PC]

[19] James Joyce has pointed out to me that the zero-one rule might not be as decisive as Bennett claims. As an example, he suggests that when sitting in a room with his dog, he might be certain that Obama is not in the room with him, but nevertheless believe that that if Obama is in the room with him then he is brilliantly disguised as a dog. And the latter, Joyce suggests, is an indicative conditional. Perhaps the zero-one rule should thus be used as a rule of thumb, rather than as a clear-cut and decisive rule for how to classify conditionals. In any case, the rule does not play a crucial role in my argument. But I will, nevertheless, occasionally use it as a heuristic.

[20] See [Hájek and Hall, 1994] for an overview of many of these results.

[21] 'PC' is short for 'probability calculus'.

6. $P(D) = P(D \mid B).P(B) + P(D \mid \neg B).P(\neg B)$ [Law of total probability]
7. $P(A \mapsto C) = P(A \mapsto C \mid C).P(C) + P(A \mapsto C \mid \neg C).P(\neg C)$ [from 6]
8. $P(A \mapsto C) = 1.P(C) + 0.P(\neg C).P(C)$ [from 7, 4 and 5]
9. $P(A \mapsto C) = P(C)$ [from 8]
10. So $P(C \mid A) = P(C)$ [from 9 and AT]

In other words, given our assumptions, A and C must be probabilistically independent. But that seems absurd. And given a non-trivial language – in particular, for the case in question, one in which $\neg A$, $A \wedge C$ and $A \wedge \neg C$ get assigned a positive probability – it is easy to construct a probability function for which this does not hold [Lewis, 1976]. □

Let us now consider Bradley's argument. From Adams' thesis it follows that:

Thesis 12 (Preservation Condition)**.**
If $P(A) > 0$ *and* $P(B) = 0$ *then* $P(A \mapsto B) = 0$*.*

For if $P(\mathrm{B}) = 0$ then $P(\mathrm{A} \wedge \mathrm{B}) = 0$; and if $P(\mathrm{A} \wedge \mathrm{B}) = 0$ and $P(\mathrm{A}) > 0$ then $P(\mathrm{A} \wedge \mathrm{B})/P(\mathrm{A}) = 0$. So the Preservation Condition is logically weaker than Adams' thesis. Moreover, it is arguably more intuitively plausible than AT, as Bradley points out ([Bradley, 2000] : 220):

> You cannot, for instance, hold that we might go to the beach, but that we certainly won't go swimming and at the same time consider it possible that if we go to the beach we will go swimming! To do so would reveal a misunderstanding of the indicative conditional (or just plain inconsistency).

Unfortunately, it turns out that the Preservation Condition generates a triviality result similar to Lewis':

Theorem 2 (Preservation triviality)**.** *If the Preservation Condition holds for all probability measures on a language* $\mathcal{L}$*, then* $\mathcal{L}$ *is trivial in the sense of not containing three sentences that have positive probability but are pairwise logically independent.*

(I will not provide a proof of Theorem 3, since the proof itself does not, I think, add anything to our understanding of the problem.)

It would be very restricting to have to work with only trivial languages. Moreover, I have been treating '$\mapsto$' as an ordinary language indicative conditional and claimed that AT usually holds for this conditional. But ordinary language is of course not trivial in the above sense. Hence, some assumption(s) in the above proofs must be false if the ordinary language indicative conditional satisfies AT. Lewis himself argued we should give up Adams' thesis, as a theory of the probability of truth of indicative conditionals, but suggested we should still accept AT as a criteria of *assertibility*. (Frank Jackson takes a similar stance [Jackson, 1991]). Others have taken the result to show that indicative conditionals do not express propositions (see e.g. [Gibbard, 1981]), and thus do not not satisfy the Boolean properties and/or the probability axioms that are assumed in the triviality proofs. In Lewis's proof above, this allows us to reject step 3 and perhaps the use of probability calculus (in Bradley's proof, this invalidates Lemma 1).

Later I will argue for none of these ways out of triviality. Instead, I will rely on a multidimensional semantics (to be discuss in next chapter) for which it is not true that conditional sentences are connected to factual sentences in the simple way that the triviality proofs assume. (A result of this is that the semantics implies that (sometimes) $P(\mathrm{A} \mapsto \mathrm{B} \mid$

B) = $P^{+}_{\mathrm{B}}(A \mapsto B) \neq P(\mathrm{B} \mid \mathrm{A} \wedge \mathrm{B})$. Hence, AT only holds for simple conditionals.) This avoids triviality in a similar way as denying that conditionals express propositions. However, the multidimensional semantics shows how this way out of triviality is compatible with maintaining that these conditionals are propositions (with truth values). Hence, we don't have to explain away their apparent truth-aptness. Moreover, it turns out that this way out of triviality coheres with my argument in subsequent chapters, where my examination of the desirability of conditionals establishes, or so I argue, that we cannot reduce conditionals to factual propositions.

2.2.3 Subjunctive conditionals

As already discussed, we can accept the Ramsey test for subjunctive conditionals without having to accept Adams' thesis for such conditionals, since the way in which we suppose the antecedent of a subjunctive conditional, when evaluating the conditional in accordance with RT, is generally not *evidential*. Instead, the supposition in such cases is *subjunctive* (or counterfactual). Two important attempts have been made at formalising the Ramsey test for subjunctive conditionals. Firstly, there is the thesis that subjunctive suppositions go by expected conditional chances, which given the Ramsey test implies Skyrms' thesis. Secondly, there is David Lewis's method of *Imaging* [Lewis, 1976], which I will discuss in the next subsection.

First, it might be useful to get Skyrms' thesis again on the table. Recall that on this view, for any rational person and any propositions A and B, the probability of A $\boxdot\!\!\rightarrow$ B equals the expected chance of B given A:

$$P(\mathrm{A} \boxdot\!\!\rightarrow \mathrm{B}) = \sum_{W_i} P(W_i).Ch_{w_i}(\mathrm{B} \mid \mathrm{A})$$

where W_i is the proposition that world w_i is actual. So for each world w_i that an agent considers possible, she calculates the conditional chance of B given A in that world, weighs that by how probable she finds w_i, and sums the results of this calculation for all worlds.

It is interesting to notice that various authors have suggested ways of measuring rational credence assignments to counterfactuals that, on a closer look, turn out to be formally equivalent to Skyrms' thesis. Robert Williams [Williams, 2012] suggests a way of formalising the subjunctive Ramsey test, based on what I previously called the Principal Suppositional Principle, that turns out to be identical to Skyrms' thesis (on a natural interpretation). And already mentioned, Stefan Kaufmann [Kaufmann, 2004] suggests a way of measuring our credence assignment to conditionals that fail to satisfy Adams' thesis, which also turns out to be formally equivalent to Skyrms' thesis. I will not discuss this further here, but one might take this as suggesting that Skyrms must have been on to something.

Lewis's *Imaging*

The second main attempt at formalising subjunctive suppositions (or something like it) is Lewis's method of *Imaging* [Lewis, 1976]. If w is the actual world, let w^{A} represent the world that would be actual if A were true; for convenience, let's say that w^{A} is the *closest* A-world to w. For any probability function P and any proposition A, the Image of P on A is a new probability function P^{A}, such that for any proposition B:

Definition 1 (Imaging). $P^{A}(B) \doteq \sum_{W_i} P(W_i).P(B \mid W_i^{A})$

(Recall that W_i is the proposition that world w_i is actual; in addition, let W_i^{A} be the proposition that w_i^{A} is the closest A-world to actuality.) Now if $w_i^{\mathrm{A}} \notin \mathrm{B}$ then $P(\mathrm{B} \mid W_i^{\mathrm{A}}) = 0$, but if $w_i^{\mathrm{A}} \in \mathrm{B}$ then $P(\mathrm{B} \mid W_i^{\mathrm{A}}) = 1$. So in calculating $P^{\mathrm{A}}(\mathrm{B})$, each w_i only counts for something if B is true in its closest A-world; and if it does count for something, it counts according to the probability that it is actual.

An example may make the idea clearer. Assume that we have four possible worlds, w_1, w_2, w_3 and w_4, and that $w_1, w_2 \in \mathrm{A}$ but $w_1, w_3 \in \mathrm{B}$. Moreover, assume that $w_1^{\mathrm{A}} = w_4^{\mathrm{A}} = w_1$ and $w_2^{\mathrm{A}} = w_3^{\mathrm{A}} = w_2$ (i.e. w_1 is the closest A-world to itself and also to w_4, etc.), and that $P(W_1) = P(W_4) = 0.3$ and $P(W_2) = P(W_3) = 0.2$. Then $P(\mathrm{B}) = 0.5$.[22] Finally, $P^{\mathrm{A}}(\mathrm{B})$ is then calculated as follows:

$$\begin{aligned} P^{\mathrm{A}}(\mathrm{B}) = P(W_1).P(\mathrm{B} \mid W_1^{\mathrm{A}}) + P(W_2).P(\mathrm{B} \mid W_2^{\mathrm{A}}) \\ + P(W_3).P(\mathrm{B} \mid W_3^{\mathrm{A}}) + P(W_4).P(\mathrm{B} \mid W_4^{\mathrm{A}}) \\ =0{,}3(1)+0{,}2(0)+0{,}2(1)+0{,}3(1)=0{,}6 \end{aligned} \tag{2.3}$$

Lewis [Lewis, 1976]) shows that given a particular semantics of counterfactuals, namely Robert Stalnaker's [Stalnaker, 1968], we have:

Thesis 13 (Imaging thesis)**.** $P(A \boxrightarrow B) = P^A(B)$

Since the semantics I use in chapter 4 to introduce counterfactuals to Jeffrey's decision theory entails (given certain additions I endorse) Skyrms' thesis but not the Imaging thesis, I will not discuss Imaging further.

Subjunctive triviality

Triviality results similar to those discussed in section 2.2.4 have been constructed for subjunctive versions of the Ramsey test. Williams [Williams, 2012] constructs one based on the assumption that $Ch(\mathrm{A} \boxrightarrow \mathrm{B}) = Ch(\mathrm{B} \mid \mathrm{A})$. I will not reproduce Williams' proof, since it is exactly Lewis' triviality proof that I discussed above, with the exception that it applies to counterfactuals rather than indicatives, and P is substituted for Ch and $P(\cdot \mid \cdot)$ for $Ch(\cdot \mid \cdot)$. Thus what I later say about ways to avoid Lewis' triviality result applies also to Williams'.

Hannes Leitgeb ([Leitgeb, 2012a], section 2) constructs a triviality result that is based on the same principles as Lewis', but whose structure differs from Lewis' in an illuminating way (and is directly aimed at Skyrms' thesis). It is thus worth reproducing it here:

Theorem 3 (Leitgeb triviality)**.** *If subjunctive conditional $A \boxrightarrow B$ has classical truth values at world w_i (and is determined by the facts of w_i), then Skyrms' thesis (as formulated above) implies that if $P(W_i) > 0$, then the conditional chance of B given A in w_i must be either 0 or 1; and is 1 if and only if $A \boxrightarrow B$ is true at w_i.*

Proof.

Lemma 1. *First we prove that $P((A \boxrightarrow B) \wedge W_i) = P(W_i).Ch_{w_i}(B \mid A)$*

Proof.

[22] Assuming that we can calculate the probability of proposition A by:

$$P(\mathrm{A}) = \sum_{W_i} P(W_i).v(\mathrm{A}, w) \tag{2.2}$$

1. $P((\text{A} \boxdot\!\!\rightarrow \text{B}) \wedge W_i) = P(W_i).P(\text{A} \boxdot\!\!\rightarrow \text{B} \mid W_i)$ [def. of $P(\cdot \mid \cdot)$]
2. $= P(W_i).\sum_{W_j} P(W_j \mid W_i).Ch_{w_j}(\text{B} \mid \text{A})$ [Skyrms' thesis]
3. $= P(W_i).Ch_{w_i}(\text{B} \mid \text{A})$ [for $\sum_{W_j} P(W_j \mid W_i)$ reduces to $P(W_i \mid W_i) = 1$]

□

Now either A $\boxdot\!\!\rightarrow$ B is (a) true or (b) false at w_i. Assuming (a) we have:

1. $P((\text{A} \boxdot\!\!\rightarrow \text{B}) \wedge W_i) = P(W_i).Ch_{w_i}(\text{B} \mid \text{A}) = P(W_i)$ [from Lemma 1 and (a)]
2. $Ch_{w_i}(\text{B} \mid \text{A}) = 1$ [from 1 since $P(W_i) > 0$]

Assuming (b) we have:

1. $P((\text{A} \boxdot\!\!\rightarrow \text{B}) \wedge W_i) = P(W_i).Ch_{w_i}(\text{B} \mid \text{A}) = 0$ [from Lemma 1 and (b)]
2. $Ch_{w_i}(\text{B} \mid \text{A}) = 0$ [from 1 since $P(W_i) > 0$]

□

So Leitgeb's proof shows that, given the above assumptions (in particular, that counterfactuals are either true or false at a world), that conditional chances (at worlds) are 'trivial' (or 'crisp'): always either 0 or 1. Leitgeb's way out of triviality involves making a distinction between, on one hand, rational degree of belief in a counterfactual, and, on the other hand, the acceptability of a counterfactual. And he assumes that an agent's degree of belief in a counterfactual can come a part from the degree to which she accepts the counterfactual ([Leitgeb, 2012a], section 2).

I will not discuss Leitgeb's solution in detail, but simply state that I find this assumption unwelcome and believe it should be avoided if possible. And as I will show in next chapter, the multidimensional semantics that I will adopt in this thesis invalidates one of the assumptions of the triviality result: the assumption that any counterfactual is either true or false in any particular world. But once we give up that assumptions, we can no longer assume, as Leitgeb does in his proof, that when A $\boxdot\!\!\rightarrow$ B is true at w_i, $P((\text{A} \boxdot\!\!\rightarrow \text{B}) \wedge W_i) = P(W_i)$. And indeed, this will not be true in general in the multidimensional semantics (as we will see in next chapter).

2.3 Logic of Counterfactuals

Independently of any formal semantics, there are various logical principles that we might want counterfactuals to satisfy (or fail to satisfy). In this section I will only discuss three such principles: Centring, Modus Ponens and the Conditional Excluded Middle. The motivation for focusing on these three is that getting clear on what we think about these principles is important for the discussion of the semantics of conditionals in next section.

2.3.1 Centring and Modus Ponens

The term Centring in the present context comes (I believe) from David Lewis [Lewis, 1986a], whose *semantics* has a Centring condition with certain logical implications. I will however start by discussing these logical implications, and will, for simplicity, refer to them as *Strong*

and *Weak Centring* (and later use the prefix 'Semantic' to refer to Lewis' semantic Centring conditions). These are the two logical principles:[23]

Thesis 14 (Strong Centring)**.** *For all A, B:* $(A \wedge B) \supset (A \rightarrow B)$

Thesis 15 (Weak Centring)**.** *For all A, B:* $(A \rightarrow B) \supset (A \supset B)$

Below I will first examine the logical relationship between the two Centring conditions and their relationship to Modus Ponens. I will then consider arguments against Centring and Modus Ponens.

Logical relations between WC, SC and MP

Recall that Weak Centring states that that $(A \rightarrow B) \supset (A \supset B)$ is a logical truth. Substituting $\neg B$ for B throughout the axiom makes it easier to see the relation to Strong Centring. With this substitution, the axiom is:

$$(A \rightarrow \neg B) \supset (A \supset \neg B) \tag{2.4}$$

By the truth conditions of the material conditional, WC is identical to:

$$(A \rightarrow \neg B) \supset \neg(A \wedge B) \tag{2.5}$$

By contraposition:

$$(A \wedge B) \supset \neg(A \rightarrow \neg B) \tag{2.6}$$

With this formulation of WC it is easy to show how it follows from SC, given that we accept the following condition (as well as Modus Ponens for the material conditional which is of course uncontroversial):

Thesis 16 (Conditional Consistency (CC))**.**
For any consistent A: $(A \rightarrow B) \supset \neg(A \rightarrow \neg B)$

CC seems to hold at least for the kind of indicatives for which AT is most plausible; i.e. indicative conditionals whose antecedent gets assigned a probability greater than 0 (the principle is e.g. endorsed in [Bradley, 2007a]). It should not, however, be taken to hold for conditionals with inconsistent antecedents, at least if we assume (as is done in classical logic) that anything follows from a contradiction (sometimes called the *Principle of Explosion*).

Theorem 4. *SC and CC imply WC.*

Proof.

(1) $A \wedge B \supset (A \rightarrow B)$ [SC]
(2) $A \wedge B$ [Assumption]
(3) $A \rightarrow B$ [1, 2 and MP for '$\supset$']

[23] Both are entailed by Lewis' strongly centred (semantic) system of spheres, but only the latter by his weakly centred system. Recall that on Lewis' semantics, $A \boxrightarrow B$ is (non-vacuously) true at a world w just in case B is true in *all* the closest A-worlds to w. (A counterfactual is vacuously true at w when its antecedent is impossible at w.) In Lewis' strongly centred semantical system, when A is true at w, w itself is *the* closest A-world to w. Hence, if both A and B are true at w, then so is $A \boxrightarrow B$, as Strong Centring states; but if A is true and B is false, then $A \boxrightarrow B$ is false at w, as Weak Centring states (by contraposition and given the truth conditions for $\supset$). In Lewis' weakly centred system, if A is true at w then w is *one of* the closest A-worlds to w. Hence, if A is true but B is false at w, then $A \boxrightarrow B$ is false w.

(4) $\neg(A \rightarrow \neg B)$ [3 and CC]
Hence, $\emptyset \vdash (A \wedge B) \supset \neg(A \rightarrow \neg B)$

□

How can we then get from Weak to Strong Centring? One way to do so is to ad the Conditional Excluded Middle. Recall from last section that CEM states that for any propositions A and B, it is a logical truth that either $A \rightarrow B$ or $A \rightarrow \neg B$ is true. (So from $\neg(A \rightarrow B)$ one can infer $A \rightarrow \neg B$.) I will discuss the plausibility of CEM in in section 2.3.2. CEM and Conditional Consistency may seem closely related, but actually neither principle implies the other: CEM says that *at least* on one of $A \rightarrow B$ and $A \rightarrow \neg B$ is true; CC that (for a consistent A) *at most* one of $A \rightarrow B$ and $A \rightarrow \neg B$ is true; so together they imply that (for a consistent A), *exactly* on one of $A \rightarrow B$ and $A \rightarrow \neg B$ is true.

Theorem 5. *WC and CEM imply SC.*

Proof. WC can be formulated as: $(A \wedge B) \supset \neg(A \rightarrow \neg B)$. By CEM, $\neg(A \rightarrow \neg B)$ entails $A \rightarrow B$. So by WC and CEM, $(A \wedge B) \supset (A \rightarrow B)$, which is just SC. □

Finally, let us examine what happens if we give up WC. The main logical consequence of giving up this axiom is that that Modus Ponens (MP) is no longer guaranteed to hold, even for *simple* conditionals (i.e. conditionals with factual antecedents and consequents). In fact, WC is both necessary and sufficient for MP.

Theorem 6. *MP holds for '$\rightarrow$' if and only if WC holds for '$\rightarrow$'.*

Proof.

MP implies WC
(1) $(A \wedge (A \rightarrow B) \vdash B$ [MP]
(2) $\emptyset \vdash (A \wedge (A \rightarrow B)) \supset B$ [from 1]
(3) $\emptyset \vdash (A \rightarrow B) \supset (A \supset B)$ [exp.]
So MP implies WC.

WC implies MP
(1) MP fails only if $((A \rightarrow B) \wedge (A \wedge \neg B)) \nvdash \bot$
(2) $(A \rightarrow B) \supset (A \supset B)$ [WC]
(3) $(A \rightarrow B) \supset \neg(A \wedge \neg B)$ [from 2]
So WC implies MP (from 1 and 3) □

In next section we will see arguments that suggest that, intuitively, SC may be too strong. However, since the semantics I use throughout this thesis entails CEM, this section shows that I cannot abandon SC without also giving up WC. But that is equivalent to abandoning MP which I think is a too high price to pay (for reasons that will become apparent). Hence, I must to accept SC.

The plausibility of Centring and MP

Given the role that indicative conditionals play in evidential reasoning, it is hard to deny Strong Centring for such conditionals. And in fact, SC must hold for indicatives if Adams' thesis does. The same is true for Weak Centring.

Theorem 7. *AT implies SC.*

Proof. Probabilistic SC (for indicatives): $P(\mathrm{A} \wedge \mathrm{B}) \leq P(\mathrm{A} \mapsto \mathrm{B})$. SC given AT: $P(\mathrm{A} \wedge \mathrm{B}) \leq \frac{P(\mathrm{A} \wedge \mathrm{B})}{P(\mathrm{A})}$, which must holds since it is equivalent to $P(\mathrm{A}).P(\mathrm{A} \wedge \mathrm{B}) \leq P(\mathrm{A} \wedge \mathrm{B})$. □

Theorem 8. *AT implies WC.*

Proof. Probabilistic WC (for indicatives): $P(\mathrm{A} \mapsto \mathrm{B}) \leq P(\mathrm{A} \supset \mathrm{B}) = P(\neg \mathrm{A}) + P(\mathrm{A} \wedge \mathrm{B})$. WC given AT: $\frac{P(\mathrm{A} \wedge \mathrm{B})}{P(\mathrm{A})} \leq P(\neg \mathrm{A}) + P(\mathrm{A} \wedge \mathrm{B})$. Thus $P(\mathrm{A} \wedge \mathrm{B}) \leq P(\neg \mathrm{A}).P(\mathrm{A}) + P(\mathrm{A}).P(\mathrm{A} \wedge \mathrm{B})$; and hence, $1 \leq \frac{P(\neg \mathrm{A}).P(\mathrm{A})}{P(\mathrm{A} \wedge \mathrm{B})} + \frac{P(\mathrm{A} \wedge \mathrm{B}).P(\mathrm{A})}{P(\mathrm{A} \wedge \mathrm{B})}$; so $1 \leq \frac{P(\neg \mathrm{A}).P(\mathrm{A})}{P(\mathrm{A} \wedge \mathrm{B})} + P(\mathrm{A})$.

This last claim holds if and only if: $P(\neg \mathrm{A}) \leq \frac{P(\neg \mathrm{A}).P(\mathrm{A})}{P(\mathrm{A} \wedge \mathrm{B})}$. To see that that holds, notice that it can be written as $P(\neg \mathrm{A}).P(\mathrm{A} \wedge \mathrm{B}) \leq P(\neg \mathrm{A}).P(\mathrm{A})$ which is true since $P(\mathrm{A} \wedge \mathrm{B}) \leq P(\mathrm{A})$. □

So given that most indicatives satisfy Adam's thesis, most indicatives should satisfy both Strong and Weak Centring. However, in previous sections I suggested that we think of the conditionals that I will mostly be considering as subjunctive (or counterfactual) conditionals. Hence, the question we need to ask is whether SC and WC are plausible conditions on subjunctive reasoning.

Several authors writing on the logic of subjunctive conditionals have rejected Strong Centring as a logical truth for such conditionals. As pointed out by Hannes Leitgeb, "it is at least questionable to assume that the mere presence of facts which are described by A and B is capable of establishing a counterfactual dependence of B on A ... and even so by pure logic," ([Leitgeb, 2012b]: 87). It is moreover not hard to come up with examples that seem to support such rejection of subjunctive SC. Take the following example:

Conditional 14. *If I were writing a chapter of my thesis right now, the Universe would have been created by the explosion of a tiny mass 13.75 billion years ago.*

In general, when the antecedent, A, is not in any way connected to the consequent, B, a counterfactual $\mathrm{A} \mathrel{\square\!\!\rightarrow} \mathrm{B}$ may sound very strange, to say the least, even when both A and B are true.

Moreover, it can be even harder to accept conditionals with a true antecedent and a consequent that *are* connected, but connected in the 'wrong way' (as Alan Hájek points out [Hájek, ms]). Take the following conditional:

Conditional 15. *If the Labour Party had held on to their seat in the City of Durham in the 2010 General Election, the Conservative Party would have won the election.*

The Labour candidate for Durham did hold on to her seat. But *in spite of that*, the Conservatives won the General election. And this illustrates exactly why SC has implications that seem counterintuitive: normally when someone utters a subjunctive conditional, we expect the antecedent to somehow explain, or at least contribute to the truth of, the consequent.

It could perhaps be argued that the aforementioned expectation is due to a rule of *pragmatics* rather than the logic (or semantics) of counterfactuals; perhaps the rule that it is *inappropriate* to express $\mathrm{A} \mathrel{\square\!\!\rightarrow} \mathrm{B}$ unless A and B are related in the right way. Thus while the above two conditionals may be *true*, their utterance is nevertheless *inappropriate* or *odd*. This seems to have been the view of David Lewis ([Lewis, 1986a]: 28), who reminds us of the importance of not conflating oddness with falsity. Since I need to accept SC, as previously mentioned, I will have to use some version of this argument to explain how the last two conditionals can be true in spite of it seeming inappropriate to utter them.

Let me briefly mention another implication that SC and WC have together: they imply two worlds that differ in actual facts must differ in counterfactuals; in other words, actual (or non-modal) facts supervene on counterfactuals.

Theorem 9. *SC and WC together entail that non-modal facts supervene on counterfactuals.*

Proof. Let **F** denote the set of factual sentences that are true at world w_i (and suppose **F** is closed under conjunction). Let **C** denote the set of counterfactual sentences true at w_i. Suppose $A, B \in \mathbf{F}$. Then given SC, $(A \boxdot\!\!\rightarrow B) \in \mathbf{C}$. Now suppose we change the truth value of B (but not A); call the new/changed world w_i', let $\mathbf{F}'$ denote the set of factual sentences true at w_i' and $\mathbf{C}'$ the set of counterfactual sentences true at w_i'. Then $B \notin \mathbf{F}'$ and (given bivalence) $\neg B \in \mathbf{F}'$. But $A \in \mathbf{F}'$ iff $A \in \mathbf{F}$. Hence by WC, $(A \boxdot\!\!\rightarrow B) \notin \mathbf{C}'$. Thus, if two possible worlds differ in the truth value of some factual sentence, then they also differ in the truth value of some counterfactual. □

However, WC by itself does not entail that facts supervene on counterfactuals. Suppose that B is true at w_i, but no counterfactual with a true antecedent and B as consequent is true at w_i, as is consistent with WC being true but SC false. In this case, changing the truth value of B, but leaving the truth values of all other sentences unchanged, will not necessarily change what counterfactuals are true at w_i. Hence, facts do not supervene on counterfactuals.

It may be natural to think that for any tautology T, and for any sentence A, A and $T \boxdot\!\!\rightarrow A$ are logically equivalent, which would mean that (non-modal) facts supervene on counterfactuals, since then changing a (non-modal) fact would change at least one counterfactual. But without Centring that is not the case. Any semantics that does not imply the two logical Centring conditions can be used to show this, but since I haven't yet presented the semantics I will use, let me explain this in terms of Lewis' famous similarity semantics. Although A is true in the actual world w_i, there might, if either (semantic) Centring condition does not hold, be a world w_j such that: T is true at w_j, w_j is at least as similar to w_i as w_i is to itself, but A is false at w_j. Hence, if Centring fails, then A might be true but $T \boxdot\!\!\rightarrow A$ false.

Humeans like David Lewis (whose semantics implies both Strong and Weak Centring) contend that modal facts supervene on non-modal ones. (Lewis himself called his view 'Humean' [Lewis, 1987], [Lewis, 1994], after David Hume.) Some Humeans may want to make the following stronger claim: modal facts supervene on non-modal ones but the most fundamental facts, which according to most Humeans (e.g. Lewis) are non-modal, don't supervene on any less fundamental facts. But then they cannot accept SC and WC, since together these two entail that *all* non-modal facts supervene on counterfactuals (since changing the truth value of *any* factual sentence changes the truth value of some counterfactual). As I point out in [Stefánsson, 2014d] (where I first presented the above theorem), it seems that Lewis himself was only committed to the weaker of the aforementioned Humean views. (This seems apparent from his [Lewis, 1983]: 358.) I myself am far from convinced that we should accept the stronger Humean view. Hence, the lesson I take from Theorem 9 is simply that *if* we want to accept the stronger Humean view, then we must at last abandon SC.[24]

[24]Wlodek Rabinowicz has pointed out to me that the supervenience concept on which the above argument is based is very thin, and that strong Humeans might have a stronger concept of supervenience, according to which one fact supervenes on another just in case the latter is true *because* the former is. Given such a supervenience concept, it need not be the case that all non-modal facts supervene on modal ones, even if both Strong and Weak Centring are true. It seems to me, however, that the thin concept of supervenience has become standard in the literature (see e.g. http://plato.stanford.edu/entries/supervenience/), and that the strong Humean that Rabinowicz has in mind would say that non-modal facts *ground* modal facts (where 'grounding' is understood as an asymmetrical relation). But the general point of Rabinowicz's might nevertheless be true; i.e., that the argument presented above does not really

A few authors on the logic of counterfactuals also reject Modus Ponens and thus Weak Centring for such conditionals (see for instance [Gundersen, 2004], [Leitgeb, 2012a] and [Leitgeb, 2012b]). As will become apparent in chapter 5, it makes things much easier if I can assume MP. But more importantly, it seems that Modus Ponens plays an indispensable role in our use of counterfactuals in practical deliberation. I claim that one part of the desirabilistic import of counterfactuals stems from the fact that we use them to reason about the effects of various interventions (such as actions). The most simple example is when deciding what to do: if I am trying to make up my mind as to whether I should take the tube or cycle to school, I try to determine what would happen if I were to make each of these choices. If I believe, for instance, that if I were to cycle I would become wet, then that reduces the desirability of cycling (since I dislike being wet). It is hard to see how counterfactuals could play this role in practical deliberation unless we think that given the aforementioned counterfactual, a situation in which I cycle is one in which I get wet. So Modus Ponens seems to play an indispensable role in practical counterfactuals reasoning. Therefore, I will assume that MP holds for at least simple, or non-nested, counterfactuals.[25]

Let me finish this discussion of Modus Ponens by pointing out that an argument by Jonathan Bennett, which on the face of it my seem to be a counterexample to MP for counterfactuals (or subjunctives), is actually unproblematic for the principle. However, the example does illustrate the limited *applicability* of MP for counterfactuals. Bennett says:

> Because a subjunctive conditional is zero-tolerant, one can properly accept it while knowing that one would not be willing to use it in Modus Ponens. Last year I went to Spain; I am pretty sure that If I had not visited Spain last year I would have visited France. However, if I consider the implications of my discovering to my amazement that I did not visit Spain, they do not lead to the conclusion that I went to France. On the contrary, if I add 'I did not visit Spain last year' to my belief system with its multitude of memories and other evidences of my having done so, the resulting system makes me unwilling to have any opinion about what I did last year ([Bennett, 2003]: 230).

On a closer look, it is clear that this is not an argument against MP for counterfactuals (nor is it clear that the example is intended as such). MP would fail as a general rule of inference if and only if we could consistently *simultaneously* endorse $A \boxrightarrow B$ and $A \wedge \neg B$. (That is, if and only if $((A \boxrightarrow B) \wedge (A \wedge \neg B)) \nvdash \bot$.) But that is not what Bennett's example shows. Let S represent the proposition that Bennett visited Spain last year and F the proposition that he visited France. Bennett believes $\neg S \boxrightarrow F$ and S. If, however, he were to learn that $\neg S$, he

force those we would typically call 'strong Humeans' to abandon Centring, since such people are not committed to the view that fundamental facts don't supervene (in the thin sense) on less fundamental ones. Nevertheless, my (admittedly quite weak) conditional claim stands: if one endorses strong Humeanism, as the view is defined it in last paragraph, then one should abandon at least Strong Centring.

[25] My discussion will only concern simple conditionals; i.e conditionals with factual antecedents and consequents. Vann McGee was perhaps first to point out, by help of the following examples, that MP fails for even *indicative* conditionals with conditional consequents [McGee, 1985]. Opinions polls taken right before the 1980 US Presidential Election showed that the Republican candidate Reagan had a considerable lead over the Democrat Carter, with Anderson, the second Republican in the race, as a distant third. Hence, people had a good reason to believe that:

Conditional 16. *If a Republican wins, then if it is not Reagan that wins, it will be Anderson who wins.*

But despite most people believing (and having a good reason to believe) that a Republican would win, they did not (and had no reason to) believe:

Conditional 17. *If Reagan does not win, it will be Anderson who wins.*

since the second most likely presidential candidate after Reagan at the time was Carter.

would, he says, not be willing to have any opinion about what he did last year. Although he does not explicitly say so, it seems to me that this must include $\neg S \boxdot\!\!\rightarrow F$; in other words, upon learning that he did, in spite of all evidence to the contrary, actually not visit Spain, he becomes unwilling to have an opinion about *any proposition* that concerns what he did last year, including any *subjunctive* proposition. So upon learning $\neg S$, he stops affirming $\neg S \boxdot\!\!\rightarrow F$, which means that MP is not violated.

Bennett's example thus turns out to be very similar to the examples I gave to show that there may be probabilistic dependency between a subjunctive conditional and its antecedent. The examples had the following form: I believe that if the Government prints more money then inflation will rise; but I also believe that the Government is aware of this, is rational, well informed and inflation averse; hence, upon learning that the Government has printed money, I drop my believe in the conditional rather than concluding that inflation will rise.

While these examples are no counterexamples to MP, they nevertheless do point to limitations in the *applicability* of MP for counterfactuals compared with the more 'ordinary' indicative or material MP. As Bennett points out, it would be inappropriate to assert or even accept an indicative conditional $A \mapsto B$ while *knowing* that one would not be willing to use it to infer that B upon learning the truth of A (ibid). The same of course holds for the material conditional (which is only 'asserted' in formal logic or mathematics). These examples however show that this is not the case with subjunctive conditionals. Hence, while the acceptance of an indicative or material conditional is always a commitment to use it in MP whenever one learns the truth of its antecedent, this is not so with subjunctive conditionals. This should perhaps not come as a surprise. For after all, subjunctive (or *counter*factual) conditionals are very often intended to be used in circumstances where their antecedent is known, or strongly believed, to be false. Hence, it is no wonder that they are often not intended to be used in Modus Ponens.

Objective Chance, Skyrms' thesis and Centring

Certain views on chances and how credences in chances hypothesis should change imply that Skyrms' thesis is inconsistent with the two Centring conditions. Other views on these issues however make Skyrms' thesis compatible with Centring. Below I briefly describe a view that seems quite attractive, and which together with ST implies that Centring fails. Since I am, as already mentioned, committed to both ST and Centring, I will have to adopt another (and, as it happens, more popular) view on chances.

According to a theory of chance that Carl Hoefer has defended [Hoefer, 2007] (and Hoefer and Roman Frigg have further developed [Frigg and Hoefer, 2010], [Frigg and Hoefer, ta]), both the future and the past can be chancy. That the future can be chancy, if anything can, is uncontroversial. But the idea that the past can be chancy is quite controversial. Here is how Hoefer puts it ([Hoefer, 2007]: 554):

> [M]y coin flip at noon yesterday was an instance of a chance setup with two possible outcomes, each having a definite objective chance. It was a chance event. The chance of heads was 1/2. So 1/2 is the objective chance of A [the proposition that the coin I flipped at noon yesterday lands heads]. It still is; the coin flip is and always was a chance event. Being to the past of me-now does not alter that fact, though as it happens I now know A is false.

The details of Hoefer's theory does not matter for the present purposes. But to put it simply,

the idea is that chances supervene on facts about event-*types*. So although the outcome of the event *token* in question (i.e. the toss of a particular fair coin at a particular time) certainly was that the coin landed heads, that token event does not, by itself, really affect the chance of A (given there is a very great number of token events of this type), since the chance of A is determined by the chance of events of the type in question.

Assuming Hoefer's view on chance, the following seems a natural view on credence in chance hypothesis. For some chance hypotheses, e.g. those that concern systems or types of events that we do not have much experience with or knowledge of, our credence in the hypotheses should not be very robust with respect to experience. For example, if I am asked to assign credence to chance hypothesis regarding who of two tennis players, of whom I have no knowledge, will win the match, then I would presumably assign highest credence to the hypothesis that they have an equal chance of winning. However, as I observe one of them easily win two sets in a row, I will, presumably, change my credence assignment in the aforementioned chance hypothesis.

However, for other chance hypotheses, concerning systems or types of events that you have great knowledge of, your credence in chance hypotheses about that system or type of events should be more robust with respect to experience. Suppose you have built an indeterministic computer that has been programmed to generate either event of type A or B when button P is pushed, such that the chance of it generating B when the button is pushed is 1 in a million; otherwise it generates A. You programmed it yourself, and you are very confident in your abilities as a programmer. Thus the subjective probability you attach to the proposition 'the chance that the computer generates B, if the button is pushed, is 0.000001' is close to 1. That is, if B represents the proposition that the computer generates B at time t^+ and P the proposition that the button is pushed at (an earlier) time t^-, then you have close to full credence in the proposition that the actual world w is such that $Ch_w(\mathrm{B} \mid \mathrm{P}) = 0.000001$. What happens now if you observe the computer generating B at time t^+ after the button is pushed at time t^-? Presumably, that depends on what you have experienced before. If you just programmed the computer and this was its first output (and you believe that the past can be chancy) your confidence in $Ch_w(\mathrm{B} \mid \mathrm{P}) = 0.000001$ need not decrease at all, given how confident you are in your abilities as a programmer.

When the above view on chance and credence in chance hypotheses is combined with Skyrms' thesis, probabilistic versions of both Strong and Weak Centring are violated (and with the latter, Modus Ponens for subjunctive conditionals). Recall that according to the former condition, Strong Centring, $(\mathrm{A} \wedge \mathrm{B}) \supset (\mathrm{A} \boxdot\!\!\rightarrow \mathrm{B})$, while the latter, Weak Centring, states that $(\mathrm{A} \boxdot\!\!\rightarrow \mathrm{B}) \supset \neg(\mathrm{A} \wedge \neg\mathrm{B})$. Stated probabilistically, the two Centring conditions are:

Thesis 17 (Probabilistic Strong Centring)**.** $P(A \wedge B) \leq P(A \boxdot\!\!\rightarrow B)$

Thesis 18 (Probabilistic Weak Centring)**.** $P(A \boxdot\!\!\rightarrow B) \leq P(\neg A \vee B)$

To see how Skyrms' thesis violates the two Centring conditions, given the proposed understanding of chance and its relation to credence, consider again the computer example above. (Let A be the proposition that the computer generates A at time t^+.) Now if you observe that the button is pressed at time t^+ (P) and the computer generates B ($\neg$A), your credence in $\mathrm{P} \wedge \neg\mathrm{A}$ should be close to 1 (assuming that you trust your observation). However, given that you are confident in your programming abilities, and thus still confident that that the objective conditional chance of $\neg$A given P is one in a million, you should, according to Skyrms' thesis, give very low credence to the subjunctive conditional $\mathrm{P} \boxdot\!\!\rightarrow \neg\mathrm{A}$. So Prob-

abilistic Strong Centring fails. Moreover, since $P(\mathrm{P} \wedge \neg \mathrm{A}) \approx 1$, $P(\neg \mathrm{P} \vee \mathrm{A}) \approx 0$ even though according to Skyrms' thesis $P(\mathrm{P} \boxdot\!\!\rightarrow \neg \mathrm{A}) \approx 1$. So Weak Centring fails.

Given other views on chance, Skyrms' thesis is however consistent with both Centring conditions. According to David Lewis [Lewis, 1980], chances evolves such that whenever A turns out to be true the chance of A becomes 1 (and the chance of $\neg$A becomes 0). If we, moreover, suppose that rational agents' credence in chance hypothesis evolve in the same way, then whenever an agent learns $\mathrm{A} \wedge \mathrm{B}$, she should, according to Skyrms' thesis, assign full credence to $\mathrm{A} \boxdot\!\!\rightarrow \mathrm{B}$ (thus satisfying SC) and not accept $\mathrm{P} \boxdot\!\!\rightarrow \neg \mathrm{B}$ (thus satisfying WC). Since I will, in the rest of this thesis, take both Centring conditions as given, and also endorse Skyrms' thesis, I will have to assume that chances and rational credence in chance hypotheses evolve the way Lewis suggests. However, I will not defend this theory of chance. Chance is currently a very hotly debated topic, a discussion of which would take us too far afield.

I should, however, add one more thing concerning the view on chance that I will assume. Lewis famously thought that non-trivial objective chance and determinism were mutually incompatible. To take an example, coin tosses obey the laws of mechanics, which means that whether they land heads or tails is fully determined by their initial conditions and other environmental factors. Hence, some, like Lewis, want to say that the chance of a coin landing heads or tails on a particular occasion is either 0 or 1. In later chapters (in particular chapter 6), I will assume that when a fair coin is properly tossed, it has a 0.5 (objective) chance of landing either heads or tail. In this regards my view is in agreement with Frigg and Hoefer's (see in particular [Frigg and Hoefer, 2010]). However, other authors on chance have modified Lewis' theory such that it allows for 'deterministic (non-trivial) chances' without implying that the past can be chancy (see for instance [Loewer, 2001]).

To sum up: consistency requires that I adopt a view on chance that implies, firstly, that only the future is chancy, and, secondly, that determinism is compatible with non-trivial chance. I will not discuss the theory of chance in any more detail. Hoefer and Frigg's view on chance will reappear in the next chapter, where I show that contrary to appearances, the multidimensional semantics is not incompatible with their view on chance. Even though I do not, at this point in time, endorse Hoefer and Frigg's view on chance, I don't think that our semantics for conditionals should exclude the view, which some find very attractive. Hence, I do think that it is worth showing that the multidimensional semantics is consistent with Hoefer and Frigg's view.[26]

2.3.2 Conditional Excluded Middle

As already mentioned, the *Conditional Excluded Middle* (CEM) is the view that, for any propositions A and B, at least one of $\mathrm{A} \rightarrow \mathrm{B}$ and $\mathrm{A} \rightarrow \neg \mathrm{B}$ is true. More formally:

$$\text{For any A, B: } (\mathrm{A} \rightarrow \mathrm{B}) \vee (\mathrm{A} \rightarrow \neg \mathrm{B})$$

CEM is widely widely held to be true of *indicative* conditionals. In fact, a probabilistic version of CEM, $P((\mathrm{A} \rightarrow \mathrm{B}) \vee (\mathrm{A} \rightarrow \neg \mathrm{B})) = 1$, is implied by Adams' thesis, if we assume what I called the *Conditional Consistency Condition* (CC):

Theorem 10. *AT implies Probabilistic CEM (given consistent antecedents).*

[26] As an autobiographical note, it might be worth admitting that I have myself on previous occasions found myself drawn to their Hoefer and Frigg's view.

Proof.

1. $P(A \rightarrow B) = P(B \mid A)$, $P(A \rightarrow \neg B) = P(\neg B \mid A)$ [AT]
2. $P(B \mid A) + P(\neg B \mid A) = 1$ [Probability calculus (PC)]
3. $P(A \rightarrow B) + P(A \rightarrow \neg B) = 1$ [From 1 and 2]
4. $P((A \rightarrow B) \wedge (A \rightarrow \neg B)) = 0$ [CC]
5. $P((A \rightarrow B) \vee (A \rightarrow \neg B)) = 1$ [From 3 and 4 and PC] □

More controversial is whether CEM holds for *subjunctive* conditionals. In this section I will discuss arguments for and against CEM for subjunctive conditionals (admittedly mainly for); arguments that do not directly depend on any particular semantics for such conditionals.

But first a terminological note: to simplify the discussion, for any conditional $A \rightarrow B$, I will call $A \rightarrow \neg B$ the *converse* of the first. Given this terminology, the CEM implies that if a conditional is false, then its converse is true. (In other words, given CEM, the the negation of a conditional implies the conditional's converse.)

Intuitive arguments

Intuition seems to go in both directions regarding the subjunctive CEM (SCEM). Here is an intuitive argument that it holds. Let us assume the ordinary principle of excluded middle (EM), i.e. that for any A, either A or $\neg A$ is true (I don't know of any *intuitive* argument against EM). Now intuitively, when we express subjunctive conditionals, such as $A \boxrightarrow B$, we make claims about hypothetical situations where it is true that A but otherwise differ minimally from the situation in which the conditional was expressed. So to evaluate the truth of $A \boxrightarrow B$, we 'zoom in on' hypothetical situations where it is true that A. Given the ordinary EM, whatever situation we zoom in on, either B or $\neg B$ will be true in that situation. Moreover, if we zoom in on more than one situation, and perhaps find no principled way of arbitrating between them, it will be true in *any* of these situations that either B or $\neg B$. So it seems we have an intuitive argument for CEM for subjunctive conditionals.

Someone might object that the fact that we may have no principled way of determining whether the $A \wedge B$ situations or the $A \wedge \neg B$ situations are relevant for determining the truth value of $A \boxrightarrow B$, is a reason against CEM for subjunctive conditionals.[27] But that thought is in my view mistaken. There are many factual propositions, in particular propositions concerning future events and the future acts of free agents, whose truth value is (even in principle) impossible for us to determine. For instance, there is, as far as I can tell, no reasonable way for me, or anyone else, to determine the truth value of the proposition that I have a sandwich for lunch on December 1 2050. Even knowing all the facts of our world won't settle the matter, if we admit that people have, at least to some degree, freedom of will. Nevertheless, either I will have a sandwich for lunch that day or I won't. The fact that we have no principled way of deciding between future paths of our world where I do and where I don't have the sandwich is thus, in and off itself, no argument against the ordinary excluded middle: in all of these paths, the proposition in question is either true or false. Similarly, the fact that we often have no principled way of determining whether the $A \wedge B$ situations or $A \wedge \neg B$ situations are relevant for determining the truth value of $A \boxrightarrow B$ – and would not be able to do so even if we knew all the facts of our world – is, in and off itself, no reason against subjunctive CEM.

[27] I thank Wlodek Rabinowicz for pressing me on this issue.

David Lewis [Lewis, 1986a] has offered an example which many people seem to find intuitive as an argument against subjunctive CEM. My view is that Lewis' argument is an instance of the mistake discussed in last paragraph. Consider the following conditional:

Conditional 18. *If Verdi and Bizet had been compatriots, both would have been Italian.*

Since Bizet was actually French and Verdi Italian, it seems plausible that if they had been compatriots, then either Bizet had been Italian or Verdi had been French. It would be hard to give an explanation of why, if we add the antecedent hypothetically to our stock of knowledge, we would hypothetically conclude that they are both of some other nationality. So if they had been compatriots, either both or neither had been Italian. But is conditional 18 true? No, says Lewis, since when we zoom in on situations just like ours except that Bizet and Verdi are compatriots, we will find that both are Italian in some of them and both are French in others. From this Lewis concludes that 18 must be false (since on his semantics, A $\boxdot\!\!\rightarrow$ B is true only if B is true in *all* the situations that we zoom in on to evaluate A $\boxdot\!\!\rightarrow$ B). Does that mean that the following conditional is true?

Conditional 19. *If Verdi and Bizet had been compatriots, it is false that both would have been Italian.*

By the same reasoning as before – zooming in on the relevant situations, Bizet and Verdi are both French in some and both Italian in others – Lewis claims that conditional 19 is false. If that is the case, then we have $\neg(A \boxdot\!\!\rightarrow B) \wedge \neg(A \boxdot\!\!\rightarrow \neg B)$, which means that $\neg(\neg(A \boxdot\!\!\rightarrow B) \vee (A \boxdot\!\!\rightarrow \neg B))$. So subjunctive CEM fails.

I must admit that my pre-theoretical intuition, although not very strong, conflicts with what Lewis's theory predicts. In fact, Lewis' himself admits that this implication of his theory is not among its most intuitive parts.[28] Others have taken the Veri-Bizet example to be a convincing counterexample to the subjunctive CEM (see e.g. [Joyce, 1999]: 65). However, what I said about the proposition that I will have a sandwich for lunch on December 1 2050 also holds, as far as I can tell, for the the Bizer-Verdi counterfactuals. That is, although the facts of our world cannot even in principle determine their truth value, it is nevertheless the case that they are either true or false.[29]

So intuition seems to go both ways when it comes to the subjunctive CEM. Moreover, when evaluating this particular principle, I find it very hard to not first consider the more general question 'what makes a (subjunctive) conditional true?' When one has already decided on an answer to the latter question, the intuition concerning the former question becomes much clearer. For instance, given Lewis' semantics for counterfactuals, the Verdi-Bizet example does provide a counterexample to the subjunctive CEM. But that, in itself, does of course not mean that the principle is false. However, if we can derive CEM from more fundamental (and more widely accepted) principles, then that would, in my view, provide much stronger arguments for this highly abstract principle than any concrete examples can be expected to do. In next subsection I will discuss two such arguments.

[28] Lewis in fact goes so far as saying that what he wants to say about the Bizer-Verdi counterfactual intuitively sounds like a *contradiction* ([Lewis, 1986a]: 80).

[29] Robert Stalnaker suggests that since since the facts of our world do neither determine the truth nor falsity of the Bizer-Verdi counterfactual, it is *indeterminate* (see e.g. [Stalnaker, 1984], [Stalnaker, 1981]). Given a supervaluationist account of indeterminacy, he then salvages CEM. A similar argument can be (and often is) made for propositions concerning future events, and the actions of free agents, to explain why the ordinary excluded middle is true, even though the facts of our world don't often neither determine the truth or falsity of a proposition.

Formal arguments

Charles B. Cross [Cross, 2009] provides a more formal argument for subjunctive CEM that relies on a principle Cross calls 'Maximal Preservation'. Cross (ibid, 178) formulates the principle as follows:

Thesis 19 (Maximal Preservation)**.** *Every self-consistent counterfactual supposition preserves as much of the actual truth as possible while consistently accommodating the truth of what is counterfactually supposed.*

How much is 'possible' to preserve given any supposition is of course relative to the agent's total belief, the condition supposed true and the mode of the supposition. But it seems reasonable that for any particular belief set S and any particular proposition A, it should at least be determined how much it is possible to preserve of S while counterfactually (or subjunctively) supposing A. Now suppose that subjunctive CEM fails. Then it must be possible that there are propositions A and B such that both A $\boxdot\!\!\rightarrow$ B and A $\boxdot\!\!\rightarrow \neg$B are false; in other words, there *exists a possible world* where both A $\boxdot\!\!\rightarrow$ B and A $\boxdot\!\!\rightarrow \neg$B are false. For any particular world w, call the set $\{\alpha : \text{A} \boxdot\!\!\rightarrow \alpha\}$ the 'set of A's counterfactual consequents' in that world; let $\{\alpha : \text{A} \boxdot\!\!\rightarrow_w \alpha\}$ represent that set in world w. Then when subjunctive CEM fails, there is a proposition A and a world w such that the set of A's counterfactual consequents at w neither includes B nor $\neg$B. Given the ordinary excluded middle, however, either B or $\neg$B will be true at w. Now suppose that B is true at w. Then Maximal Preservation is violated, since $\{\alpha : \text{A} \boxdot\!\!\rightarrow_w \alpha\} \bigcup \{\text{B}\}$ is consistent (none of the αs are $\neg$B), accommodates A and preserves strictly more than $\{\alpha : \text{A} \boxdot\!\!\rightarrow_w \alpha\}$ of what is true at w. By same reasoning, Maximal Preservation is violated if $\neg$B is true at w. If CEM does not fail at w, however, then Maximal Preservation is not violated, since $\{\alpha : \text{A} \boxdot\!\!\rightarrow_w \alpha\} \bigcup \{\text{B}\}$ will then either be inconsistent or preserve equally as much as $\{\alpha : \text{A} \boxdot\!\!\rightarrow_w \alpha\}$ of what is true at w. So if we want to accept Maximal Preservation, we must accept subjunctive CEM.

Robert G. Williams [Williams, 2010] discusses another formal argument for subjunctive CEM based on first and second order logic. Williams' argument has three premises (which I have slightly modified):

- First premise: The following are equivalent:

  A: No student would have passed if they had goofed off
  B: Every student would have failed to pass if they had goofed off.

- Second premise: A and B can be formalised respectively as follows (where S is the predicate '... is a student', G is the predicate '... goofs off' and A is the predicate '...passes'):

  A*: $\neg\exists x(Sx \supset (Gx \boxdot\!\!\rightarrow Px))$
  B*: $\forall x(Sx \supset (Gx \boxdot\!\!\rightarrow \neg Px))$

- Third premise: $\forall F[\neg\exists x(Gx \supset Fx) \leftrightarrow \forall x(Fx \supset \neg Gx)]$

Now by premise 3, A* is equivalent to:

C*. $\forall x(Sx \supset \neg(Gx \boxdot\!\!\rightarrow Px))$

But then C* is identical to B*: C* is equivalent to A*, which by premise 2 is equivalent to A, which by premise 1 is equivalent to B, which in turn by premise 2 is equivalent to B*.

The argument can be generalised:

Proof.

The following seem to be logical truths:
1. $\forall F, G, H[\neg\exists x(Fx \supset (Hx \boxdot\!\!\rightarrow Gx)) \leftrightarrow \forall x(Fx \supset (Hx \boxdot\!\!\rightarrow \neg Gx))]$
2. $\forall \mathcal{F}, \mathcal{G}[\neg\exists x(\mathcal{F}x \supset \mathcal{G}x) \leftrightarrow \forall x(\mathcal{F}x \supset \neg\mathcal{G}x)]$
But from this it follows that:
3. $\neg\exists x(Fx \supset (Hx \boxdot\!\!\rightarrow Gx)) \leftrightarrow \forall x(Fx \supset \neg(Hx \boxdot\!\!\rightarrow Gx))$ [By 2]
4. $\forall x(Fx \supset \neg(Hx \boxdot\!\!\rightarrow Gx)) \leftrightarrow \forall x(Fx \supset (Hx \boxdot\!\!\rightarrow \neg Gx))$ [By 1 and 3]
Now since 4 holds for *all* F, G, and H, $\neg(Hx \boxdot\!\!\rightarrow Gx)$ and $(Hx \boxdot\!\!\rightarrow \neg Gx)$ are logically equivalent. □

So we have good general reasons for accepting CEM. Moreover, the conditionals that I use to motivate this thesis clearly satisfy CEM. Therefore, it might be possible to deny CEM for some subclass of the set of subjunctive conditionals, but nevertheless accept that they hold for the conditionals that I am working with. This is fortunate, since the semantics that I use when introducing counterfactuals to Jeffrey's decision theory (and discuss in next chapter) implies CEM for both indicatives and subjunctives.

2.4 Concluding Remarks

I will conclude by just briefly summarising what I have done in this chapter. I have argued that conditionals in general should satisfy the Ramsey test, Centring and the Conditional Excluded Middle; and that *most* indicative conditionals in addition satisfy both Independence and Adams's thesis, but that subjunctive conditionals in general satisfy neither of these. Instead, subjunctive conditionals satisfy Skyrms' thesis. In next chapter I will discuss and further develop a recent *Multidimensional Possible World Semantics for Conditionals*.

Appendix: AT and Independence (proof)

Theorem 11. *Given MP and CEM, AT and Independence are equivalent.*

Proof. In proving this theorem I will rely on the following Lemma:

Lemma 2. *SC and MP together entail that* $(A \wedge B)$ *is logically equivalent to* $(A \wedge (A \rightarrow B))$.

Proof.

(1) $(A \wedge B) \supset (A \rightarrow B)$ [SC]
(2) $(A \wedge B) \supset A$ [by def. of '$\wedge$']
(3) $(A \wedge B) \supset (A \wedge (A \rightarrow B))$ [from 1 and 2]
(4) $(A \wedge (A \rightarrow B)) \supset B$ [by MP]
(5) $A \supset A$ [logical truth]
(6) $(A \wedge (A \rightarrow B)) \supset (A \wedge B)$ [by 5 and 6]
Hence, $(A \wedge B)$ is logically equivalent to $(A \wedge (A \rightarrow B)$ [from 3 and 6] □

Since Adams' thesis implies both SC and MP, the Lemma shows that AT implies that $(A \wedge B)$ is logically equivalent to $(A \wedge (A \rightarrow B))$; which of course implies that $P(A \wedge (A \rightarrow B)) = P(A \wedge B)$.

Now notice that we can equivalently state Independence as follows:

$$P(\mathrm{A} \wedge \mathrm{A} \rightarrow \mathrm{B}) = P(\mathrm{A}).P(\mathrm{A} \rightarrow \mathrm{B}) \tag{2.7}$$

(since $P(\mathrm{A} \rightarrow \mathrm{B} \mid \mathrm{A}) = P(\mathrm{A} \rightarrow \mathrm{B} \wedge \mathrm{A})/P(\mathrm{A})$.)

AT by itelf implies Independence: $P(\mathrm{A} \wedge \mathrm{A} \rightarrow \mathrm{B}) = P(\mathrm{A} \wedge \mathrm{B}) = P(\mathrm{B} \mid \mathrm{A}).P(\mathrm{A}) = P(\mathrm{A} \rightarrow \mathrm{B}).P(\mathrm{A})$ (the second equality follows from the definition of conditional probability; the last equality from AT).

Independence and MP+CEM together imply AT: We have seen that CEM and Weak Centring, which as we have seen is logically equivalent to MP, together imply Strong Centring (SC). Hence, given what was said above, it follows from CEM+MP that $P(\mathrm{A} \wedge (\mathrm{A} \rightarrow \mathrm{B})) = P(\mathrm{A} \wedge \mathrm{B})$. Thus, we have given Independence, MP and SC: $P(\mathrm{A}).P(\mathrm{A} \rightarrow \mathrm{B}) = P(\mathrm{A} \wedge \mathrm{A} \rightarrow \mathrm{B}) = P(\mathrm{A} \wedge \mathrm{B})$. Hence, $P(\mathrm{A} \rightarrow \mathrm{B}) = \dfrac{P(\mathrm{A} \wedge \mathrm{B})}{P(\mathrm{A})} = P(\mathrm{B} \mid \mathrm{A})$.

So, given MP and CEM, AT and Independence are equivalent. □

Chapter 3

Multidimensional Semantics for Conditionals

3.1 Introduction

In this chapter I discuss the semantical system that I will later use to introduce counterfactual prospects to Richard Jeffrey's decision theory. The main justification for using this particular semantics is that it works quite well for these decision-theoretic purposes. But in this chapter I try to show that the semantics is, independently of such purposes, quite plausible and solves some of the problems we saw in last chapter. The main original contribution of the chapter is that I show that the semantics is more permissive and general than Richard Bradley's [Bradley, 2012] original formulation of it suggested. Moreover, I show that contrary to what Bradley's formulation may have suggested, it is consistent with a particular (Humean) metaphysical view, which some people find attractive. As I briefly explained in last chapter, I am myself no longer entirely convinced of the metaphysical view that I discuss in this chapter. However, I think that our semantics for conditionals shouldn't be inconsistent with a metaphysical view that many people endorse. Hence, I think that it is important to establish that the two are consistent. (All sections of this chapter except 3.6 and 3.8 are from [Stefánsson, 2014c].)

3.2 Probability and Metaphysics of Conditionals

Recall that according to what we might call the 'Ramsey view', there is a tight connection between the probabilities of conditionals and conditional probabilities. Moreover, recall from last chapter that there are two types of conditionals in ordinary language, roughly speaking corresponding to two types of probabilities. The probabilities of one type of conditionals, often called *indicative conditionals*, equal the *subjective* conditional probabilities of their consequents given their antecedents. (In last chapter I called this view on indicatives 'Adams' thesis'.) The other type of conditionals is often called *subjunctive conditionals*, and the probabilities of such conditionals, according to this version of the Ramsey view, equal (roughly) the *objective* conditional probabilities of their consequents given their antecedent. (Recall that for the present purposes, I take subjunctives and counterfactuals to be the same kind of conditional.) Of course, one rarely has direct access to objective chances (unlike subjective probabilities). However, this view says that one should set one's credence in a counterfactual according to ones *expectation* of the objective conditional chance of its consequent given its antecedent. (In last chapter I called this 'Skyrms' thesis'.)

According to a popular view on the metaphysics of modality, which I will call the 'Humean Supervenience view', modal facts – i.e. facts about conditionals, objective proba-

bilities, etc. – supervene on non-modal ones. In other words, if two possible worlds differ in what modal facts are true at them, then they also differ in what non-modal facts are true at them. David Lewis, who was the first to give the name 'Humean Supervenience' to the metaphysical view under discussion,[1] expressed what I take to be the two ideas that together give the Humeans Supervenience view its intuitive appeal, when he claimed that, firstly, all contingent facts are implied by the most fundamental facts, and, secondly, the most fundamental facts concern spatiotemporal relations and local qualities (see e.g. [Lewis, 1994]: 474-4). Together, these two ideas entail that all facts, modal or non-modal, are implied by a particular class of non-modal facts, namely, those concerning spatiotemporal relations and local qualities. But if that is the case, then two possible worlds[2] that differ in modal facts must differ in non-modal facts. In other words, modal facts supervene on non-modal facts.[3]

According to the version of the Humean Supervenience view that I will focus on in this chapter, both past and future events can be chancy (for a defence of such a view, see e.g. [Hoefer, 2007], [Frigg and Hoefer, 2010] and [Frigg and Hoefer, ta]). That is, both future and past events can have an objective probability that falls strictly between 0 and 1. One way to motivate the idea that the past as well as the future can be chancy, is to assume that objective probabilities – or *chances*, as I will from now on call them – do not supervene on facts about *particular* events but on facts about event-*types* (as I briefly mentioned in last chapter). Suppose E is a sentence expressing the occurrence of an event e. Then the chance of e and the chance of E's truth, on this version of Humean Supervenience, does not supervene on facts about this particular event e, but on facts about events *like* e. Hence, it might well happen that although e did not take place, so E turns out to be false, the chance of E's truth is still very high.

When combined with Skyrms' thesis, the above version of Humean Supervenience has immediate implications for the relation between non-modal facts and subjunctive conditionals. In particular, as we saw in last chapter, these two views are together incompatible with both Strong and Weak Centring. On the face of it, it might seem impossible to find a single semantical system for conditionals that implies the above views for both indicative and subjunctive conditionals, while being at the same time consistent with Humean Supervenience. Firstly, many philosophers are of the opinion that David Lewis' famous *triviality results* (proved in [Lewis, 1976] and [Lewis, 1986b]) imply that indicative conditionals cannot both supervene on non-modal facts and satisfy Adams' thesis (see for instance [Bradley, 2012]: 557). Thus, *any* Humean Supervenience view might seem incompatible with the version of the Ramsey view discussed above. Secondly, we have seen that Adams' thesis implies Centring. Hence, it might seem impossible for the same semantical system to imply both Adams' and Skyrms' thesis, while being consistent with the version of the Humean Supervenience view discussed above.

Below I show that Richard Bradley's recent *Multidimensional Possible World Semantics* (MD-semantics) for conditionals [Bradley, 2012] can simultaneously satisfy the above views. In other words, I will show that contrary to what e.g. Bradley himself suggests, we can have a semantics of conditionals that implies both Adams' and Skyrms' theses (without

[1] The view is named after David Hume, "the greater denier of necessary connections" ([Lewis, 1987]: ix).

[2] Or, on Lewis' view, two worlds *like ours* (see e.g. [Lewis, 1987]). For the present purposes, I will treat supervenience theses as being either necessarily true or necessarily false.

[3] Various empirical and metaphysical objections have been raised against Humean Supervenience. I will not try to defend the view against these objections, but will rather focus on trying to make it compatible with the aforementioned view on the probability of conditionals.

encountering triviality problems) while satisfying Humean Supervenience. I will do so by showing that contrary to what Bradley's own formulation suggests, we can derive the so-called *Ramsey test* from the MD-semantics without first adding Centring. Having derived the Ramsey test, we can, given an additional assumption, derive Adams' thesis for indicatives and Skyrms' thesis for subjunctives. This entails that Centring must hold for indicatives, but will imply the failure of Centring for subjunctives given the above interpretation of chance. Finally, I show that we can add a supervenience constraint on the MD-semantics, such that conditionals supervene on non-modal facts, without having to worry about triviality (again, contrary to what Bradley suggests).

3.3 The Multidimensional Semantics

Bradley's MD-semantics is based on the observation that a person's uncertainty as to the truth of sentences is of (at least) two kinds: uncertainty about what is the case and uncertainty about what would be the case under a supposition. In the language of possible worlds, people are both uncertain as to which world is actual and uncertain as to which worlds are counter-actual[4] under different suppositions. The basic ingredients in the MD-semantics are n-tuples of worlds $\langle w_i, w_{Aj}, w_{Bk}, w_{Cl}, ...\rangle$, where w_i is a candidate for the actual world, w_{Aj} a *counter-actual* world under the supposition that A (where A is a factual sentence), w_{Bk} a counteractual world under the supposition that B, etc. Thus we can interpret each n-tuple as the *event* that w_i is the actual world, w_{Aj} a counteractual world under the supposition that A, etc. I will for now focus on a single supposition, a supposition that A, but the results I state for the framework hold for multiple suppositions as well (see [Bradley, 2012]: sec. 7). And I will use an ordered pair $\langle w_i, w_j\rangle$ to represent the event that the first element, w_i, is the actual world and the second element, w_j, the counteractual world under the supposition that A.

To keep things simple, let us for the moment assume that there are only four possible worlds to consider: w_1 to w_4. Let us moreover assume that sentence A is true at worlds w_1 and w_2 but sentence B at worlds w_1 and w_3. The semantic content, or meaning, of the factual sentence A is then the set of pairs (constructed from these four worlds) whose first element is either w_1 or w_2; and the semantic content of the conditional sentence $A \rightarrow B$ is the set of pairs whose second member is w_1. Putting this information in tabular form makes it much easier to grasp. The first table represents how we have matched up worlds and sentences:

	B	$\neg B$
A	w_1	w_2
$\neg A$	w_3	w_4

The next table represents the semantic content of sentences. The columns of the table give the content of conditional sentences; the first column the content of the sentence $A \rightarrow B$, the second the content of the sentence $A \rightarrow \neg B$. Thus, the first column gives the *truth conditions* of $A \rightarrow B$; any of the pairs in this column make the conditional true. The rows represent the semantic content of factual sentences; the first and second row, for instance, the content of sentence A but the first and third the content of sentence B. Thus the first two rows give the truth conditions of A; any of the pairs in these rows make this factual sentence true.

[4]This terminology is borrowed from [Bradley, 2012].

	W_1	W_2
W_1	$\langle w_1, w_1 \rangle$	$\langle w_1, w_2 \rangle$
W_2	$\langle w_2, w_1 \rangle$	$\langle w_2, w_2 \rangle$
W_3	$\langle w_3, w_1 \rangle$	$\langle w_3, w_2 \rangle$
W_4	$\langle w_4, w_1 \rangle$	$\langle w_4, w_2 \rangle$

The content of a sentence, on my understanding, is a proposition. Let α represent the proposition expressed by the sentence A and β the proposition expressed by B. Then the proposition α is just the set of pairs of worlds in the first and second row and the proposition $\alpha \rightarrow \beta$ the set of pairs of worlds in the first column.

As the above table indicates, the MD-semantics entails the truth of the *Conditional Excluded Middle* (CEM); that is, the principle that for all sentences A and B, either $A \rightarrow B$ or $A \rightarrow \neg B$ is true. For given the ordinary excluded middle, either B or $\neg B$ is true at each counteractual A-world. But the semantics also respects the intuition, discussed in last chapter, that for some sentences B, the facts of the actual world might determine the truth of neither $A \rightarrow B$ nor $A \rightarrow \neg B$. (For instance, the facts of the actual world might not determine whether Bizet would be Italian or not, had he and Verdi been compatriots.) In general it is not necessarily the case, on the MD-semantics, that the (non-modal) facts of a world determine what conditionals are true (as further discussed in section 3.7). Nevertheless, it must be true that whatever world is counteractual under the supposition that A, either B or $\neg B$ is true at that world.[5]

3.4 The Ramsey Test without Centring

Many authors on the logic and semantics of conditionals have taken some version of the *Ramsey test thesis* (RT) as one of the desiderata that any theory of conditionals must satisfy. In this section I will focus on what I called the *strong* Ramsey test, which, recall, says that for any (propositions or sentences) A, B, and any rational credence function P:

$$P(A \rightarrow B) = P_A(B) \tag{3.1}$$

In other words, our degree of belief in the conditional $A \rightarrow B$ should be identical to our degree of belief in B upon the supposition that A.

In Bradley's own presentation of the MD-semantics, he derives the Strong Ramsey test after having added *Centring* to his semantics (see [Bradley, 2012]: sec. 5). Recall that Strong and Weak Centring are the respectively following two principles: $\emptyset \vdash (A \wedge B) \supset (A \rightarrow B)$, $\emptyset \vdash (A \rightarrow B) \supset (A \supset B)$. Stated probabilistically, the two principles respectively say that for any A, B: $P(A \wedge B) \leq P(A \rightarrow B)$, $P(A \rightarrow B) \leq P(A \supset B)$.

On the MD-semantics, and given our toy-model from above, adding the semantical Centring condition eliminates the possibility that w_1 is the actual world and w_2 counteractual under the supposition that A (i.e. $\langle w_1, w_2 \rangle$). For A is true at w_1, which with Centring means that if w_1 is the actual world, then the counteractual world under the supposition that A must also be w_1. By the same reasoning, Centring eliminates (from our toy-model) the possibility

[5] Below I suggest certain generalisations of the MD-semantics. Wlodek Rabinowicz has pointed out to me that one could generalise the semantics even further, in a way that would not make the CEM come out true: instead of assuming that given any elementary possibility, each supposition corresponds to a single world, one could assume that each supposition corresponds to a set of worlds. This is certainly something that I intend to explore further.

$\langle w_2, w_1 \rangle$. Thus when Centring is added, the table above which represented the semantic content of sentences becomes:

	W_1	W_2
W_1	$\langle w_1, w_1 \rangle$	
W_2		$\langle w_2, w_2 \rangle$
W_3	$\langle w_3, w_1 \rangle$	$\langle w_3, w_2 \rangle$
W_4	$\langle w_4, w_1 \rangle$	$\langle w_4, w_2 \rangle$

Below I show that contrary to what Bradley's original presentation of the MD-semantics suggested, the Strong Ramsey test follows from the MD-semantics *with or without* Centring. The importance to Humeans of being able to derive the Ramsey test without assuming Centring will be made clear in next section. But it might be worth noting again that even those who are not motivated by a Humean metaphysics might find it desirable to be able to derive RT without assuming Centring. Many authors on the logic of conditionals have pointed out that the strong Centring condition seems particularly counterintuitive. Are we for instance willing to infer from the fact that David Cameron was Prime Minister of the UK in 2013 and England made it to the 2014 World Cup, that "had Cameron been the UK Prime Minister in 2013, England would have made it to the 2014 World Cup"? As I said in last chapter, I must assume that the answer to this question is 'yes'. However, I do admit that that might be counterintuitive.

Before deriving the Ramsey test, we need some additional assumptions and notation. Let $\mathcal{P}$ be a probability mass function on the set W of worlds that measures the probability of any world being the actual one; $\mathcal{Q}$ a probability mass function on the subset of W where A is true (which I will call W_A) that measures the probability of any world being counteractual under the supposition that A; and $\mathcal{P}r$ a joint probability mass function on pairs of worlds, where e.g. $\mathcal{P}r(\langle w_i, w_j \rangle)$ measures the probability that w_i is the actual world and w_j the counteractual world under the supposition that A.

Bradley assumes the following relationship between $\mathcal{P}r$ and $\mathcal{P}$ and $\mathcal{Q}$:

Assumption 2 (Marginalisation)**.**

1. $\sum_{w_i \in W} \mathcal{P}r(\langle w_i, w_j \rangle) = \mathcal{Q}(w_j)$
2. $\sum_{w_j \in W_A} \mathcal{P}r(\langle w_i, w_j \rangle) = \mathcal{P}(w_i)$

Now let $\mathcal{P}r(\cdot \mid w_i)$ be the conditional probability mass function on W_A, given that w_i is the actual world. In other words, the function measures, for any w_j, the conditional probability that w_j is counteractual under the supposition that A, given that w_i is actual. From the two marginalisation properties it follows that:

$$\mathcal{P}r(\langle w_i, w_j \rangle) = \mathcal{P}(w_i).\mathcal{P}r(w_j \mid w_i) \tag{3.2}$$

Properties like these are standard in probability theory and statistics. I will thus not discuss their justification. Given Marginalisation, the uncertainty in our toy-model above can be represented by:[6]

[6]Since it follows from centring, like the discussion above illustrates, that $\mathcal{P}r(w_1 \mid w_1) = 1$ and $\mathcal{P}r(w_2 \mid w_1) = 0$, an agent's state of uncertainty given Centring can be represented as:

	W_1	W_2
W_1	$\mathcal{P}(w_1).\mathcal{P}r(w_1 \mid w_1)$	$\mathcal{P}(w_1).\mathcal{P}r(w_2 \mid w_1)$
W_2	$\mathcal{P}(w_2).\mathcal{P}r(w_1 \mid w_2)$	$\mathcal{P}(w_2).\mathcal{P}r(w_2 \mid w_3)$
W_3	$\mathcal{P}(w_3).\mathcal{P}r(w_1 \mid w_3)$	$\mathcal{P}(w_3).\mathcal{P}r(w_2 \mid w_3)$
W_4	$\mathcal{P}(w_4).\mathcal{P}r(w_1 \mid w_4)$	$\mathcal{P}(w_4).\mathcal{P}r(w_2 \mid w_4)$

We can now show that the Ramsey test follows from the MD-semantics with the above assumption. (Notice that Centring is nowhere assumed in the following argument.) Given the MD-semantics, and how we have paired up worlds and sentences, the truth conditions of the sentence $A \rightarrow B$ are given in the first column of the table. From the second marginalisation property and the marginalisation implication it follows that the total probability of this column is $\mathcal{Q}(w_1)$, i.e. the probability of w_1 being the counteractual world under the supposition that A. So $P(A \rightarrow B) = \mathcal{Q}(w_1)$. But w_1 is the only A-world where B is true. Hence, $\mathcal{Q}(w_1)$ simply measures the probability of B under the supposition that A. So the probability of $A \rightarrow B$ is the probability of B under the supposition that A, in accordance with the Quantitative RT.

The above result holds in general. In the simple example we have been working with, there is only one A-world where B holds. Dropping that assumption does not undermine the result.

Theorem 12. *The MD-semantics entails the Quantitative RT.*

Proof. Assume that there are multiple A-worlds where B holds, call them w_i, w_j, ... w_m. Then the truth condition of $A \rightarrow B$ is given by the columns W_i, W_j, ... W_m. Each of these columns has probability $\mathcal{Q}(w_i)$, $\mathcal{Q}(w_j)$, etc. Thus, we have $P(A \rightarrow B) = \mathcal{Q}(w_i) + \mathcal{Q}(w_j) + ...\mathcal{Q}(w_m) = \mathcal{Q}(w_i \vee w_j \vee ...w_m)$ (since worlds are mutually exclusive). But $\mathcal{Q}(w_i \vee w_j \vee ...w_m)$ just measures the probability of B under the supposition that A. Hence, $P(A \rightarrow B) = P_A(B)$. □

3.5 Adams' Thesis, Skyrms' Thesis and Centring

Recall that there are different ways one can suppose something to be true. In particular, there is a difference between supposing that, *as a matter of fact*, A is true; versus supposing that, *contrary to fact*, A is true (see e.g. [Joyce, 1999]: ch. 6). The probability of B under the supposition that A often varies significantly depending on the mode of supposition. When we suppose that, as a matter of fact, Oswald did not kill Kennedy, most of us (hypothetically) conclude that someone else killed Kennedy. For we are pretty certain that Kennedy was in fact murdered. Similarly, we attach high probability to the indicative conditional: "If Oswald did not kill Kennedy, then someone else did". However, when we suppose that, contrary to fact, Oswald did not kill Kennedy, most of us (hypothetically) conclude that nobody killed Kennedy. For most of us (presumably) do not believe that there was a conspiracy to murder Kennedy. Similarly, we attach low probability to the counterfactual: "If Oswald had not killed Kennedy, someone else would have."

	W_1	W_2
W_1	$\mathcal{P}(w_1)$	0
W_2	0	$\mathcal{P}(w_2)$
W_3	$\mathcal{P}(w_3).\mathcal{P}r(w_1 \mid w_3)$	$\mathcal{P}(w_3).\mathcal{P}r(w_2 \mid w_3)$
W_4	$\mathcal{P}(w_4).\mathcal{P}r(w_1 \mid w_4)$	$\mathcal{P}(w_4).\mathcal{P}r(w_2 \mid w_4)$

One of the appeals of the Ramsey test, as mentioned in last chapter, is that since it is general enough to allow for different types of suppositions, it can be interpreted such that it becomes appropriate for either indicative or subjunctive conditionals. Recall that I am assuming that with indicative conditionals the appropriate supposition is matter-of-factual – or, as it is often called, *evidential* – and that such suppositions are characterised by Bayesian conditioning. Adding this assumption to the strong Ramsey test gives us Adams' thesis, i.e. the view that for any A, B and any rational credence function P: $P(A \mapsto B) = P(B \mid A)$.

Bradley shows that we can derive Adams' thesis from the MD-semantics, when we add Centring to the semantics and two conditions that are appropriate for evidential suppositions ([Bradley, 2012]: 562-563). Implying AT is arguably the main advantage of the MD-semantics, since it shows that contrary to a popular view, Lewis' famous triviality results do not show that it is impossible to give truth conditions for conditionals that satisfy AT in a non-trivial language. Bradley himself suggests that his semantics can imply AT while avoiding the triviality results *because* the semantics, as he formulates it, violates the thesis of Humean Supervenience ([Bradley, 2012]: 557): what counterfacts are true, on the MD-semantics as he formulates it, does not supervene on the actual (or non-modal) facts. However, as I show in next section, even if we add a Humean Supervenience constraint onto the MD-semantics, it sill implies AT without running into triviality problems. So the violation of Humean Supervenience cannot be the reason the MD-semantics can validate AT in a non-trivial language. Rather, it seems the reason simply is that even if we assume that counterfacts supervene on on the facts, the counterfacts are not, in this semantics, determined by the facts *in the particular way* that is commonly assumed in the triviality results (see e.g. overview in [Hájek and Hall, 1994]). In particular, the truth of B does not determine that $A \mapsto B$ is true, as these results assume.

If we first add what I have called the "Principal Suppositional Principle" (PSP) to the MD-semantics, then we can (as we saw in last chapter) derive AT from the semantics without first adding Centring. For the present purposes, we need to formalise the PSP as follows:

Assumption 3 (PSP). $\boldsymbol{Q}(B \mid W_i) = Ch_{w_i}(B \mid A)$

Generalised to the case where there is uncertainty about the chances, this version of PSP says:

$$\boldsymbol{Q}(B) = \sum_i Ch_{w_i}(B \mid A).\boldsymbol{Q}(W_i) \tag{3.3}$$

Recall from last chapter that to get AT from equation 3.3 and the Ramsey test, we assume that when the supposition is evidential, $\boldsymbol{Q}(W_i)$ equals $P(W_i \mid A)$. We also saw in last chapter that from an equation like 3.3 and RT we can derive Skyrms' thesis, if we assume that when the supposition is contrary-to-factual, $\boldsymbol{Q}(W_i)$ equals $P(W_i)$.

So we have derived Adams's thesis for indicative conditionals and Skyrms' thesis for subjunctive conditionals from the MD-semantics with the addition of the PSP. And we have done so without adding Centring as a semantic constraint. Nevertheless, it is clear that Centring must hold for indicatives in this system, given that Adams' thesis holds for such conditionals, since Adam's thesis implies both Strong Centring and Weak Centring, as I proved in last chapter. The same is not true for Skyrms' thesis. Without any particular interpretation of chance, ST is compatible Centring being either true or false.

However, as we saw in last chapter, the view that chances supervene on facts about event-types, which implies that the past can sometimes be chancy, means that Centring is

inconsistent with Skyrms' thesis. This should convince those who accept this particular Humean view of the importance of being able to derive Skyrms' thesis from the MD-semantics without adding first a semantic Centring condition. For this shows that those who accept the above view on chance and also believe that subjunctive conditionals express objective-chance dependencies, can view the MD-semantics as providing a semantical underpinnings for their philosophical views. More generally, it is useful to know that we can accept this intuitively plausible notion of chance without giving up the hope of a unified theory of conditionals that implies AT for indicatives and ST for subjunctives.

3.6 Avoiding triviality

We have seen that the MD-semantics implies Adams' thesis while giving truth conditions for indicative conditionals as well as subjunctives. So it must somehow avoid Lewis' triviality result. Considering our possible world toy-model above, we can see why Lewis' result is not sound given this semantics: we cannot assume (as Lewis does) that $P(A \mapsto B \mid B) = 1$ (this is what Hajek and Hall call Lewis' "key maneuver" [Hájek and Hall, 1994]: 89). So contrary to what is standardly assumed, the truth of B does not determine, according to this semantics, that $A \mapsto B$ is true.

Let's see more precisely how the model I have been working with violates Lewis' key maneuver. With the assignments of worlds to sentences I used above, we have $B = \{< w_3, w_1 >\} \bigcup \{< w_3, w_2 >\} \bigcup \{< w_1, w_1 >\} \bigcup \{< w_1, w_2 >\}$. But the conditional $A \mapsto B$ is neither true in $< w_3, w_2 >$ nor $< w_1, w_2 >$, so as long as these have a positive probability, then $P(A \mapsto B \mid B) \neq 1$. Now of course, Centring, which I have suggested we must accept, implies that $P(< w_1, w_2 >) = 0$ (given our example). But we have no justification for assuming that $P(< w_3, w_2 >) = 0$. Hence, one of the assumptions of Lewis' triviality result is not satisfied in this model.

It follows from the axioms of probability that $P(p \mid p \wedge q) = 1$.[7] Hence, in the MD-semantics, the equality between $P(A \mapsto B \mid B)$ and $P(B \mid A \wedge B)$ must not hold, in general, in spite of the equality between $P(A \mapsto B)$ and $P(B \mid A)$. Given the last equality, $P(A \mapsto B \mid B)$ is identical to $P(B \mapsto (A \mapsto B))$. So the lesson to be learnt is that while the MD-semantics implies Adams' thesis for simple conditionals, it does not do so for conditionals with conditional consequents.

To see that Leitgeb's triviality result is unsound for MD-semantics, recall that he assumes that any counterfactual $A \boxdot\!\!\rightarrow B$ is either true or false at w_i; and that if the counterfactual is true at w_i, then $P(A \boxdot\!\!\rightarrow B \wedge w_i) = P(w_i)$. Now given MD-semantics, a counterfactual is not simply true or false at a world; it depends also on what are the relevant counter-worlds. Moreover, even if, say, w_i happens to be the actual world and $A \boxdot\!\!\rightarrow B$ is actually true, it does not follow, on the MD-semantics, that $P(A \boxdot\!\!\rightarrow B \wedge w_i) = P(w_i)$. For instance, say that $< w_3, w_1 >$ is actually true. Given our toy-model, $A \boxdot\!\!\rightarrow B$ is then a true counterfactual. But our model also has the possibility $< w_3, w_2 >$, and w_2 is an A-world where B is false. So $P(w_3)$ (which I understand as the probability of w_3 being the actual world) is not identical to $P(A \boxdot\!\!\rightarrow B \wedge w_3)$ even though the counterfactual happens to be true at w_3.

However, Leitgeb's proof can be made sound given MD-semantics if, instead of talking of a counterfactual being either true or false at w_i, we assume in the proof that a counterfactual is either true or false at any pair $< w_i, w_j >$. But then we no longer have a *triviality* result. For

[7] $P(p \mid p \wedge q) = \dfrac{P(p \wedge p \wedge q)}{P(p \wedge q)} = \dfrac{P(p \wedge q)}{P(p \wedge q)} = 1.$

the triviality consisted in conditional chances *at a world* being either 0 or 1. However, since Leitgeb's proof, to be sound given MD-semantics, can only assume that a counterfactual is either true or false given any pair consisting of a world and a counter-world, but not true or false given simply any world, the result is that at any particular *world and counter-world pair*, the conditional chance of B given A must be either 0 or 1. But that is as it should be, and does not trivialise conditional chance: given any pair consisting of an actual world and a counterfactual A-world, the conditional chance of B given A should be either 0 or 1. For once we have fixed the attention on a particular counteractual A-world, B is either true or false at that world. But this does not mean that at any actual actual world w_i, the conditional chance of B given A must be either of these extremes, since w_i may take a number of counteractual A-worlds, and in some of these B may be true but in others B may be false.

3.7 Facts, counterfacts and Supervenience

The strongest metaphysical assumption of the MD-semantics, as originally presented by Bradley, is that counterfacts do not supervene on non-modal facts. (Nor does the converse hold.) To be precise, here is how I will be using the term 'supervenience':

Definition 2 (Fact/Counter-fact Supervenience)**.** *Counter-facts supervene on non-modal facts just in case no two worlds can differ in counterfacts without also differing in non-modal facts, but non-modal facts supervene on counterfacts just in case no two worlds can differ in non-modal facts without also differing in counter-facts.*

In our toy-model above, for instance, even after adding centring, we have both $\langle w_3, w_1\rangle$ and $\langle w_3, w_2\rangle$ as possibilities. The first element of these two pairs is the same, and thus these two possibilities share all non-modal facts. Nevertheless, differ in what conditionals they make true. For instance, in the first case, $A \rightarrow B$ is true, but in the second case, $A \rightarrow \neg B$ is true, despite no difference in non-modal facts. Thus, conditionals do not supervene on (non-modal) facts – and hence, given Skyrms' thesis, chances do not supervene on (non-modal) facts either[8] – contrary to the intuitively plausible Humean Supervenience view.

As I previously discussed, it is the failure of the facts of the actual world to determine the counterfacts – and hence the failure of the counterfacts to supervene on the facts – that that explains how the Conditional Excluded Middle can hold even thought the facts of the world might neither determine the truth of $A \rightarrow B$ nor $A \rightarrow \neg B$. Bradley, moreover, seems to imply that it is this failure of supervenience that makes the accommodation of Adams' thesis possible within a truth-conditional semantics for non-trivial languages ([Bradley, 2012]: 557). However, as I will now show, it is possible to make counterfacts supervene on (non-modal) facts in the MD-semantics without encountering triviality. Unfortunately, then we no longer have an explanation for the aforementioned intuition about CEM. But that is perhaps a price that Humeans are willing to pay.

To show that we can add a supervenience constraint on the MD-semantics without having to worry about triviality, I will focus on showing that Lewis' "key manoeuver" ([Hájek and Hall, 1994]: 89) does still not hold. The key manoeuver, recall, is to show that Adams' thesis implies that for all A, B, such that $P(A \wedge B) > 0$, $P(A \mapsto B \mid B) = P(B \mid A \wedge B) = 1$. This is not an assumption that we can, in general, make in the MD-semantics, as we have

[8] ST implies that two worlds cannot differ in counterfactuals without differing in chances; in other words, counterfactual conditionals supervene on chances. But then if chances supervene on facts, counterfactuals supervene on facts (since supervenience is a *transitive* relation). So if conditionals in general do not supervene on facts, then, given Skyrms' thesis, chances cannot supervene on facts either.

seen. Now let us add to the model the assumption that each world is only compatible with *one* possible counteractual A-world (but the converse does not hold). In other words, we assume that counterfacts supervene on (non-modal) facts. Suppose for instance that w_4 is only compatible with w_1 as a counteractual A world and w_3 is only compatible with w_2 as a counteractual A world. We can represent this by the following table:

	W_1	W_2
W_1	$\langle w_1, w_1 \rangle$	
W_2		$\langle w_2, w_2 \rangle$
W_3		$\langle w_3, w_2 \rangle$
W_4	$\langle w_4, w_1 \rangle$	

From the above table we should see that we can still not assume $P(A \rightarrow B \mid B) = 1$. For given our assignment of sentences to worlds (which was perfectly legitimate, given this semantics), we now have $B = \{\langle w_1, w_1 \rangle\} \bigcup \{\langle w_3, w_2 \rangle\}$. And recall that $A \rightarrow B$ is not true in $\langle w_3, w_2 \rangle$ (since B is not true in the A-world w_2). Hence, $P(A \rightarrow B \mid B) \neq 1$. But all ingredients needed to derive the Ramsey test (and Adams' thesis) are still in place. So after adding this supervenience relation, we still have a semantical model for indicative conditionals where AT holds for a non-trivial language (contrary to what Bradley suggests).

In fact, something even stronger holds: if counterfacts supervene on facts *and facts supervene on counter-facts*, then the MD-semantics does still not lead to triviality.[9] To establish this stronger claim, I will enrich our toy-model slightly, but use only four basic sentences: A, B, $\neg A$ and $\neg B$. The n-tuple $\langle w_1, w_1, w_1, w_3, w_2 \rangle$ now represents the possibility that w_1 is the actual world, w_1 counteractual under the supposition that A, w_1 also counteractual under the supposition that B, w_3 counteractual under the supposition that $\neg A$ and w_2 counteractual under the supposition that $\neg B$. Then for each candidate for actuality, the following table might represent counter-worlds under different suppositions:

	$\langle W, W_A, W_B, W_{\neg A}, W_{\neg B} \rangle$
W_1	$\langle w_1, w_1, w_1, w_3, w_2 \rangle$
W_2	$\langle w_2, w_2, w_1, w_3, w_2 \rangle$
W_3	$\langle w_3, w_2, w_3, w_3, w_2 \rangle$
W_4	$\langle w_4, w_1, w_1, w_4, w_4 \rangle$

The above possibility space satisfies the constraint that counterfacts supervene on non-modal facts – no two possibilities differ in counter-worlds without differing in actual worlds – and the constraint that non-modal facts supervene on counterfacts – no possibilities differ in actual worlds without differing in some counter-worlds. In addition, it satisfies Centring. But here we should see that we still cannot assume that $P(A \rightarrow B \mid B) = 1$. For instance, if world w_2 is actual, and thus $\langle w_2, w_2, w_1, w_3, w_2 \rangle$ is the n-tuple that describes the actual and counteractual worlds under different suppositions, then B is actually true but nevertheless $A \rightarrow B$ is false. Hence, we certainly cannot assume that in general, $P(A \rightarrow B \mid B) = 1$. But everything needed to derive AT is still in place.

[9] Given my proof in last chapter that Strong and Weak Centring together imply supervenience in the latter direction, the possibility of strengthening the result from last paragraph in this way should not come as a surprise, given that the weaker result holds even with SC and WC.

Should we then, when using the MD-semantics to represent conditionals and rational attitudes to them, in general work with models containing either or both supervenience constraints discussed above? I think not. Firstly, such constraints only make sense if we are working with models where the worlds are sufficiently fine-grained. Even Humeans should accept that a description of the actual world only implies what conditionals are true if the description contains *all* (non-modal) facts of the world. But even the most rational agents are of course unaware of some facts of their worlds. Therefore, when constructing possible world models to represent rational attitudes to conditionals, no harm is done if the worlds are not specific enough to sustain the supervenience relations that we might think hold between facts and counter-facts. And indeed such coarse-grained models are much more realistic, even if the aim is to model the attitudes of perfectly rational agents. Secondly, I am not convinced that we should think that e.g. conditionals do supervene on non-modal facts. Indeed, I will suggest in chapter 5 that the desirability of counterfactuals can in some cases not be determined by the desirability of any factual propositions alone, which casts doubt on the Humean view that counterfactuals (and perhaps other modal propositions) supervene on the non-modal facts. Finally, if we add a (modal-on-non-modal) supervenience constraint, the semantics no longer captures the intuition that for for some A and B, the facts of our world don't determine whether $A \boxdot\!\!\rightarrow B$ or $A \boxdot\!\!\rightarrow \neg B$. Nevertheless, it is good to know that we *can* use these more fine grained models, if we like, without having to worry about triviality.

3.8 Concluding remarks

In next chapter I will use the multidimensional possible world semantics for conditionals to introduce counterfactuals to Richard Jeffrey's decision theory, in a way that allows for desirability measures of counterfactuals propositions, as well as conjunctions of counterfactual and factual propositions. I hope to have explained why the multidimensional possible world semantics provides a plausible theory of conditionals. However, the main justification for focusing on that particular theory of conditionals in this PhD thesis, is that it allows for an elegant way to introduce counterfactuals to Jeffrey's theory.

Chapter 4

Counterfactual Desirability

4.1 Introduction

As I mentioned in the introductory chapter, the main motivation behind this PhD thesis is the observation that the desirability of what actually occurs is often influenced by what *could have been*. Recall the two examples I discussed in the introduction. Suppose you have been offered two jobs, one very exciting but with a substantial risk of unemployment, the other less exciting but more secure. If you choose the more risky option, and as a result become unemployed, you might find that the fact that you *could have* chosen the risk-free alternative makes being unemployed even worse. For in addition to experiencing the normal pains of being out of job, you may be filled with *regret* for not having chosen the risk-free alternative. Not all occasions where what could have been influences the desirability of what actually occurs involve anything like regret. Suppose that a patient has died because a hospital gave the single kidney that it had available to another patient. Suppose also that the two patients were in equal need of the kidney, had equal rights to treatment, etc. Now if we learn that a fair lottery was used to determine which patient was to receive the kidney, then most of us find that this makes the situation less undesirable than had the kidney simply been given to one of them. For that at least means that the patient who died for lack of a kidney had a chance to acquire it. In other words, *had* some random event turned out differently than it actually did, the dead patient *would have* lived. And that somehow makes the situation less undesirable.

This desirabilistic dependency between what is and what could have been creates well known paradoxes for the traditional theory of rational choice, as for instance formulated by John von Neumann and Oskar Morgenstern [von Neumann and Morgenstern, 1944] and Leonard Savage [Savage, 1972]. The first example is just a simplified version of Maurice Allais' infamous paradox [Allais, 1953], [Allais, 1979], whereas the latter is an instance of a decision theoretic problem identified decades ago by Peter Diamond [Diamond, 1967]. In this chapter, which is based on a joint working paper with Richard Bradley [Bradley and Stefánsson, 2015], I use a framework based on a combination of Richard Jeffrey's decision theory [Jeffrey, 1983] and Bradley's semantics for conditionals [Bradley, 2012] to explore the above dependency.

Section 4.2 explains the two paradoxes and why they cast doubt on a rationality postulate known as *Separability*. Unlike expected utility theory, Richard Jeffrey's decision theory does not encode this Separability property. But as is explained in section 4.3, the lack of counterfactual prospects in Jeffrey's theory nevertheless means that it cannot represent the preferences discussed by Allais and Diamond. To overcome this problem, section 4.4

introduces counterfactuals into Jeffrey's theory and section 4.5 then establishes that this makes it possible to represent Allais' and Diamond's preferences as maximising the value of a Jeffrey desirability function. In section 4.6 I explain what axioms must be added to the multidimensional framework to obtain a standard expected utility representation.

There are three important results in this chapter that I would like to emphasise. Firstly, as already mentioned, the chapter establishes that we can represent Diamond's and Allais' preferences as maximising Jeffrey-desirability if we extend Jeffrey's framework to counterfactuals in the way I suggest. Secondly, I show that contrary to what is standardly claimed, the aforementioned preferences violate *two*, not one, axioms that are implied by standard expected utility theory. Neither axiom is required for Jeffrey-desirability representation. Finally, I explain why the aforementioned axioms imply certain very implausible *epistemic* norms for counterfactual reasoning. So in addition to being inconsistent with preferences that, I contend, are practically rational, the two axioms are inconsistent with epistemically rational counterfactual reasoning. I believe that these results seriously undermine the claim that expected utility theory is our best theory of practical rationality.

I omit proofs of some of the claims in this chapter. These proofs can however be found in a joint working paper with Richard Bradley [Bradley and Stefánsson, 2015] (which can be found online here: http://www.lse.ac.uk/ CPNSS/research/ currentResearchProjects/ ChoiceGroupworkingPapers.aspx).

4.2 Two Paradoxes of Rational Choice

The Allais Paradox has generated a great deal of discussion, both amongst philosophers and behavioural economists and psychologists. The paradox is generated by offering people a pair of choices between different lotteries, each of which consists in tickets being randomly drawn. First, people are offered a choice between a lottery that is *certain* to result in the decision maker receiving a particular prize, say £2400, and a lottery that could result in the decision maker receiving nothing, but could also result in the decision maker receiving either as much as or more than £2400. The situation can be represented as a choice between the lotteries L_1 and L_2 below, where, for instance, L_1 results in the decision maker receiving a prize of £2500 if one of the tickets number 2 to 34 is drawn:

	1	2 – 34	35 – 100
L_1	£0	£2500	£2400
L_2	£2400	£2400	£2400

Having made a choice between L_1 and L_2, people are asked to make a choice between lotteries L_3 and L_4:

	1	2 – 34	35 – 100
L_3	£0	£2500	£0
L_4	£2400	£2400	£0

When presented with this pair of choices, many people choose, and *strictly* prefer, L_2 over L_1 and L_3 over L_4. (See [Kahneman and Tversky, 1979] for discussion of an early experiment of the Allais Paradox.) One common way to rationalise this preference, which I will refer to as 'Allais' preference', is that when choosing between L_1 and L_2, the possibility of ending up with nothing when you could have received £2400 for sure outweighs the possible extra

gain of choosing the riskier alternative, since receiving nothing when you could have gotten £2400 for sure is bound to cause considerable *regret* (see e.g. [Loomes and Sugden, 1982] and [Broome, 1991]). When it comes to choosing between L_3 and L_4, however, the desire to avoid regret does not play as strong role, since decision makers reason that if they choose L_3 and end up with nothing then they would, in all likelihood, have received nothing even if they had chosen the less risky option L_4.

Intuitively rational as it seems, Allais' preference is inconsistent with the most common formal theories of rational choice. (Assuming, that is, that the probabilities of each ticket is the same in the two choice situations. That is, the probability of a ticket being drawn from e.g. tickets 2-34 is the same in both choice situations.) Let us continue, throughout this section, to think of the alternatives between which people have preferences as *lotteries*, with the understanding that some lotteries result in the same consequence (or 'prize') in all states of the world. According to the standard theory of rational choice, *Expected Utility theory* (EU theory), all rational preferences can be represented as maximising the expectation of utility, where the expected utility of a lottery L is given by:

$$EU(L) = \sum_{s_i \epsilon \mathbf{S}} u(L(s_i)).P(s_i)$$

where $\mathbf{S}$ is a partition of the possible states, $L(s_i)$ is the consequence of L if s_i happens to be the actual state of the world, u a utility measure on consequences, and P a probability measure on states.

In the usual manner let $\geq$ represent the relation '... is at least as preferred as ...', and $>$ and $\sim$ the corresponding strict preference and indifference relations. (When more convenient, I will use $<$ for the same relation, except of course in the other direction, i.e. representing '... is less preferred than...'.) Then EU theory states that for any rational agent:

$$L_i > L_j \quad \text{iff} \quad EU(L_i) > EU(L_j) \tag{4.1}$$

(When the above holds for a person's preferences, we say that the *EU* function *represents* the person's preferences.)

The problem that the Allais Paradox poses to standard decision theory, is that there is no way to represent Allais' preference as maximising the value of a function with the EU form. To see this, let us assume that in both choice situations the decision maker considers the probability of each ticket being drawn to be 1/100. Then if Allais' evaluation of the alternatives is in accordance with the EU equation, Allais' preferences implies that both:

$$U(£0) + 33U(£2500) + 66U(£2400) < 100U(£2400) \tag{4.2}$$

and:

$$66U(£0) + 34U(£2400) < 67U(£0) + 33U(£2500) \tag{4.3}$$

But the latter implies that:

$$\begin{aligned} & 34U(£2400) + 66U(£2400) = 100U(£2400) \\ < \quad & U(£0) + 33U(£2500) + 66U(£2400) \end{aligned}$$

in contradiction with 4.2. Hence, there is no EU function that simultaneously satisfies $U(L_1) < U(L_2)$ and $U(L_4) < U(L_3)$. In other words, there is no way to represent a person

who (strictly) prefers L_2 over L_1 and L_3 over L_4 as maximising utility as measured by an EU function. Since all rational preferences should, according to EU theory, be representable as maximising expected utility, this suggests that either Allais' preference is irrational or EU theory is incorrect. (Hence the 'paradox': many people both want to say that Allais' preference is rational and that EU theory is the correct theory of practical rationality.)

Another way to see that Allais' preference cannot be represented as maximising the value of an EU function, is to notice that (given the suggested and perhaps most natural description of the lotteries) the preference violates a *Separability axiom* which a preference needs to satisfy for it to be possible to represent it by an EU function. (This Separability axiom is perhaps best known in the form of Savage's *Sure Thing Principle*). The axiom implies that when comparing two alternatives whose consequences depend on what state is actual, rational agents only consider the states of world where the two alternatives differ. More formally, a simple version of the axiom states that:

If:

	s_1	s_2
L_i	x	z
L_j	y	z

then $L_i \succ L_j$ iff $x \succ y$.

In the choice problem under discussion, this means that you only need to consider the tickets that give different outcomes depending on which alternative is chosen. Hence, you can ignore the fourth column, i.e. tickets 35-100, both when choosing between L_1 and L_2 and when choosing between L_3 and L_4, since these tickets give the same outcome no matter which alternative is chosen. When we ignore this column, however, alternative L_1 becomes identical to L_3 and L_2 to L_4. Hence, by simultaneously preferring L_2 over L_1 and L_3 over L_4, the decision maker seems to have revealed an inconsistency in her preferences.

The second example discussed in last section generates a paradox similar to Allais' if we assume that there is nothing irrational about strictly preferring a lottery that gives the patients an equal chance of receiving the kidney to giving the kidney to either patient without any such lottery being used. If we call the patients Ann and Bob, and let *ANN* represent the outcome where Ann receives the kidney and *BOB* the outcome where Bob receives the kidney, then to represent the aforementioned attitude, which I will refer to as 'Diamond's preference', as maximising the value of an EU function, it must be possible to simultaneously satisfy:

$$U(ANN) < 0.5U(ANN) + 0.5U(BOB) \tag{4.4}$$

$$U(BOB) < 0.5U(ANN) + 0.5U(BOB) \tag{4.5}$$

But that is of course impossible: an average of the values $U(ANN)$ and $U(BOB)$ can never be greater than *both* values $U(ANN)$ and $U(BOB)$.

Again, we can see the tension between Diamond's preference and standard theories of rational choice by noticing that it violates Separability (given the suggested and arguably most natural description of the relevant alternatives). An implication of this axiom is that, given the prospects displayed below, where E represents the outcome of some random event (e.g. a coin toss), $L \succ L_A$ iff $L_B \succ L_A$ and $L \succ L_B$ iff $L_A \succ L_B$. Hence, Diamond's preference in conjunction with Separability implies a contradiction.

	E	$\neg E$
L	*ANN*	*BOB*
L_A	*ANN*	*ANN*
L_B	*BOB*	*BOB*

The fact that both Allais' and Diamond's preferences violate Separability (given this natural description of the alternatives) without seeming irrational (as I argued in the introductory chapter), casts doubt on Separability as a rationality postulate. Moreover, both preferences suggest that the value of actual outcomes often depend on counterfactual ones, since the two preferences violate Separability since the agents in question judge that outcomes in mutually incompatible states are not desirabilistically independent. Both the desire to avoid regret, as manifested in Allais' preference, and the concern for giving each patient a 'fair chance', which seems to be what underlies Diamond's preference, have something to do with counterfactuals. Regret, at least in the situation under discussion, is a bad feeling associated with knowing that one *could have* acted differently, and that if one had, things would have been better. And to say that even if Bob did not receive a kidney he nevertheless had a chance, seems to mean that there is a meaningful sense in which things could have turned out differently – for instance, a coin could have come up differently – and if they had, Bob would have received the kidney. So both Allais and Diamond violate the formal Separability requirement of standard decision theories since they judge that the value of what actually occurs at least partly depends on what could have been.

Perhaps for the reason discussed above, some economists and philosophers have thought that Separability as a requirement on preference is implied by an evaluative assumption we call *Ethical Actualism* (EA). Informally put, EA is the assumption that *only the actual world matters*, so that the desirability of combinations of what actually occurs and what could have occurred only depends on the desirability of what actually occurs. In a well-known defence of Separability, Nobel Laureate Paul Samuelson claims that it would be irrational to violate Ethical Actualism,[1] and since he thinks that EA implies Separability, he takes this to show that it would be irrational to violate Separability. The Separability postulate Samuelson was defending, which is implied by what we above called Separability, states that if some outcome $(A)_1$ is at least as good as $(B)_1$ and $(A)_2$ is at least as good as $(B)_2$, then an alternative that results in either $(A)_1$ or $(A)_2$ depending on whether a coin comes up heads or tails, is at least as good as an alternative that results in $(B)_1$ or $(B)_2$ depending on how the coin lands. Here is Samuelson's informal justification of the axiom:

> [E]ither heads or tails must come up: if one comes up, the other cannot; so there is no reason why the choice between $(A)_1$ and $(B)_1$ should be 'contaminated' by the choice between $(A)_2$ and $(B)_2$. ([Samuelson, 1952]: 672-673)

In other words, the *reason* an evaluation or ordering of alternatives should satisfy Separability, is that there should be no desirabilistic dependencies between mutually incompatible outcomes; in other words, our preferences should satisfy Separability since our evaluation of outcomes should satisfy ethical actualism.

Various philosophers and decision theorists have cited Samuelson's remark favourably. John Broome, who takes it to at least provide a "prima facie presumption in favour of [Separability]", rhetorically asks: "How can something that never happens possibly affect

[1] Samuelson does not really provide an argument for why it would be irrational to violate EA. Instead, he seems to take it to be obvious, and in no need for a special explanation, that it would be irrational to violate EA.

the value of something that does happen?" ([Broome, 1991]: 96). But intuitive as the link between ethical actualism and Separability may be, the former does not (by itself) imply the latter, nor vice versa, as I explain in 4.5.1. Indeed, as we shall see, *both* an axiom that encodes ethical actualism and one that encodes Separability is required for expected utility representation. So even if Samuelson and Broome are right about the intuitive appeal of ethical actualism, that does not in any way establish that Separability is rationally required.

4.3 Jeffrey Desirability

Not all decision theories assume Separability. In particular the version of decision theory developed by Richard Jeffrey makes do with a much weaker condition on preference, which he calls Averaging. (See section 4.6 for a formal statement of Averaging.) In Jeffrey's theory the objects of both the agent's beliefs and desires are propositions, with her degrees of belief in the truth of propositions rationally required to be probabilities and her degrees of desire for their truth required to be desirabilities, where:

> "*the desirability of a proposition is a weighted average of the desirabilities of the cases in which it is true, where the weights are proportional to the probabilities of the cases*" ([Jeffrey, 1983]: 78).

Let's now explore whether we can represent Allais' and Diamond's preferences as maximising Jeffrey-desirability.[2]

As already discussed, Jeffrey defines both his desirability measure, *Des*, and his probability measure, *P*, over a Boolean algebra of propositions from which the impossible proposition has been removed.[3] (In other words, the domain of *Des* and *P* is a set of propositions closed under negation and the classical logical operators.) And the desirability of a proposition, according to this measure, is a weighted average of the different ways in which the proposition can become true, where the weights of each way is given by the probability of the proposition coming true in that way rather than some other way.

Jeffrey takes a proposition to be a *set of possible worlds*. More precisely, if W is the universal set of possible worlds, and Ω the set of subsets of W (i.e. the power set of W), then Jeffrey's desirability and probability measures are defined over Ω, and any subset of W, which I will denote by (non-italic) uppercase letters (A, B, C, etc.), is a proposition according to Jeffrey. We can thus think of each way in which A can be true as a world that is compatible with the truth of A. Assuming for simplicity that there are finitely many mutually exclusive worlds compatible with A,[4] then the Jeffrey-desirability of a proposition is given by:

$$Des(\mathrm{A}) = \sum_i Des(w_i).P(w_i \mid \mathrm{A}) \tag{4.6}$$

Why should we accept this as a measure of the desirability of a proposition? As I mentioned in the first chapter, one way to see the appropriateness of this measure is to think of desirability as *news value*; that is, a proposition A is desirable to an agent to the extent that it would be valuable for her to learn that A is true. Intuitively, it seems that the desirability of

[2] The possibility of representing Allais' preference as maximising desirability would probably not have impressed Jeffrey himself, who was satisfied with Savage's view that Allais' preference reveals some sort of 'error' of judgement ([Savage, 1972]: 102-103; [Jeffrey, 1982]: 722).

[3] For the quasi-uniqueness of the Bolker-Jeffrey representation theorem for Jeffrey's theory, the algebra has to be non-atomic. That is, we must always be able to partition each element into smaller elements.

[4] If we want to assume infinitely many (non-atomic) worlds, we take the integral instead of the sum.

learning the truth of A depends on how desirable are the different ways in which A could be true and the probabilities that it comes true in any one of these ways rather than another.

There are, of course, also some formal justifications for this way of measuring the value of propositions, one of which being that the measure is *partition invariant*. That is, if a proposition A can be expressed as the disjoint disjunction of both $\{B_1, B_2, B_3...\}$ and $\{C_1, C_2, C_3...\}$, then $\sum_{B_i \in A} P(B_i \mid A).Des(B_i) = \sum_{C_i \in A} P(C_i \mid A).Des(C_i)$ (see e.g. [Joyce, 1999], Theorem 4.). The same is not true of the expected utility equation: the same alternative will get assigned different utilities depending on how we partition the state and outcome spaces. In fact, unlike Jeffrey's equation, the expected utility equation can only be used when the state and outcome spaces have been partitioned finely enough to account for *everything* the agent cares about. In other words, our partition needs to be such that given each alternative and state, there is no uncertainty as to the utility value of the outcome associated with that alternative-state pair. If we do not have such fine partitions, as ordinary decision makers rarely (if ever) do, then different partitions will lead to alternatives being assigned different values. Hence, when using expected utility theory to decide how to act, different partitions may recommend different courses of action, as James Joyce points out ([Joyce, 1999], [Joyce, 2000]).

Another formal property of Jeffrey's measure that is very important for the argument below, is hat it does not imply the same (strong) separability property as the expected utility measure. The reason Jeffrey's theory does not imply the Separability axiom is that the contingencies that affect how an alternative turns out are not assumed to be probabilistically independent of the alternative itself. Recall that Separability as previously defined states that:

If:

	s_1	s_2
L_i	x	z
L_j	y	z

then $L_i \succ L_j$ iff x$\succ$ y.

We should expect Separability to fail in the kind of circumstances where Jeffrey's desirability equation has practical implications that differ from the expected utility equation: namely circumstances where there is a probabilistic dependency between the alternative that is chosen and the facts of our world that affect what consequences the alternative has (i.e. the way in which that alternative is actualised). For instance, if z is considered a more desirable outcome than both x and z, and L_i makes s_2 more likely than does L_j, then L_i might be preferred to – and be assigned a higher desirability than – L_j even when the consequence x is not preferred to y. In that case, Jeffrey's equation can, but the expected utility equation cannot, represent the preference in question.

Unfortunately, although Jeffrey's theory does not imply Separability, which seemed to be the property of expected utility theory that generated the tension between that theory and the two preferences under discussion, Jeffrey's theory is also in tension with the two preferences. In fact, in the special case when there is probabilistic independence between the proposition that is being evaluated and the different possible worlds that are compatible with that proposition – as seems to be the case in Allais' and Diamond's examples – Jeffrey's theory becomes a *non-conditional* expected utility theory, and thus also entails Separability.

Moreover, given that we take the desirability of a lottery to be a weighted average of the desirabilities of its possible prizes, as is standardly done when applying Jeffrey's theory, it

is evident that the theory cannot handle Allais' and Diamond's preferences. For instance, given that a person conceptualises the four alternatives that generates the Allais paradox in the way that is standardly assumed (that is, assuming that the first two tables in section 4.1 actually represent the way in which she sees the four alternatives), then (even if we now think of the four alternatives as propositions) there is no Jeffrey desirability function that assigns greater desirability to L_2 than L_1 and also a greater desirability to L_3 than L_4. The same holds for the Diamond paradox: given standard assumptions, there is no desirability function such that L gets assigned greater desirability than both L_A and L_B.

Let's focus on the Diamond paradox to see why the above is true. For Diamond's preference to be compatible with Jeffrey's theory, given the above assumption about how to evaluate lotteries with Jeffrey's theory, there has to be a function *Des* such that:

$$Des(ANN) < Des(ANN).P(ANN \mid \mathrm{L}) + Des(BOB).P(BOB \mid \mathrm{L}) \tag{4.7}$$

$$Des(BOB) < Des(ANN).P(ANN \mid \mathrm{L}) + Des(BOB).P(BOB \mid \mathrm{L}) \tag{4.8}$$

which implies that:

$$Des(ANN) < 0.5Des(ANN) + 0.5Des(BOB) \tag{4.9}$$

$$Des(BOB) < 0.5Des(ANN) + 0.5Des(BOB) \tag{4.10}$$

But again, an average of the desirabilities of *ANN* and *BOB* can never exceed the desirability of both *ANN* and *BOB*.

This shows that there is more at play than just the failure of Separability in the explanation of Allais' and Diamond's preferences. For the standard representation of the two problems, and also the standard application of Jeffrey's theory, implicitly builds in the aforementioned assumption of ethical actualism. Without this assumption (but assuming that the desirability of Ann or Bob getting the kidney is independent of the random event E), Jeffrey's theory just says that:

$$Des(\mathrm{L}) = Des(ANN \wedge \mathrm{L}).Prob(ANN \mid \mathrm{L}) + Des(BOB \wedge \mathrm{L}).Prob(BOB \mid \mathrm{L})$$

and nothing requires that $Des(ANN \wedge \mathrm{L}) = Des(ANN)$ or $Des(BOB \wedge \mathrm{L}) = Des(BOB)$.

So one way to accommodate Allais' and Diamond's preferences within Jeffrey's framework, is to describe the consequences differently from what is standardly done. To some extent, this is Broome's suggestion (although stated within Savage's framework), which I briefly discussed in the introductory chapter, and will discuss in more detail in chapter 6. However, solutions of this kind will be unsatisfactory if they involve introducing new primitive consequences in the representation of the decision problem, without explaining their relationship to the available actions. In particular, they must explain what it is about the form of the lottery L that makes $Des(ANN \wedge \mathrm{L}) > Des(ANN)$. Moreover, to avoid trivialising decision theory by making it allow that *any possible* choice is rational, we should require that exercises of this kind, where new propositions (or consequences) are created to make seemingly problematic preferences compatible with decision theory, adhere to some independently plausible principles.

In the context of Jeffrey's framework, avoiding these objections requires a specification of the propositional structure of lotteries, and how they give rise to the attitudes of Allais' and Diamond's. I do so below by widening the domain of Jeffrey's theory to include coun-

terfactual propositions. And I show that the properties that generate Allais' and Diamond's paradoxes, respectively regret and fairness, then emerge as a relationship between factual and counterfactual conditional propositions. This is not an ad hoc solution to the problems under consideration, I believe, since decision theorists should independently of these problems allow for the value dependencies one often finds between actual and counterfactual outcomes.

But let me first briefly mention why introducing indicative conditionals to Jeffrey's theory (as e.g. done in [Bradley, 1998] and [Bradley, 2007b]) will not solve the problem of representing Allais' and Diamond's preferences. An indicative conditional is generally considered to be what Jonathan Bennett calls *zero intolerant*, as discussed in chapter 2, "meaning that such a conditional is useless to someone who is really sure that its antecedent is false" ([Bennett, 2003]: 45). In other words, we use indicative conditional sentences such as $A \mapsto B$ to make statements about worlds where we think A might be true (where 'might' is understood epistemically, not merely logically or metaphysically). But $A \mapsto B$ provides no information about a world where we *know* A to be false. Hence, it is 'uselessness' to someone who is certain that A is false. So assuming for now that we can formulate desirability and probability as a measure on sentences (which should be understood as derivative of the propositions that make true the sentences in question), it is therefore plausible to assume, as Bradley does, that $Des(\neg A \wedge (A \mapsto B)) = Des(\neg A)$: if A is actually false, then $A \mapsto B$ makes no desirabilistic difference. Thus the conditionals that generate the paradoxes discussed in section 4.1 cannot be indicative conditionals, since the problems they generate consist exactly in the fact that they have desirabilistic impact when their antecedents are believed to be false.

What we need to do, therefore, is introduce *counterfactual* conditionals into Jeffrey's theory. Jeffrey himself tried to solve the problem of providing an account of counterfactuals, but by his own account did not succeed.

> (If I had, you would have heard of it. There's a counterfactual for you.) In fact, the problem hasn't been solved to this day. I expect it's unsolvable. ([Jeffrey, 1991]: 161)

I believe we need not be as pessimistic as Jeffrey in this respect. In fact, I think the account discussed in last chapter is quite useful for the task at hand. In next section I use that account to introduce counterfactuals to Jeffrey's decision theory.

4.4 Counterfactuals

I will not, in this chapter, explain the multidimensional semantics in detail again.[5] However, it might be useful to remind the reader of the main terminology. A possible *counteractual* world under the supposition that A is true, is just a way things might be, or might have been, were A true.[6] If world w_A could be the case under the supposition that A, then we say that w_A is a possible counteractual A-world. If A is false, w_A will be said to be *strictly counterfactual*. But counteractual worlds are not always strictly counterfactual: if A is true then w_A may not only be a possible way things are under that supposition that A, but the way things actually are.

[5] Some technical details about defining probability and desirability for the multidimensional framework, that are not discussed here, can be found in [Bradley and Stefánsson, 2015].

[6] Keep in mind that in this chapter it is useful to think of the semantics in terms of propositions rather than sentences, since we assume that the contents of agent's desires and beliefs are propositions.

Also recall that we now represent the basic possibilities by n-tuples of worlds, e.g. $\omega = \{w_1, w_2, w_3, ...\}$, where by convention the first element (w_1 in this case) always represents a potential actual world, but the reminder potential counteractual worlds under different suppositions. (In what follows, I will for convenience sometimes use ω as a variable for such n-tuples; and sometimes as a variable for a maximally specific proposition that is only true at one n-tuple.) And propositions are now taken to be sets of n-tuples of worlds. The factual proposition A, for instance, is the set of all n-tuples where A is true at the first element. Suppose now that, by convention, the second element in the n-tuples (world w_2 in the n-tuple above) always represents a potential counteractual world under the supposition that A. Then the conditional proposition A $\rightarrow$ B is the set of all n-tuples where B is true in the second element.[7]

In what follows W will be the set of possible worlds, Ω the set of subsets of W. To reduce complexity, let $\{S_1, ...S_n\} \subseteq \Omega$ represent a set of n suppositions, and I will sometime use $\langle X, Y_1, ..., Y_n \rangle$ to denote the proposition that X is the case and Y_i is or would be if S_i. Let $\digamma$ represent the set of elementary possibilities, i.e. the set of all n-tuples, and Γ the set of subsets of $\digamma$; i.e. the set of *multidimensional* propositions.

4.4.1 Probability

Recall from last chapter that the multidimensional semantics is based on the observation that an agent can be uncertain both about what is actually the case and about what is or would be the case if some condition is or were true. One might be pretty sure that the match is to be played tomorrow, for instance, but quite unsure as to whether it would be played were it to rain. In what follows it is important to keep in mind that we have one probability mass function measuring the probability that any world is actual, and one probability mass function for each supposition, each measuring the probability that a world is counteractual given that supposition.

These probability mass functions induce a probability measure, P, on the full set of propositions. Recall from before that I used $\mathcal{P}$ as a probability mass function on W measuring the probability that each world is actual, and $\mathcal{Q}$ as a probability mass function on W_A measuring the probability that each world is counteractual under the supposition that A. We define:

$$P_0(\mathrm{A}) \doteq \sum_{w_i \in A} \mathcal{P}(w_i) \tag{4.11}$$

And as we saw in last chapter, the multidimensional semantics entails that:

$$P_1(\mathrm{A} \rightarrow \mathrm{B}) \doteq \sum_{w_i \in B} \mathcal{Q}(w_i) \tag{4.12}$$

So in our multidimensional possible world model, P_0 serves as a measure of the agent's degrees of belief in the facts, P_i the agent's degrees of belief in the counterfacts under the supposition that S_i. Finally P encodes the agent's state of belief regarding both the facts and the counterfacts, with $P(\langle X, Y_1, ..., Y_n \rangle)$ measuring the joint probability that X is the case and that Y_i is or would be the case if S_i. For P to serve as a measure of such joint probability in

[7] Note that this means that single worlds are truth makers for factual propositions. If w represents the actual world and w is a member of A, then it is true that A. However, the truth value of 'if A then B' is determined by whether the n-tuple that correctly represents both facts and counterfacts is a member of A $\rightarrow$ B.

the manner suggested it must be related to its marginals, P_0, ... P_n, in accordance with the Marginalisation condition discussed in last chapter.

Some much stronger probability principles than Marginalisation will play an important role in the argument that follows, since it turns out that they are implied by expected utility theory, as the theory is typically conceived, but are neither implied by Jeffrey's original theory nor the extension of it that I discuss below. The first condition requires probabilistic independence between what is the case and what merely could have been, while the second requires independence between counterfacts under mutually exclusive suppositions (recall that the non-subscript variables represent what is or could be actual):

Fact-Counterfact Independence: If $X \cap S_i = \varnothing$, then:

$$P(\langle X, Y_i \rangle) = P(X).P(Y_i)$$

Counterfact Independence: If $S_i \cap S_j = \varnothing$, then:

$$P(\langle Y_i, Y_j \rangle) = P(Y_i).P(Y_j)$$

Both conditions are very demanding and it is not difficult to think of counterexamples. Suppose I am meeting my partner and am almost certain that she said we should meet either at location A or B, but I cannot remember which. If I go to A and discover that she is not there, then I become quite certain that if I had gone to location B I would have found her there. So what is actually true clearly influences the probabilities I assign to the different counterfactual possibilities, in violation of Fact-Counterfact Independence.[8] Similarly, suppose I am about to go to location A but am told by my friend that were I to go to location B I would meet my partner. From that I can of course infer that if I go to the location I intended then I won't find my partner there. So the counterfacts under the supposition that I go to one location are not probabilistically independent of those under the supposition that I go to the other, in violation of Counterfact Independence.

It seems clear then that counterfactual reasoning does not typically satisfy these two conditions. And rationality surely does not require that they be satisfied. In fact, certain theories of rational decision-making assume that rational agents *violate* both principles. In game theory with imperfect information, which is a theory about rational strategic decision-making for agents who are uncertain about what moves other 'players' have already made, it is for instance standardly assumed that a rational strategy for figuring out whether a player P has made a particular move M, is to ask oneself what were to happen if P did not make that move. If it turns out that not making move M would lead to a bad outcome for P, then that might reasonably lead one to increase one's credence in the proposition that P has made move M. In other words, it is typically assumed in game theory that rational players violate Fact-Counterfact Independence.

Nonetheless, as we shall see, the above principles are implied by Actualism and Separability (but not by Jeffrey's theory). I believe that a good theory of practical rationality should, if possible, avoid such implausible epistemic implications. Moreover, it seems particularly problematic if a theory of rational individual decision-making contradicts an assumption

[8]Dorothy Edgington has pointed out to me that the following (often used) example might be more useful to explain why Fact-Counterfact Independence does not generally hold. Suppose I am wondering if my friend has gone out. I drive by her house, and see that the lights are not on. This increases my credence in the proposition that she has gone out, since I reason that if she had not gone out, the lights would be on.

that is standardly made in the theory of rational strategic decision-making. Hence, this result casts doubt on the claim that Separability and Actualism is rationally required. I will revisit this point in section 4.6.

4.4.2 Desirability and counterfactual value

Beliefs about counterfactual possibilities play an important role in our reasoning about what we should do (as briefly discussed in chapter 2). So too do our evaluative attitudes to counterfactual possibilities, for instance, through the regret we anticipate if we forego opportunities that would have led to desirable outcomes. And just as our uncertainty about what is the case can be different from our uncertainty about what would be the case if some or another condition were true, so too can our assessment of how desirable something is differ from our assessment of how desirable its truth is on the supposition of some condition or other.

To reflect this we should introduce measures of value on both the facts and the counterfacts in the same way that we introduced probability measures on both. The desirability of possible actual worlds will be represented here by a utility function u_0 on W, while the desirability of possible counteractual worlds under the supposition that S_i will be represented by a utility function u_i on the set W_i induced by that supposition. Finally, the joint desirability of worlds will be measured by a utility function, u, on n-tuples of worlds. For example, $u(\langle w, w_1, ..., w_n\rangle)$ will measure the desirability that w is the actual world, that w_1 is/would be the counteractual world on the supposition that S_1, ..., and that w_n is/would be the counteractual world on the supposition that S_n. (See [Bradley and Stefánsson, 2015] for a discussion of the relationship between the different utility measures.) For convenience, let's assume (as Jeffrey does) that the measures are all zero-normalised in the sense that:[9]

$$\sum_{w\in W} u_0(w).p_w(w) = \sum_{w_i\in W_i} u_i(w_i).p_i(w_i) = \sum_{\omega\in\digamma} u(\omega).p(\omega) = 0$$

From the function u we can determine a corresponding desirability function *Des* on all propositions by defining the desirability of any $\alpha \in \Gamma$ to be the conditional expectation of utility given α. Formally:

$$Des(\alpha) := \sum_{\omega\in\alpha} u(\omega).P(\omega \mid \alpha) \tag{4.13}$$

As can be seen, *Des* is still a Jeffrey-desirability measure: all I have done is to replace the single worlds that appear Jeffrey's measure with n-tuples of worlds (which means that *Des* is now defined over Γ rather than Ω). But now *Des* represents an agent's preferences for the truth of both facts and counterfacts by measuring, for any proposition of the form $\langle X, Y_1, ..., Y_n\rangle$, the desirability that X is the the case and that Y_i is/would be the case if S_i is/were. This concludes the extension of Jeffrey's decision theory to counterfactuals.

4.5 Counterfactual-Dependent Preferences

One reason Allais' and Diamond's preferences cannot be represented as maximising the value of an EU function is that the EU equation implies that given any fixed description of the possibly outcomes of an alternative, the value of what actually occurs never depends on what merely could have been. But for people with Allais' preferences, the desirability of

[9] This may be a controversial assumption. But since I spent much of chapter 1 defending the assumption, I will not discuss it further here.

receiving nothing is not independent of whether or not one could have chosen a risk-free alternative; whereas for people with preferences like Diamond's, the desirability of either patient receiving the kidney is not independent of what would have occurred had some random event turned out differently. So both Allais' and Diamond's preferences, on this interpretation, are dependent on the truth of counterfactuals. Moreover, the part that is causing the violation of expected utility theory – and Jeffrey's theory, when it includes the actualism assumption that is usually made when applying the theory – can in both cases be formalised as a relationship between a set of worlds, C, and a set of *strictly counter-actual* worlds $B_{\bar{A}}$ where $\bar{A} \cap C = \varnothing$. (I will use $\bar{A}$ to denote the complement of set A.)

To make the above claim more precise, let's look at Diamond's preference first and suppose that Diamond wants to use a coin toss to decide who receives the kidney. Let A be the set of worlds where the coin comes up heads and $\bar{A}$ the set of worlds where the coin comes up tails. Let B be the set of worlds where Bob receives the kidney and $\bar{B}$ the set of worlds where Ann receives the kidney. We have thus made two simplifying assumptions. Firstly, it might seem more natural to let A ($\bar{A}$) be the set of worlds where the coin comes up heads (tails) *if* tossed. But nothing is lost by this simplification. Secondly, we have limited our attention to only situations where either Ann or Bob receives the kidney. But what is distinctive about Diamond's preference is what it has to say about situations where a number of individuals have an equal claim to an indivisible good that *some* but not all of them get. (Any kind of welfarism for instance condemns a situation where *none* of the patients receive the kidney.) Hence, since we want to focus on the core of this preference, it seems justifiable to limit our attention to situations where one of Ann and Bob receives the kidney.

The part of Diamond's preference that leads to violation of expected utility theory can then be formulated thus:

$$\langle A \cap B, \bar{B}_{\bar{A}} \rangle \succ \langle A \cap B, B_{\bar{A}} \rangle \tag{4.14}$$

In other words, Diamond prefers the proposition that the coin comes up heads and Bob receives the kidney but Ann would have gotten it had tails come up, to the proposition that the coin comes up heads and Bob receives the kidney and would also have gotten it had the coin come up tails.

Let us then turn to Allais' preferences. Now let A represent the set of worlds where Allais chooses the risky option (which will be L_1 or L_3 depending on the context) and B the set of worlds where Allais is *not guaranteed* to win anything. Unlike when representing Diamond's preference, we need a third (basic) set of worlds to represent Allais' preferences, since the worlds where Allais is not guaranteed to win anything are not necessarily the same as the worlds where Allais wins nothing. And it is relative to a situation where Allais has won nothing that the fact that he could have chosen a risk-free alternative makes a difference. Let D denote the set of worlds where Allais wins nothing. Then the preference that causes Allais to violate expected utility theory can be represented thus:

$$\langle A \cap B \cap D, B_{\bar{A}} \rangle \succ \langle A \cap B \cap D, \bar{B}_{\bar{A}} \rangle \tag{4.15}$$

In other words, according to Allais, winning nothing after having made a risky choice is made worse when it is true that *had he* chosen differently he would definitely have won something.

4.5.1 Preference Actualism and Separability

We have seen that both Diamond's and Allais' preferences exhibit a non-trivial sensitivity to counterfactual states of affairs that is manifested in the violation of a condition that I will call Preference Actualism: the requirement that preferences for prospects be independent of the strict counterfacts. Formally:

Preference Actualism: For all sets of worlds A, B, C such that $C \cap \bar{A} = \varnothing$:

$$\langle C, B_{\bar{A}} \rangle \sim \langle C, \bar{B}_{\bar{A}} \rangle$$

Preference Actualism is of course just a version of the doctrine of ethical actualism that was informally introduced earlier on. In the appendix to [Bradley and Stefánsson, 2015], we prove that preferences that violate Preference Actualism cannot be represented as maximising expected utility, given how expected utility is standardly defined (this will be further discussed in section 4.6 below). And we formally show that a preference might violate Preference Actualism without violating Separability, so this result does not simply follow from the fact that Separability is a necessary condition for expected utility maximisation. To see intuitively how Preference Actualism and Separability differ, consider again the choice problem Diamond is faced with:

	E	$\neg E$
L	*ANN*	*BOB*
L_A	*ANN*	*ANN*
L_B	*BOB*	*BOB*

Recall that in this case, Separability tells us that $L \succ L_A$ iff $L_B \succ L_A$ and $L \succ L_B$ iff $L_A \succ L_B$. Suppose that according to Bill, $L_B \succ L \succ L_A$, in accordance with Separability. But contrary to Preference Actualism, conditional on E being the case, Bob is not indifferent between L and L_A. Bill thinks that it is better if Bob receives the kidney than Ann, but in addition he cares, even after the result of the lottery becomes apparent, about what chance his preferred outcome had (before the lottery was over). Bill is an example of a person who violates Preference Actualism while (in this case) satisfying Separability.

Carl is quite different from Bill. According to Carl, before he knows whether E takes place or not, his preference is $L \succ L_A \sim L_B$, in violation of Separability. However, the reason for this preference is that Carl wants expected welfare to be distributed equally, where the expectation is calculated according to Carls' beliefs. So once Carl has learnt that E, he is indifferent between L and L_A, in accordance with Preference Actualism. Thus, Carl is an example of a person who violates Separability while satisfying Preference Actualism. (This example shows that those who share Diamond's judgements due to *ex ante* egalitarianism, do not necessarily violate Preference Actualism.)

Although Preference Actualism and Separability are logically independent, in the absence of further assumptions, it turns out that the two are logically equivalent given certain assumptions that are either implicitly or explicitly part of standard formulations of expected utility theory such as Savage's [Savage, 1972], and which do seem to be satisfied in Allais' and Diamond's examples; in particular, Centring and an assumption about the probabilistic independence of counterfacts under disjoint suppositions. (This is proved in [Bradley and Stefánsson, 2015].) Hence, given the background of Savage's framework, Allais' and

Diamond's violation of Preference Actualism can be seen as explaining why they violate Separability.

4.5.2 Preference Actualism and desirability maximisation

While expected utility maximisation (as EU is standardly defined) requires adherence to Preference Actualism, the principle is not stated in any axiomatisation of the theory. Therefore, one might worry that although Preference Actualism is also not part of the so-called Bolker's axiomatisation of Jeffrey's theory, which I state in next section, the principle is nevertheless implied by these axioms. But there is no need to worry, as can be seen by the fact that we can construct a Jeffrey desirability measure that violates Preference Actualism. I will work with a simple model based on the set $W = \{w_1, w_2, w_3, w_4, w_5\}$ of five possible worlds and the corresponding set Ω of its subsets, including the events $A = \{w_1, w_2, w_3\}$, $\bar{A} = \{w_4, w_5\}$, $B = \{w_1, w_2, w_4\}$ and $\bar{B} = \{w_3, w_5\}$. For present purposes we only need to focus on one supposition, namely the supposition that A is false. Then the set of elementary possibilities is given by $\mathcal{W} = \{w_1, w_2, w_3, w_4, w_5\} \times \{w_4, w_5\}$, and in particular, $\langle A \cap B, \bar{B}_{\bar{A}} \rangle = \{\langle w_1, w_5 \rangle, \langle w_2, w_5 \rangle\}$ and $\langle A \cap B, B_{\bar{A}} \rangle = \{\langle w_1, w_4 \rangle, \langle w_2, w_4 \rangle\}$ (where the first element in each 2-tuple, i.e. in each pair, represents a potential actual world, but the second element a counterfactual world in the complement to A).

To induce the preferences required, we define a pair of probability and utility mass functions, p and u, on this set of world pairs, by setting $p(\langle w_4, w_5 \rangle) = p(\langle w_5, w_4 \rangle) = 0$ and assigning the values to remaining possibilities displayed in Table 2.

World Pairs	Probability	Utility
$\langle w_1, w_4 \rangle$	0.125	-1
$\langle w_1, w_5 \rangle$	0.125	1
$\langle w_2, w_4 \rangle$	0.125	-1
$\langle w_2, w_5 \rangle$	0.125	1
$\langle w_3, w_4 \rangle$	0.125	-1
$\langle w_3, w_5 \rangle$	0.125	1
$\langle w_4, w_4 \rangle$	0.125	0
$\langle w_5, w_5 \rangle$	0.125	0

Table 2: Probability-Utility Values

Let P and Des be pair of probability and desirability functions on the set of subsets of $\mathcal{W}$ constructed from p and u in the manner previously outlined. It is easy to see that preferences induced by Des will violate Preference Actualism. In particular, they will be such that:

$$\langle A \cap B, \bar{B}_{\bar{A}} \rangle \succ \langle A \cap B, B_{\bar{A}} \rangle \tag{4.16}$$

But by construction they satisfy the standard preference axioms of Jeffrey decision theory. So it is follows that preferences violating Preference Actualism, although not representable as utility maximising, may nonetheless be desirability maximising.

4.5.3 Modelling Allais' and Diamond's preferences

Strictly speaking, 4.14 does not fully represent Diamond's preference in full. Recall that Diamond's preference consists in preferring a lottery (say a coin toss) that results in either Bob or Ann receiving a kidney (alternative L) to giving the kidney to Ann without using a fair lottery (L_A) and also to giving the kidney to Bob without using a fair lottery (L_B). This is how Diamond might evaluate the 'constant' alternatives:

$$Des(L_B) = Des(\langle \mathrm{A} \cap \mathrm{B}, \mathrm{B}_{\bar{\mathrm{A}}} \rangle)$$

$$Des(L_A) = Des(\langle \mathrm{A} \cap \bar{\mathrm{B}}, \bar{\mathrm{B}}_{\bar{\mathrm{A}}} \rangle)$$

But since the lottery can turn out in more than one way, Diamond must, if he is to satisfy Jeffrey's equation, evaluate its desirability as a weighted sum of the ways in which it might turn out, for instance:

$$Des(L) = 0.5Des(\langle \mathrm{A} \cap \mathrm{B}, \bar{\mathrm{B}}_{\bar{\mathrm{A}}} \rangle) + 0.5Des(\langle \bar{\mathrm{A}} \cap \bar{\mathrm{B}}, \mathrm{B}_{\mathrm{A}} \rangle)$$

assuming that (he believes that) a *fair* coin is (properly) tossed to decide who receives the kidney.

There is thus a Jeffrey-desirability function representing Diamond's preference as long as there is a function Des that simultaneously satisfies:

$$Des(\langle \mathrm{A} \cap \mathrm{B}, \mathrm{B}_{\bar{\mathrm{A}}} \rangle) < 0.5Des(\langle \mathrm{A} \cap \mathrm{B}, \bar{\mathrm{B}}_{\bar{\mathrm{A}}} \rangle) + 0.5Des(\langle \bar{\mathrm{A}} \cap \bar{\mathrm{B}}, \mathrm{B}_{\mathrm{A}} \rangle) \tag{4.17}$$

$$Des(\langle \mathrm{A} \cap \bar{\mathrm{B}}, \bar{\mathrm{B}}_{\bar{\mathrm{A}}} \rangle) < 0.5Des(\langle \mathrm{A} \cap \mathrm{B}, \bar{\mathrm{B}}_{\bar{\mathrm{A}}} \rangle) + 0.5Des(\langle \bar{\mathrm{A}} \cap \bar{\mathrm{B}}, \mathrm{B}_{\mathrm{A}} \rangle) \tag{4.18}$$

Since what motivates Diamond's preference is his concern for fairness, he should be indifferent between Bob and Ann actually receiving the kidney. Moreover, the value generated by having used the lottery, or the disvalue generated by not having used the lottery, is according to Diamond independent of whether Ann or Bob actually receives the kidney. Hence, for Diamond:

$$0.5Des(\langle \mathrm{A} \cap \mathrm{B}, \bar{\mathrm{B}}_{\bar{\mathrm{A}}} \rangle) + 0.5Des(\langle \bar{\mathrm{A}} \cap \bar{\mathrm{B}}, \mathrm{B}_{\mathrm{A}} \rangle) = Des(\langle \mathrm{A} \cap \mathrm{B}, \bar{\mathrm{B}}_{\bar{\mathrm{A}}} \rangle) = Des(\langle \bar{\mathrm{A}} \cap \bar{\mathrm{B}}, \mathrm{B}_{\mathrm{A}} \rangle) \tag{4.19}$$

$$Des(\langle \mathrm{A} \cap \mathrm{B}, \mathrm{B}_{\bar{\mathrm{A}}} \rangle) = Des(\langle \mathrm{A} \cap \bar{\mathrm{B}}, \bar{\mathrm{B}}_{\bar{\mathrm{A}}} \rangle) \tag{4.20}$$

Therefore, to be able to represent Diamond's preference as maximising Jeffrey-desirability, all that is required is that there is a Jeffrey-desirability function such that:

$$Des(\langle \mathrm{A} \cap \mathrm{B}, \mathrm{B}_{\bar{\mathrm{A}}} \rangle) < Des(\langle \mathrm{A} \cap \mathrm{B}, \bar{\mathrm{B}}_{\bar{\mathrm{A}}} \rangle) \tag{4.21}$$

And in last section we saw that such functions exist.

The preference exhibited in 4.15 also only *partly* captures Allais' preference. But again, it is not hard to show that in Allais' case all we need to establish is that there is a desirability function such that:

$$Des(\langle \mathrm{A} \cap \mathrm{B} \cap \mathrm{D}, \bar{\mathrm{B}}_{\bar{\mathrm{A}}} \rangle) < Des(\langle \mathrm{A} \cap \mathrm{B} \cap \mathrm{D}, \mathrm{B}_{\bar{\mathrm{A}}} \rangle)$$

And that is the only part of Allais' preference that has the potential to cause trouble; and indeed does cause trouble for EU theory. But an argument like the one provided above

can just as well be made to show that we can represent Allais' preference as maximising desirability.

To sum up: we have now seen that when we use the multidimensional semantics to extend Jeffrey's desirability measure to counterfactual prospects, we can formally express the desirabilistic dependencies that often exist between actual and counterfactual outcomes, and, as a result represent Allais' and Diamonds' preferences as maximising desirability.

4.6 Representation Theorems

In this last section I turn to the question of what axioms a preference relation has to satisfy to be representable as maximising desirability, and what additional axiom the relation has to satisfy to be representable as maximising expected utility.

Here are the two most important axioms that are necessary for desirabilistic representation:

Averaging: If $\alpha \perp \beta$ then $\alpha \preceq (\alpha \vee \beta) \preceq \beta \Leftrightarrow \alpha \preceq \beta$

Impartiality: Suppose $\alpha \approx \beta$ and that for some $\gamma \not\approx \alpha, \beta$ such that $\alpha \perp \gamma$ and $\beta \perp \gamma$, it is the case that $\alpha \vee \gamma \approx \beta \vee \gamma$. Then for all such γ, $\alpha \vee \gamma \approx \beta \vee \gamma$.

Impartiality concerns the connection between an agent's beliefs and her preferences, and, in particular, is a preference test for equiprobability. Although this condition is far from being obvious as a requirement of rationality (as Jeffrey admits [Jeffrey, 1983]: 147), I will not discuss it here. Averaging is a weak version of Separability, and says that a disjunction of two propositions can never be strictly preferred to each disjunct. This seems quite compelling as a rationality requirement, at least when prospects described in sufficient detail, such that they include, for instance, counterfactual properties.

Recall that we say that *Des* and *P* represent $\preceq$ just in case for any A, B: $A \preceq B \Leftrightarrow$ *Des*(A) $\leq$ *Des*(B). Ethan Bolker proved the following theorem, which as we show in [Bradley and Stefánsson, 2015], also holds for the multidimensional framework developed in this chapter:

Theorem 13 ([Bolker, 1966])**.** *Let $\langle \Gamma, \subseteq \rangle$ be a complete, atomless Boolean algebra of prospects. Let $\preceq$ be a complete, transitive and continuous relation on $\Gamma - \{\perp\}$ that satisfies Averaging and Impartiality. Then there exists a pair of desirability and probability functions, Des and P, respectively on $\Gamma - \{\perp\}$ and Γ that jointly represent $\preceq$.*

(Let's say that when the above holds, *Des* is a 'Jeffrey representation' of the preference relation $\preceq$. If, however, an expected utility function *EU* represents $\preceq$, then we say that *EU* is an 'expected utility representation' of $\preceq$.)

Before determining what additional axioms a preference relation between multidimensional prospects has to satisfy to be representable as maximising expected utility, we need to decide how to define expected utility in the multidimensional framework. Given how people normally think of EU theory, a natural suggestion, that stays close to Savage's original theory, is to think of the prospects to be evaluated as vectors of the form $\langle Y_1,...,Y_n \rangle$, where Y_i is the consequence of $\langle Y_1,...,Y_n \rangle$ if state S_i is actual, and define expected utility as follows:

Definition 3 (Expected Utility (EU))**.** *A desirability function Des is an expected utility iff:*

$$Des(\langle Y_1,...,Y_n \rangle) = \sum_{i=1}^{n} Des(Y_i \mid S_i).P(S_i)$$

It should be noted that this definition of expected utility is more general than the usual one in that it allows that the desirabilities of consequences be dependent on the state of the world in which they are realised. In the event that state-independence holds, $Des(\mathrm{Y}_i|\mathrm{S}_i) = Des(\mathrm{Y}_i)$. Then if we let act f be the proposition $\langle \mathrm{Y}_1, ..., \mathrm{Y}_n \rangle$ and $f(\mathrm{S}_i) = \mathrm{Y}_i$, we obtain the familiar Savage formulation of expected utility theory: $Des(f) = \sum_{i=1}^{n} Des(f(\mathrm{S}_i)).P(\mathrm{S}_i)$.

I have already mentioned, in various places, that Separability, in some form or another, is necessary for expected utility maximisation. The following Separability property, which is equivalent to Savage's *Sure Thing Principle* (given the multidimensional framework), is entailed by expected utility maximisation, as EU is defined above:

Separability: If $\mathrm{H} \cap \mathrm{T} = \varnothing$, then $\langle \mathrm{X}_\mathrm{H}, \mathrm{Z}_\mathrm{T} \rangle \preceq \langle \mathrm{Y}_\mathrm{H}, \mathrm{Z}_\mathrm{T} \rangle \Leftrightarrow \langle \mathrm{X}_\mathrm{H}, \mathrm{Z}'_\mathrm{T} \rangle \preceq \langle \mathrm{Y}_\mathrm{H}, \mathrm{Z}'_\mathrm{T} \rangle$

As we have seen, Diamond and Allais both violate Separability (given the most natural description of the alternatives they are faced with). Recall also Preference Actualism from before, which says that whenever $\mathrm{C} \cap \bar{\mathrm{A}} = \varnothing$, $\langle \mathrm{C}, \mathrm{B}_{\bar{\mathrm{A}}} \rangle \sim \langle \mathrm{C}, \bar{\mathrm{B}}_{\bar{\mathrm{A}}} \rangle$. As we have seen, Diamond and Allais also violate Preference Actualism.

As previously mentioned, Separability and Preference Actualism, while expressing a similar intuition, are logically independent of each other except in the context of certain further assumptions. It turns out moreover that *both* Preference Actualism and Separability are necessary for an expected utility representation as EU was defined above (this is proved in [Bradley and Stefánsson, 2015]):

Theorem 14 (Necessity)**.** *Assume Centring. For there to be an expected utility representation of the preference relation $\preceq$ it has to satisfy Separability and Preference Actualism.*

Theorem 15 (Sufficiency)**.** *Assume Centring. If there is a Jeffrey representation of preference relation $\preceq$ that satisfies Separability and Preference Actualism, then there is an expected utility representation of $\preceq$.*[10]

Theorem 14 again shows that expected utility theory (as EU is defined above) is inconsistent with both Allais and Diamond's preferences (given how the relevant outcomes are usually described). Given how reasonable these preferences are, I think we must accept that people can be practically rational without satisfying the postulates of this version of expected utility theory.

To see that this expected utility theory implies implausible epistemic requirements on counterfactual reasoning, notice that Separability and Actualism respectively imply Counterfact Independence and Fact-Counterfact Independence. This is formally shown [Bradley and Stefánsson, 2015], but there is a very intuitive explanation of this implication. What actually occurs cannot be desirabilistically independent of what could have occurred (as Actualism claims) unless the latter contains no information about what the actual world is like, and hence, no information about what is likely to be actual (as Fact-Counterfact Independence states). Similarly, counterfacts under disjoint suppositions cannot be desirabilistically independent of each other (as Separability claims) unless counterfacts under disjoint suppositions provides no information about each others, including probabilistic information (as Counterfact Independence states).

As we have seen, these two independence conditions impose implausible epistemic constraints on counterfactual reasoning. Hence, even if those who endorse Separability

[10] In [Bradley and Stefánsson, 2015] we actually prove a sufficiency theorem for an axiom that is slightly weaker than Preference Actualism (but which togehter with Separability implies Preference actualism). The difference won't matter for the present argument.

and Actualism somehow manage to avoid the problems posed by Allais and Diamond (e.g. by refining their outcome space as Broome [Broome, 1991] suggests), their theory still has implausible epistemic implications. I take it that a good theory of practical rationality should not imply counterintuitive principles of epistemic rationality. Hence, I think this provides evidence for the view that a good theory of practical rationality should neither entail Actualism nor Separability.

Before concluding this chapter, I should point out that one could define expected utility in the multidimensional framework quite differently from the way I define it above.[11] Rather than thinking of the prospects to be evaluated, i.e., the 'acts', as vectors of the form $\langle Y_1,...,Y_n\rangle$, one could think of them as functions **f** from a set of coarse grained descriptions of facts and counterfacts into the set $F^{\sqsubset}$ of n-tuples of worlds. In other words, the the consequences are then thought of as n-tuples of worlds that represent both what is the case and what could have been. Expected utility can then be defined thus:[12]

Definition 5 (Expected Utility* (EU*))**.** *A desirability function Des is an expected utility* iff:*

$$Des(f) = \sum_{i=1}^{n} Des(\langle w_1, ..., w_n\rangle \mid \mathcal{S}_i).P(\mathcal{S}_i)$$

EU* maximisation neither requires Separability nor Preference Actualism as I have defined the two conditions. (In what follows, everything I say about EU* also holds for EU**, as defined in last footnote, except that 'n-tuples of worlds' should be replaced by 'n-tuples of propositions'.) Formally, however, EU* maximisation does require a version of Separability and Actualism. Unlike desirability maximisation, EU* maximisation requires that the n-tuples of worlds be separable. The Actualism requirement of EU* maximisation is, however, rather trivial: it simply requires that once one knows which n-tuple **f** results in, the other n-tuples the prospect could have resulted in do not affect the desirability of **f**. The reason this is a trivial requirement, is that the n-tuple that **f** actually results in itself expresses what other ways **f** could have turned out. The type of Actualism that is required for EU* maximisation could thus be thought of as a requirement that 'only' the modal and non-modal facts determine the desirability of a prospect. But, I would claim, there are no facts beyond the modal and non-modal facts. So this version of Actualism says that the only thing that matters desirabilistically is everything there is.

EU* theory is not inconsistent with Allais' and Diamond's preferences. Nor does it entail the implausible epistemic principles that I discussed above. Therefore, I would suggest that one lesson of this chapter is that those who insist on using some version of (unconditional) expected utility theory, rather than Jeffrey's theory, should take the consequences to be n-tuples of worlds (or n-tuples of propositions as defined in fn. 12), rather than singleton worlds (or one-dimensional propositions). However, the problem remains that this theory is *partition dependent* (as Richard Bradley and I further discuss in [Bradley and Stefánsson, 2015]).

[11] I thank Jim Joyce for pressing me on this issue.

[12] Alternatively, one could have a more coarse grained version of expected utility:

Definition 4 (Expected Utility** (EU**))**.** *A desirability function Des is an expected utility** iff:*

$$Des(f) = \sum_{i=1}^{n} Des(\langle X_1,...,X_n\rangle \mid \mathcal{S}_i).P(\mathcal{S}_i)$$

where each X_i represents what is or would be the case if $\mathcal{S}_i$ is or would be actual.

4.7 Concluding Remarks

I have now completed the primary aim of this thesis: to develop a decision theory that allows us to express desirabilistic dependencies between actual and counterfactual outcomes and is consistent with Diamond's and Allais' preferences. The final two chapters build on this result. In this chapter the focus was on the way in which a counterfactual can matter desirabilistically in worlds in which its antecedent is (believed to be) false. As we will see in chapter 5, there are other ways in which a counterfactual can have a desirabilistic impact, not all of which concern a world in which the counterfactual's antecedent is false. Jeffrey's desirability measure, extended to counterfactuals, provides natural measures for each of these ways in which counterfactuals matter desirabilistically. In chapter 6, I apply this discussion to moral philosophy. Diamond's example illustrates that what could have been is often highly relevant for the moral value of an actual outcome. I use this observation to distinguish between modal and non-modal versions of consequentialism, and suggest that while the best-known versions of consequentialism (e.g. classical utilitarianism and welfare economics) belong to the latter category, some entrenched moral intuitions make the modal version seem more attractive.

Chapter 5

Desirability of Conditionals

5.1 Introduction

In last chapter we saw how we can extend Richard Jeffrey's desirability measures to counterfactuals, and as a result avoid certain well-known paradoxes in the theory of rational choice. There are other and perhaps less paradoxical (or problematic) ways in which counterfactuals, and conditionals in general, influence the desirability of factual propositions. A conditional can, for instance, carry undesirable information by telling us that what we find ceteris paribus desirable implies something we find undesirable, or by informing us that some proposition we desire is unlikely to actualise. The aim of this chapter is to explore the different ways in which a conditional can influence the desirability of actual outcomes; or, to put it differently, the different ways in which conditionals can be carriers of desirability.

The chapter is structured as follows. In the next section, I discuss three ways in which a conditional can be (un)desirable. Firstly, a conditional, A $\rightarrow$ B, may be (un)desirable because it informs an agent about the consequences or implications, and hence the all things considered desirability, of the factual proposition A. Secondly, the above conditional may be (un)desirable because of the information it carries about the probability of A occurring. Finally, the conditional may be (un)desirable because it tells the agent that *had* A occurred, then B would also have occurred. For each of these, I discuss examples that I hope show how common these desirabilistic impacts of conditionals are to ordinary, evaluative reasoning. (As far as I can tell, all rational and reasonable ways in which a conditional can be desirable or undesirable are special cases of these three.)

In the third section I show that measures for these different kinds of desirabilistic impacts of conditionals can all be derived from a formula that measures the 'total' desirability of a counterfactual. This formula is a natural application of Richard Jeffrey's notion of *news value*. The derived measures differ from each other in one very important respect: only measures for the first two kinds of desirability can be reduced to measures of the desirability of *factual* propositions. I discuss this difference in more detail in section 5.4, where I argue that counterfactual desirability is *primitive* in the sense that it cannot be derived from the desirability of factual propositions only.

Some of the conditionals I discuss are clearly counterfactuals, a few of them are clearly indicatives, but for some of the conditionals it is more questionable how we should classify them. And the first two ways in which a conditional can be desirable that I discuss apply both to indicatives and counterfactuals, whereas the last applies only to counterfactuals. However, as will become clear in section 5.3.1, it is safe to use the measure for counterfactuals for all of these. The measure for counterfactuals is more general than the measure

for indicatives, in the sense that when we use a measure for counterfactuals to measure the desirability of an indicative conditional, certain assumptions that hold for indicatives automatically transform the measure into the (more simple) measure that is appropriate for indicatives. Hence, for the present purposes, we need not worry whether the conditionals I discuss are counterfactuals or indicatives.

5.2 Three Kinds of Desirability

5.2.1 Factual, desirabilistic information

Example 1. *After having taken his girlfriend to the airport, Bob briefly thinks that it would be nice to see her again that same night. Soon however Bob realises that seeing his girlfriend that same night would mean that she had missed her flight. Hence he comes to the conclusion that all things considered, it is not desirable that he sees her again that same night.*

Call the proposition that Bob sees his girlfriend 'that same' night A. Let B denote the proposition that Bob's girlfriend misses her flight. Bob finds A ceteris paribus desirable. But when it occurs to him that $A \rightarrow B$, he concludes that all things considered, A is not so desirable. In cases like this I will say that the conditional in question carries (negative) *factual, desirabilistic information*; or simply that it has (negative) *factual desirability*.

The next example is also an instance of a conditional carrying such information:

Example 2. *A policy maker is considering whether to increase the minimum wage or not. She prefers, ceteris paribus, to increase the minimum wage, but since she believes*[1] *that if the minimum wage is increased unemployment will rise, and considers protecting the current level of employment to be more important than raising the minimum wage, she all things considered prefers to not raise the minimum wage.*

This type of desirability of conditionals is exemplified in a vast number of cases, that have in common that the desirability of some factual proposition either increases or decreases through the realisation of what is implied (or will be made true) by the proposition given the truth of some conditional proposition. Moreover, as the second example shows, it is because conditionals have desirability of this kind that they are indispensable to rational decision making.

5.2.2 Probabilistic information

Example 3. *After taking his friend to the airport, it occurs to Bill that it would be nice to see him again that same night. However, Bill realises that if he were to see his friend that same might, then that would mean that the latter had missed his flight. While Bill does not really care if the friend misses his flight or not, he knows that the friend is not the type of person that is likely to miss a flight, and that they are hence unlikely to see each other that night.*

Call the proposition that Bill sees his friend 'that same' night A. Let B denote the proposition that Bill's friend misses his flight. In this example, it is not the case that the truth of the conditional $A \rightarrow B$ makes A less all things considered desirable to Bill. However, the

[1] For the present purposes, it does not matter whether the conditionals that we will discuss are actually *true* or not; all that matters is whether the agents in question *believe* them to be true. (However, I will sometimes be talking about agents 'knowing' some conditionals to be true, which may indicate that they are true. But whether or not they are true is not important.) The same holds below when I discuss the probabilistic information that a conditional carries and the counterfactual desirability of a conditional.

conditional carries information about the probability of the factual proposition A. In cases like this I will say that the conditional A → B carries (negative) *probabilistic information*.

The conditional in the next example carries information of similar sort:

Example 4. *Professor K knows that her university desperately needs more government funding. She also knows that of the political parties currently fighting in a government election, only the Progressive Party (PP) has increasing government funding for universities on its agenda. In other words, she knows that if the university gets more funding, then PP must have won the election. Unfortunately for K, it is highly unlikely that PP wins the election.*

This type of desirability of conditionals is also exemplified in a vast number of cases. Common to these cases is that the truth of a particular conditional makes a factual proposition, that the agent in question either finds desirable or undesirable, either more or less likely to occur.

5.2.3 Counterfactual desirability

The structure of the next example should by now be familiar:

Example 5. *Alice is offered job E. Although the job is quite risky, and could lead to unemployment after a few years, it is also very exciting and is on the top of Alice's preference ranking over possible jobs. Before she accepts the job, however, she is offered a less exciting job, S, that comes with a guaranteed job security until retirement. She still prefers E over S, but nevertheless the fact that she has been offered the job S lowers the expected value of E, since she reasons that if she accepts E and then loses her job, she will think to herself 'had I chosen S I would have been guaranteed a stable, long term job', which she predicts will make unemployment feel even worse.*

Call the proposition that Alice chooses the more secure job A. Let B denote the proposition that Alice is in employment. Then the truth of the conditional A → B reduces, according to Alice, the desirability the proposition ¬B that she ends up unemployed (which, given our story, is only compatible with having chosen the more risky option, i.e. with proposition ¬A). In cases like this, I will say that the conditional carries negative *counterfactual, desirabilistic information*; or simply that it has (negative) *counterfactual desirability*.

The counterfactual in the next example (which should also be familiar) has *positive* counterfactual desirability.

Example 6. *A hospital has only one kidney but two patients, Ann and Bob, who are in equal need of a kidney, are equally deserving, have equal rights to treatment, etc. The head of the organ transplant department decides to flip a fair coin to decide which patient gets the kidney. He finds this more desirable than giving the kidney straight away to either patient. The reason is that if the coin toss decides, then even if Ann (Bob) gets the kidney, there is a meaningful sense in which Bob (Ann)* could have *gotten it; which, according to the department head, makes either outcome fair.*

It is perhaps worth stressing that although I call this kind of desirability, unlike the other two discussed above, *counterfactual* desirability, that does not mean that the conditionals we discussed when explaining the other kinds of desirability cannot be called 'counterfactual conditionals' as well. In the second example, of the policy maker who is considering whether to raise the minimum wage or not, the relevant conditional is being used to decide what to do, which according to many people make it a counterfactual (or subjunctive) conditional. Moreover, in this particular case, we assumed that the policy maker prefers, given the truth

of the conditional, not to raise the minimum wage. Hence, the antecedent will not become true, which may provide even further justification for calling the conditional a *counterfactual* conditional. However, the crucial difference between that example and those discussed in the current section, is that when the policy maker is deliberating, she is considering the conditional as being *open*; i.e., it has not yet been determined whether the antecedent will become true or not. By deliberating with the conditional open, however, she comes to a conclusion which turns the conditional into what we might call a 'properly' counterfactual conditional – i.e., a conditional where it has already been determined that the antecedent is false. In contrast, when e.g. Alice is deliberating about what difference the truth of $A \rightarrow B$ makes to a situation where $\neg A \wedge \neg B$, she is supposing that the antecedent is false, and from that standpoint she evaluates the desirability of the conditional as a properly counterfactual conditional.

The two examples of counterfactual desirability have in common that the truth of a counterfactual increases or decreases the all things considered desirability of the negation of the antecedent and/or the consequent. Counterfactual desirability of this sort is not uncommon in both practical and normative reasoning, and poses a problem for expected utility theory, as we saw in last chapter.[2]

5.3 A General Measure for the Desirability of Conditionals

5.3.1 News value of conditional

Recall that according to Jeffrey's decision theory, the desirability of any proposition, A, is a weighted average of the different (mutually exclusive) ways in which the proposition can be true, where the weight on each way A_i that proposition A can be true is given by $P(A_i \mid A)$. Last chapter established that we can extend the theory to counterfactuals. As we have seen, this means that we can measure the desirability of a counterfactual (which will of course usually be done against the background of what factual propositions the agent in question believes to be true). Given multidimensional theory of conditionals I have been discussing, $Des(A \boxrightarrow B)$ then measures the desirability that B is true in the world that is counteractual under the supposition that A.

To calculate the news value of a counterfactual, we need some partition of the ways in which counterfactuals can be true. What counts as such a partition depends on the semantics of conditionals. Things will be much simpler if we assume that both Strong and Weak Centring holds for conditionals in general (respectively, $A \wedge B$ implies that $A \rightarrow B$ is true and $A \wedge \neg B$ implies that $A \rightarrow B$ is false). So for the present purposes, I will rely on the following partition of the propositions compatible with $A \boxrightarrow B$: $\{\{A \wedge B\}, \{\neg A \wedge B \wedge (A \boxrightarrow B)\}, \{\neg A \wedge \neg B \wedge (A \boxrightarrow B)\}\}$. Given this partition, the desirability of $A \boxrightarrow B$ is given by:

$$\begin{aligned} Des(A \boxrightarrow B) = {} & Des(A \wedge B).P(A \wedge B \mid A \boxrightarrow B) \\ & + Des(\neg A \wedge B \wedge (A \boxrightarrow B)).P(\neg A \wedge B \mid A \boxrightarrow B) \\ & + Des(\neg A \wedge \neg B \wedge (A \boxrightarrow B)).P(\neg A \wedge \neg B \mid A \boxrightarrow B) \end{aligned} \quad (5.1)$$

The above equation is not always satisfactory, since what we are often interested in is how valuable news it is that the conditional is true compared with the news that it is not true.

[2] Amartya Sen [Sen, 1985] discusses two additional examples where what (merely) could have been influences the desirability of actual states of affairs.

Equation 5.1 can be easily extended to measure that contribution:[3]

$$\begin{aligned} Des(\mathrm{A} \boxrightarrow \mathrm{B}) - Des(\mathrm{A} \boxrightarrow \neg\mathrm{B}) = {} & Des(\mathrm{A} \wedge \mathrm{B}).P(\mathrm{A} \wedge \mathrm{B} \mid \mathrm{A} \boxrightarrow \mathrm{B}) \\ & - Des(\mathrm{A} \wedge \neg B).P(\mathrm{A} \wedge \neg\mathrm{B} \mid \mathrm{A} \boxrightarrow \neg\mathrm{B}) \\ & + Des(\neg\mathrm{A} \wedge \mathrm{B} \wedge (\mathrm{A} \boxrightarrow \mathrm{B})).P(\neg\mathrm{A} \wedge \mathrm{B} \mid \mathrm{A} \boxrightarrow \mathrm{B}) \\ & - Des(\neg\mathrm{A} \wedge \mathrm{B} \wedge (\mathrm{A} \boxrightarrow \neg\mathrm{B})).P(\neg\mathrm{A} \wedge \mathrm{B} \mid \mathrm{A} \boxrightarrow \neg\mathrm{B}) \\ & + Des(\neg\mathrm{A} \wedge \neg\mathrm{B} \wedge (\mathrm{A} \boxrightarrow \mathrm{B})).P(\neg\mathrm{A} \wedge \neg\mathrm{B} \mid \mathrm{A} \boxrightarrow \mathrm{B}) \\ & - Des(\neg\mathrm{A} \wedge \neg\mathrm{B} \wedge (\mathrm{A} \boxrightarrow \neg\mathrm{B})).P(\neg\mathrm{A} \wedge \neg\mathrm{B} \mid \mathrm{A} \boxrightarrow \neg\mathrm{B}) \end{aligned} \quad (5.2)$$

This measures the desirability that the truth of the conditional contributes, compared with its falsity, to the possible states of the world compatible with the conditional. What have we accomplished by this? As I try to show below, the different kinds of desirability exemplified at the beginning of this paper can be measured by natural assignments of values to the variables in equation 5.2.

Recall that above I said that no harm is done when the above measure for the desirability of a counterfactual is applied to an indicative conditional. When the measure is applied to such conditionals, then, given the assumptions about indicatives that I defended in chapter 2, formula 5.1 becomes considerably simpler. Firstly, recall that indicative conditionals are usually taken to be *zero-intolerant*, which implies that $\mathrm{A} \mapsto \mathrm{B}$ tells us nothing about worlds where A is false. But then $\mathrm{A} \mapsto \mathrm{B}$ cannot make any desirabilistic difference to worlds where $\neg\mathrm{A}$ is true, so $Des(\neg\mathrm{A} \wedge (\mathrm{A} \mapsto \mathrm{B})) = Des(\neg\mathrm{A})$. Secondly, setting aside McGee's counterexample (which I discussed in chapter 2), it is generally accepted that unlike counterfactual conditionals, indicative conditionals are probabilistically independent of their antecedents. In other words, $P(\mathrm{A} \mid \mathrm{A} \mapsto \mathrm{B}) = P(\mathrm{A})$.

If we assume the above conditions for a conditional $\mathrm{A} \mapsto \mathrm{B}$, then when we apply measure 5.1 to $\mathrm{A} \mapsto \mathrm{B}$, it becomes:

$$Des(\mathrm{A} \mapsto \mathrm{B}) = Des(\mathrm{A} \wedge \mathrm{B}).P(\mathrm{A}) + Des(\neg\mathrm{A}).P(\neg\mathrm{A}) \quad (5.3)$$

which is exactly Bradley's formula for calculating the desirability of indicative conditionals (see e.g. [Bradley, 1998]).

Given the discussion above and in previous chapters, it should be evident that that the first assumption (i.e., that $\mathrm{A} \mapsto \mathrm{B}$ makes no desirabilistic difference in worlds where $\neg\mathrm{A}$ is true) does not hold for the conditionals that are involved in Allais' and Diamond' reasoning, as was discussed in last chapter. For according to that reasoning, the conditionals in question do have a desirabilistic impact on worlds where their antecedent is false. We have also seen, in chapter 2, that the assumption of probabilistic independence does not in general hold for counterfactuals. Unfortunately, I also argued that some indicatives do not satisfy probabilistic independence, which means that we cannot in general use equation 5.3 to measure the desirability of indicative conditionals. There is, however, no need to worry about that here, since I will in general use the more complicated measure for all conditionals. And of course, when applied to conditionals for which either or both of the above assumptions hold, the measure automatically simplifies in the way that is appropriate for the conditional in question.

[3]Throughout this chapter, I will assume the Conditional Excluded Middle, since it is implied by the theory of conditionals I am using, but nothing of importance hangs on that assumption.

5.3.2 Savage and the evaluation of conditionals

In last chapter we saw that introducing counterfactuals into expected utility theory has undesirable consequences: it implies norms for epistemic reasoning about counterfactuals that are clearly implausible. Focusing on Savage's version of expected utility theory, I will in this section discuss further reasons for why Jeffrey's theory is more natural than Savage's if we want value functions to be defined over conditionals.

It is unclear how we could even talk about the goodness or badness of conditionals within Savage's framework. According to Savage, there are three elements of a decision problem, and similarly three types of things that agents have attitudes about: acts, consequences, and states of the world that determine the consequences of each act. And an agent has conative attitudes towards consequences and acts, but cognitive attitudes toward states of the world. The most natural way to make room for conditionals, within Savage's framework, would be to think of them as *acts*. But since an act, according to Savage, is a function from the set of mutually exclusive and collectively exhaustive states of the world into the set consequences, we must then think of these acts as *partitioning conditionals* [Bradley, ms]. That is, we think of an act as a prospect of the form $A = (\mathrm{A}_1 \rightarrow \mathrm{B}_1)(\mathrm{A}_2 \rightarrow \mathrm{B}_2)...(\mathrm{A}_n \rightarrow \mathrm{B}_n)$ where the A_is are mutually incompatible and collectively exhaustive states of the world and the B_is are maximally specific consequences. Then B_j is, for instance, the consequence of act A if state of the world A_j occurs. In some cases this seems quite natural (in particular when we are evaluating the goodness of *indicative conditionals*). For instance, imagine that I am considering how desirable is the conditional that I get £100 if a fair coin comes heads up but nothing if the coin comes tails up. Then it seems natural to represent the relevant 'act' as $(H \rightarrow £100)(\neg H \rightarrow £0)$. The value of the conditional, given Savage's framework, is given by $P(H).U(£100) + P(\neg H).U(£0)$.

In other cases this method to evaluate the goodness of a conditional doesn't work. Firstly, in some cases the 'states' (antecedents) will not be probabilistically independent of the 'acts' (partitioning conditionals), as Savage's theory requires. Recall the conditional from the third example: 'if I see my friend tonight (A), then he will have missed his flight (B)'. To evaluate the goodness of this conditional, within Savage's framework, we might think of it as the partitioning conditional $(\mathrm{A} \rightarrow \mathrm{B})(\neg\mathrm{A} \rightarrow \neg\mathrm{B})$.[4] But in this case it is not true that the 'states', A and ¬A, are probabilistically independent of the 'act', $(\mathrm{A} \rightarrow \mathrm{B})(\neg\mathrm{A} \rightarrow \neg\mathrm{B})$.

Secondly, in some cases the 'consequences' (consequents) will not, on this approach, be desirabilistically independent of the 'states' (antecedents), contrary to what Savage requires. Consider the conditional: 'if I work hard, then I will ace the exam; otherwise I will not'. Let *wh* denote the antecedent and *ace* the consequent. To evaluate the goodness of this conditional, within Savage's framework, we might think of it as the 'act' $(wh \rightarrow ace)(\neg wh \rightarrow \neg ace)$. The consequence *ace* is desirabilistically independent of the state *wh* if and only if $Des(ace \mid wh) = Des(ace)$, which means that $Des(ace \wedge wh) - Des(wh) = Des(ace)$. Now imagine someone who derives great pleasure from achievements that she know she 'earned', but not as much pleasures from the achievements as such (i.e. $Des(ace \wedge wh) > Des(ace)$). However, she finds working hard, in and of itself, not so desirable (i.e. $Des(wh) < 0$). We can think of acing and working as 'complementary goods' from the perspective of this agent. In that case, $Des(ace \wedge wh) - Des(wh) > Des(ace)$. Hence, the 'consequence' is not desirabilistically

[4] We might want to partition the antecedent and consequent further to allow for the possibility that our friend misses his flight but, nevertheless, does not come and see us.

independent of what 'state' of the world brings it about.[5]

So it seems that we cannot really discuss the desirability of conditional propositions within Savage's framework by interpreting conditionals as acts, and acts as partitioning conditionals. Perhaps we should then add conditionals as consequences? The problem with this is that conditionals are usually neither desirable nor undesirable in and of themselves, but rather undesirable or desirable given what else is true. In Savage's theory, however, it is assumed that consequences are maximally specific with regards to everything that matters for their utility. Hence, we cannot simply add each conditional as one consequence, but, for each conditional α that we are interested in, we would need to have one consequence for each possible but mutually incompatible combination of α and non-modal facts that are relevant for the evaluation of α. Suppose that for conditional α there are only two contingencies that effect its utility: a and $\neg a$. As far as I can tell, the most plausible way in which we could assign it a utility value, given Savage's framework, would be to think of it as an act, despite what was previously said, and its utilities as being given by: $u(\alpha(a)).P(a)+u(\alpha(\neg a)).P(\neg a)$. But now we have the same problem as before, since for many of the conditionals we are interested in, the relevant contingency is not probabilistically independent of the conditional. Suppose, for instance, that α is the conditional 'if more money is printed, inflation will rise'. Then a and $\neg a$ might respectively be that the central bank has and has not decided to print more money, which is clearly not probabilistically independent of α.

5.4 Specific Measures for the Desirability of Conditionals

Reasoning to a large extent involves supposing certain conditions to be true and trying to establish what follows from the supposition. To evaluate the factual desirability of the conditional A $\rightarrow$ B, for instance, we supposed that A, and ask ourselves, from this point of view, how (un)desirable it is that the conditional is true rather than false. To evaluate the counterfactual desirability of A $\rightarrow$ B, on the other hand, we suppose that $\neg$A, and ask ourselves, from this point of view, how (un)desirable it is that the conditional is true. Thus the factual desirability of A $\rightarrow$ B is measured by one particular update on the probability function in equation 5.2 and the counterfactual desirability of A $\rightarrow$ B by another update on the probability function in 5.2. To evaluate the probabilistic information of A $\rightarrow$ B, on the other hand, we do not start by any updating on a probability measure P, but rather try to figure out what assuming the conditional, compared with its negation, means for P. Using the above (Ramsian) observation, let's now turn to deriving measures for the ways in which a conditional can be (un)desirable.

5.4.1 Measure for factual desirability

In the first example, about Bob who was taking his girlfriend to the airport, we interpreted A as the proposition that Bob sees his girlfriend that same night, and B the proposition that she misses her flight. When Bob reasons about how desirable it would be to learn that he sees his girlfriend that same night, he evidentially supposes A and tries to evaluate what

[5]It might of course seem unnatural to interpret working hard as a state of the world. But keep in mind that the aim here is to see to what extent it is possible to formulate the conditional in question within Savage's theory. The above arguments were meant to show that Savage's probabilistic and desirabilistic independence assumptions make this task quite difficult. His reliance on a clean distinction between acts, states and consequences makes the task even more difficult.

follows.[6] Now the change in the Bob's attitude w.r.t. the overall desirability of A that takes place once he realises the truth of $A \rightarrow B$ can be measured by:

$$Des_A(A \rightarrow B) - Des_A(A \rightarrow \neg B) = Des_A(A \wedge B) - Des_A(A \wedge \neg B) \tag{5.4}$$

which we get from 5.2 when we conditionalise on A (since in general we have $P^+_\alpha(\alpha \mid \beta) = 1$ and $P^+_\alpha(\neg\alpha \mid \beta) = 0$).

Now recall from chapter 1 that I am assuming that the desirability of A under the (evidential) supposition that B is given by: $Des_B(A) = Des(A \mid B) - Des(B)$. Hence:

$$Des_A(A \rightarrow B) - Des_A(A \rightarrow \neg B) = Des(A \wedge B) - Des(A \wedge \neg B) \tag{5.5}$$

This, I suggest, measures the difference that the truth of the conditional makes, from the perspective of Bob, to the overall desirability of seeing his girlfriend that same night. In other words, this measures the *factual, desirabilistic information* that the conditional carries. (And the same formula can measure the factual desirability of the conditional in example 2.)

It is perhaps worth stressing that while we have been talking about 'changes' in an agent's attitudes to the *all things considered* desirability of some proposition A that take place when the agent realises what other propositions are implied by A (together with a true conditional $A \rightarrow B$), this does not (necessarily) mean that there has been a change in attitude to A in and of itself. We might have a *preference change* with respect to A and $\neg A$, since by realising the truth of the conditional the agent's preference for $\neg A$ compared to A may change. However, it may still be the case – and by assumption still is the case in examples 1 and 2 – that they *would* still prefer the alternative that they did prefer before learning the truth of the conditional, had they not gotten (or realised) this new piece of information. In other words, we should not represent the changes that take place in these cases by a new *conditional* desirability function. Rather, we have an update on the old desirability function that takes place due to changes in the *beliefs* of the agent in question.

5.4.2 Measure for probabilistic desirability

In the third example, about Bill and his friend, we denoted the proposition that Bill sees his friend that same night by A, and used B for the proposition that the friend misses his flight. I suggested an assumption to isolate the probabilistic desirability of the conditional: we assume that Bill does not really care if his friend misses his flight or not. Hence, we have $Des(A \wedge (A \rightarrow B)) = Des(A)$. We can then also assume that in a world where Bill does not see his friend that same night, it makes no difference to Bill whether or not it is true that if he had seen his friend the latter would have had missed his flight; nor does it make a difference to Bill, in such a world, whether the friend actually has missed a flight or not. Hence, $Des(\neg A \wedge B \wedge (A \rightarrow B)) = Des(\neg A \wedge \neg B \wedge (A \rightarrow \neg B)) = Des(\neg A \wedge (A \rightarrow B)) = Des(\neg A \wedge (A \rightarrow \neg B)) = Des(\neg A)$. These equalities mean that equation 5.2 (which measures

[6] For my argument to go through, the supposition has to be evidential, since I am assuming that $P_A(A \mid A \rightarrow B) = 1$, but for contrary-to-factual suppositions, we presumably have: $P^{\Box}_A(A \mid A \rightarrow B) = P(A \mid A \rightarrow B)$. When reasoning about what to do, we therefore need to be very careful about how we partition the state space we are working with. In particular, to avoid making suboptimal choices in Newcomb's problem and similar choice situations where an alternative is probabilistically, but not causally, related to good or bad outcomes, we could for instance partition our state space into causally homogenous cells, and compare the alternatives cell-by-cell (as e.g. Nancy Cartwright suggests [Cartwright, 1979]).

the desirability of a conditional being true rather than false) is equivalent to:[7]

$$Des(\mathrm{A} \rightarrow \mathrm{B}) - Des(\mathrm{A} \rightarrow \neg\mathrm{B}) = Des(\mathrm{A}).P(\mathrm{A} \mid \mathrm{A} \rightarrow \mathrm{B}) - Des(\mathrm{A}).P(\mathrm{A} \mid \mathrm{A} \rightarrow \neg\mathrm{B}) + Des(\neg\mathrm{A}).P(\neg\mathrm{A} \mid \mathrm{A} \rightarrow \mathrm{B}) - Des(\neg\mathrm{A}).P(\neg\mathrm{A} \mid \mathrm{A} \rightarrow \neg\mathrm{B}) \quad (5.7)$$

And this captures how (un)desirable it is, from the point of view of Bill, that the conditional, $\mathrm{A} \rightarrow \mathrm{B}$, is true rather than false. For given the above assumptions, the desirability of the conditional in question, according to Bill, is entirely determined by, firstly, the desirability of A (relative to $\neg\mathrm{A}$) and, secondly, the impact the conditional has on the probability of A. Strictly speaking, equation 5.7 measures more than what I above called the 'probabilistic information' of a conditional, for it measures both the probabilistic information of $\mathrm{A} \rightarrow \mathrm{B}$ and how good or bad it is that the conditional makes this probabilistic difference. To keep things simple, I will occasionally refer to what equation 5.7 measures as the 'probabilistic desirability' of the conditional $\mathrm{A} \rightarrow \mathrm{B}$. If we, however, wanted to measure just the probabilistic information of $\mathrm{A} \rightarrow \mathrm{B}$ for C, it seems most natural to use the ratio:

$$\frac{P(\mathrm{C} \mid \mathrm{A} \rightarrow \mathrm{B})}{P(\mathrm{C} \mid \mathrm{A} \rightarrow \neg\mathrm{B})} \quad (5.8)$$

Similarly, if we imagine that professor K from the fourth example only cares about which party wins the election to the extent that it influences whether or not her university gets more funding, then we can also use equation 5.7 to measure how (un)desirable she finds the truth of the conditional 'if we get more funding, then PP must have won the election'.

5.4.3 Measure for counterfactual desirability

Recall that in the fifth example, we are assuming that having been offered the stable job, *S*, reduces according to Alice the expected value of the more exciting job, *E*. For the fact that she could choose a stable, long term job, makes the prospect of unemployment (which is only compatible with choosing *E*) seem even worse. We used A for the proposition that Alice accepts the more secure job *S*. Let $\neg\mathrm{A}$ denote the proposition that Alice accepts job *E*. In other words, we assume that Alice is only considering these two jobs and will take one of them. Let B be the proposition that Alice is employed. Then when Alice reasons about what it would be like to find herself in a situation of unemployment after having turned down the more secure job, she is supposing two things. Firstly, she is supposing that she has already chosen job *E* over job *S*; that is, she is supposing that $\neg\mathrm{A}$. Secondly, she is supposing that she is unemployed; i.e. supposing that $\neg\mathrm{B}$.

Combining these two suppositions with equation 5.2 gives us:[8]

[7] The above equalities in conjunction with equation 5.2 directly give us:

$$Des(\mathrm{A} \rightarrow \mathrm{B}) - Des(\mathrm{A} \rightarrow \neg\mathrm{B}) = Des(\mathrm{A}).P(\mathrm{A} \mid \mathrm{A} \rightarrow \mathrm{B}) - Des(\mathrm{A}).P(\mathrm{A} \mid \mathrm{A} \rightarrow \neg\mathrm{B}) + Des(\neg\mathrm{A}).P(\neg\mathrm{A} \wedge \mathrm{B} \mid \mathrm{A} \rightarrow \mathrm{B}) - Des(\neg\mathrm{A}).P(\neg\mathrm{A} \wedge \mathrm{B} \mid \mathrm{A} \rightarrow \neg\mathrm{B}) + Des(\neg\mathrm{A}).P(\neg\mathrm{A} \wedge \neg\mathrm{B} \mid \mathrm{A} \rightarrow \mathrm{B}) - Des(\neg\mathrm{A}).P(\neg\mathrm{A} \wedge \neg\mathrm{B} \mid \mathrm{A} \rightarrow \neg\mathrm{B}) \quad (5.6)$$

From this we get equation 5.7, since in general $P(\alpha \wedge \beta) + P(\alpha \wedge \neg\beta) = P(\alpha)$.

[8] Here I am assuming the definition of conditional desirability that I defended in chapter 1. On that definition, $Des_{\neg\mathrm{A}\wedge\neg\mathrm{B}}(\mathrm{A} \rightarrow \mathrm{B}) - Des_{\neg\mathrm{A}\wedge\neg\mathrm{B}}(\mathrm{A} \rightarrow \neg\mathrm{B}) = Des(\neg\mathrm{A} \wedge \neg\mathrm{B} \wedge (\mathrm{A} \rightarrow \mathrm{B})) - Des(\neg\mathrm{A} \wedge \neg\mathrm{B}) - [Des(\neg\mathrm{A} \wedge \neg\mathrm{B} \wedge (\mathrm{A} \rightarrow \neg\mathrm{B})) - Des(\neg\mathrm{A} \wedge \neg\mathrm{B})] = Des(\neg\mathrm{A} \wedge \neg\mathrm{B} \wedge (\mathrm{A} \rightarrow \mathrm{B})) - Des(\neg\mathrm{A} \wedge \neg\mathrm{B} \wedge (\mathrm{A} \rightarrow \neg\mathrm{B}))$. Incidentally, Jeffrey's formula, which I criticised in chapter 1, delivers the same result in this case.

$$Des_{\neg A \wedge \neg B}(A \rightarrow B) - Des_{\neg A \wedge \neg B}(A \rightarrow \neg B) =$$
$$Des(\neg A \wedge \neg B \wedge (A \rightarrow B)) - Des(\neg A \wedge \neg B \wedge (A \rightarrow \neg B)) \quad (5.9)$$

This measures the difference, according to Alice, that having been offered the stable job makes to the desirability of being unemployed after having declined the offer; i.e. the counterfactual desirability of the conditional in question, w.r.t. this particular situation. Given the story we told and the situation that Alice is imagining, the value of equation 5.9 is negative: the truth of the (counterfactual) conditional in question makes the the situation less desirable than the truth of its converse. Hence, since $\neg A \wedge \neg B$ is one of the possible outcomes of choosing option E (i.e. of $\neg A$), the conditional in question reduces the expected value of this option.

Equation 5.9 also measures the desirability of the relevant (counterfactual) conditional in the kidney example (which will be further discussed in next chapter).

The second part of equation 5.2 – $Des(\neg A \wedge B \wedge (A \rightarrow B)).P(\neg A \wedge B \mid A \rightarrow B) - Des(\neg A \wedge B \wedge (A \rightarrow \neg B)).P(\neg A \wedge B \mid A \rightarrow \neg B)$ – also measures some sort of counterfactual desirability. However, this is counterfactual desirability of a different kind than what we have been discussing so far, since this time we are measuring what difference the counterfactual makes to a situation where the antecedent is false but the consequent is true anyway. The interpretation of this part of the equation is perhaps most straightforward in the example of the policy maker (i.e. example 2). If she decides not to increase the minimum wage ($\neg A$) but unemployment nevertheless rises (B), then she might regret not having increased the minimum wage. But this regret will presumably be much larger if it had been the case that increasing the minimum wage would lead unemployment *not* to increase (i.e. if $A \rightarrow \neg B$). Hence, the conditional $A \rightarrow B$ has positive counterfactual desirability, according to the policy maker, w.r.t. a situation where $\neg A \wedge B$. The conditional also has positive counterfactual desirability, according to the policy maker w.r.t. a situation where $\neg A \wedge \neg B$. For if she does not increase minimum wage, contrary to what she, ceteris paribus, desires, and the level of unemployment stays the same, then (given the truth of $A \rightarrow B$) she can take comfort in the thought that if she had increased the minimum wage, the unemployment level would have risen. Hence, $A \rightarrow B$ has overall *positive* counterfactual desirability to the policy maker, even though we assumed that it has *negative* factual desirability.

The last point – that the same conditional can be regarded as either desirable or undesirable depending on whether it is being evaluated as an open or 'properly' counterfactual conditional – is worth emphasising. We just saw that in the second example the relevant conditionals has negative factual desirability but positive counterfactual desirability. The opposite seems to be true in the fifth example, of Alice who is offered a job. Before she has decided whether to accept the secure job or not – i.e., when $A \rightarrow B$, 'if I choose the more secure job then I will be guaranteed a stable, long term job', is still an *open conditional* – she might be glad to learn the truth of the conditional in question. But when she evaluates the conditional under the supposition that the antecedent is false – i.e. when she evaluates the desirability of $A \rightarrow B$ as a *properly counterfactual conditional* – and also supposes that the consequent is false, then she finds the conditional *un*desirable, since it makes her regret not having chosen the less risky alternative.

5.4.4 Combining the measures

When I introduced the examples at the beginning of this chapter, I spoke as if in each example the relevant conditional had only one of the three kinds of desirabilities we have discussed. We could to isolate the first and the third kinds of desirability by having the agents in the relevant examples *suppose* either the truth or falsity of the antecedents. And we managed to isolate the second kind by assuming that the agents in those examples care about the truth of of the antecedents but are indifferent between the truth and falsity of the consequents. But it is of course possible, and will often be the case, that a conditional has all three kinds of desirabilities at the same time; in which case we need the full equation 5.2 to calculate the 'overall' desirabilistic difference that the truth of the conditional makes (compared with its falsity).

Let's focus on Bob, in our first example, who has just taken his girlfriend to the airport. The conditional that if he sees his girlfriend that night then she will have missed her flight, $A \rightarrow B$, carries negative factual desirabilistic information, we said, but presumably *positive* counterfactual desirability, since if he doesn't see his girlfriend he can take comfort in the thought the she would have missed her flight if he did. It seems equally natural to assume that the conditional in question carries some probabilistic information. For given the story we told, i.e. that the girlfriend is trying to catch a flight, we may assume that the truth of the conditional reduces the probability of the antecedent becoming true. Taken together, then, the 'total' desirability of $A \rightarrow B$, for Bob, is the sum of, firstly, the factual, desirabilistic information of $A \rightarrow B$, and, secondly, the counterfactual, desirabilistic information of $A \rightarrow B$; weighted by the probabilistic information of $A \rightarrow B$.

5.5 More on Counterfactual Desirability

So far I have derived measures for different ways in which a conditional can be desirable or undesirable, from Jeffrey's original desirability measure. Before concluding this chapter, I first discuss a fundamental difference between, on one hand, factual and probabilistic desirability of a conditional, and, on the other hand, counterfactual desirability. I then briefly discuss how others have treated what I call 'counterfactual desirability'.

5.5.1 Primitiveness of counterfactual desirability

There is a fundamental difference, as we have seen, in the measures that I have proposed for the different kinds of desirability of conditionals: counterfactual desirability is never determined entirely by the desirability of factual propositions. (That is, the formula for measuring counterfactual desirability has a counterfactual in the desirability measure on the right hand side of the equality.) At first sight, it might seem regrettable that we cannot get rid of conditionals in the formula we use to measure counterfactual desirability. However, I think this is exactly as it should be, since this type of desirability fundamentally concerns the *relationship* between what is and what would have been. Below I try to give an informal argument for why this type of desirability is *primitive*, in the sense that the 'goodness' that this relationship generates is not a function the goodness of the factual propositions involved. On the other hand, the factual desirability of a conditional is always, and the probabilistic desirability is sometimes, determined by the desirability of the factual propositions the conditional in question carries information about.

Let me first explain why factual desirability (or undesirability) of a conditional is not primitive in the same way as counterfactual desirability. In these cases a conditional, $A \rightarrow B$, has desirability because it tells us that A will not become true unless B does. The factual desirability of a conditional is, thus, a function of the desirabilities of the factual propositions that it carries information about. Once we know, for some particular agent *i*, the values of $Des(A \wedge B)$ and $Des(A \wedge \neg B)$, we can infer the factual, desirabilistic information that $A \rightarrow B$ carries for *i*. Hence, we do not need to know the desirability of any conditionals – nor the *probability* of conditionals – in order to establish the factual desirability of a conditional.

Counterfactual desirability, on the other hand, is primitive. The desirability or undesirability of $\neg A \wedge \neg B \wedge (A \rightarrow B)$, cannot, in all cases, be inferred from the desirability of factual propositions alone. Take the coin flip case, and assume that Bob got the kidney in a world where the coin lands tails. (Let's now denoted this proposition by $\neg C \wedge \neg D$.) Now the desirability of $\neg C \wedge \neg D \wedge (C \rightarrow D)$, from the point of view of the health care official, is not entirely determined by the desirability of $\neg C \wedge \neg D$ and the desirability of $C \rightarrow D$, taken on their own; nor of course the desirabilities of C, $\neg C$, D and $\neg D$. For the desirability of the conjunction, $\neg C \wedge \neg D \wedge (C \rightarrow D)$, is also partly determined by how much the department head values *fairness*, which, as I will further argue in next chapter, in this particular case fundamentally concerns a relationship between a factual proposition and a counterfactual. One way to put it is that the conjunction and the counterfactual *complement* each other, in the same way that chocolate complements coffee; and the desirability of the two, either coffee-and-chocolate or the conjunction-and-counterfactual, cannot be determined by the desirability of each item on its own.

Considering the case of Alice who is job-hunting leads to the same conclusion. Alice is reasoning about what it would be like to find herself unemployed after having chosen the more risky option (i.e. she is supposing that $\neg A \wedge \neg B$). She concludes that the truth of the counterfactual conditional – that if she had chosen the more secure option she would still be employed – makes the situation even less desirable than if the conditional had not been true. Again the desirability of $\neg A \wedge \neg B \wedge (A \rightarrow B)$ is neither determined entirely by the atomic, factual propositions involved when taken on their own; nor is it simply a sum of the desirabilities of $\neg A \wedge \neg B$ and $A \rightarrow B$. When the two conjuncts are simultaneously (believed by Alice to be) true Alice feels *regret* for not having taken the more secure job. And this regret is neither due to $\neg A \wedge \neg B$ nor $A \rightarrow B$ by themselves, but stems from an interaction between the two. As I discuss in the introductory chapter, I am claiming that Alice's feeling of regret (assuming that it is reasonable) reflects, rather than causes, the badness of her situation. And, to repeat, this badness does not reside in each conjunct, $\neg A \wedge \neg B$ nor $A \rightarrow B$, but stems from an interaction between the two.

The probabilistic desirability of a conditional is not determined in all cases by the desirability of factual propositions only. If for agent *j*, $Des(\neg A \wedge \neg B \wedge (A \rightarrow B)) \neq Des(\neg A \wedge \neg B)$, then whenever the conditional in question affects the probability of this complex proposition, $\neg A \wedge \neg B \wedge (A \rightarrow B)$, we need the value of $Des(\neg A \wedge \neg B \wedge (A \rightarrow B))$ to measure the probabilistic desirability of $A \rightarrow B$ for *j*. There is, moreover, a different sense in which probabilistic desirability cannot be reduced to factual proposition; namely, in the sense that the *probabilities* cannot be so reduced. Fundamental to probabilistic desirability is the difference between (or the ratio of) $P(C \mid A \rightarrow B)$ and $P(C \mid A \rightarrow \neg B)$. And since this formula cannot in general be reduced to the probability of factual propositions. This irreducibility of the probabilistic effects of conditionals will not, unlike the irreducibility of the desirabilistic

effects of counterfactual conditionals, be of much concern to the rest of this thesis.

To sum up: counterfactual desirability fundamentally concerns the relationship between factual and counterfactual propositions. Hence, we have an intuitive justification for the result in 5.4 that counterfactual desirability cannot, unlike factual desirability, be reduced to the desirability of factual propositions.

5.5.2 Others on 'counterfactual desirability'

Although, to the best of my knowledge, nobody has systematically discussed what I am calling 'counterfactual desirability', the paradoxes that I claim are generated by counterfactual desirability have been widely discussed, in particular Allais'. In this section I will (very briefly) explain how the main solutions to the Allais paradox differ from mine. Maurice Allais himself used the idea of *complementarity* to explain and justify what I have called the 'Allais' preference' ([Allais, 1979], pp. 88-92). But unlike what I suggest above, Allais did not seem to have had in mind that a counterfactual proposition complements a factual proposition. Rather, Allais' idea seems to have been that *states of the world* – i.e. states that are not under the agent's direct control – complement each other. That is, it seems that Allais' idea was that an outcome in one possible state of the world, conditional on an act, cannot be evaluated independently of what happens in other possible states of the world, conditional on that same act. Therefore he thought that it is a mistake to require, as Separability does, that when we compare two acts we ignore the states of the world where the two acts generate the same outcome.

My explanation of the Allais preference (and of Alice's reasoning) is more general than Allais'. I agree that what happens in one state of the world, conditional on one act, cannot always be evaluated independently of what happens in other states of the world, conditional on the same act, as e.g. the coin-flip and kidney example illustrates. However, I also claim that what happens in one state of the world, conditional on one act, cannot be evaluated independently of what happens in some (perhaps the same) state of the world, conditional on some *other* feasible act. And that, it seems to me, is what explains the Allais preference (and Alice's reasoning). For the relevant difference between, on one hand, ending up with nothing in the first of Allais' choice situations, and, on the other hand, ending up with nothing in Allais' second choice situation, crucially depends on the availability of a risk free prospect (i.e. certain act) in the first but not the second situation.

Here is another way to state the difference between my suggestion and Allais'. Let's partition the states of strictly counterfactuals worlds into set $W_{\neg A}$ and W_A, where the former is the set of possible but counterfactual worlds that differ from the actual world in ways that are *not* under the control of a particular agent i, while the latter is the set of possible but counterfactual worlds that differ from the actual world in ways that *are* under i's control. My claim is that the counterfactuals that matter to i's evaluation of factual prospects are not limited to the worlds in W_N but also concern worlds in W_A, contrary to what Allais seems to suggest.

Daniel Kahneman and Amos Tversky make another famous attempt at explaining the Allais preference. Their explanation is that people tend to "overweight outcomes that are considered certain, relative to outcomes which are merely probable" ([Kahneman and Tversky, 1979], p. 265). Recall from last chapter that if subjects have the option of choosing a risk-free prospect, the extra 33% chance at £2500 is normally not considered worth an extra 1% risk of getting nothing; but when comparing two risky prospects, the same extra chance

of getting £2500 *is* considered to be worth the same increase in the risk of getting nothing.

I think that it is a mistake to explain the Allais preference by saying that people *overweight* outcomes that are certain.[9] People are presumably right when they think that if they give up the risk-free prospect and then are unlucky enough to end up with nothing, they will feel *regret* for not having chosen the prospect that would have guaranteed them something. But if people are right in their estimation of how they will feel in such a situation, and their preference reflects this estimation, then Kahneman and Tversky are wrong to claim that people *overweight* the outcomes of the risk-free prospect.

The main disagreement I have with Kahneman and Tversky's treatment of the Allais paradox, however, is that their response does not explicitly refer to the importance of counterfactuals. Hence, their characterisation of the Allais preference, and the 'prospect theory' they offer as an alternative to expected utility (EU) theory, does not allow for a unified treatment of the different violations of EU theory that are all caused by what I call 'counterfactual desirability'. In particular, their theory cannot account for the preference for flipping the coin, in the kidney example, in the same way that they account for the Allais preference.

Graham Loomes and Robert Sugden provide an analysis of the Allais preference that comes closer to mine, because they directly refer to the importance of *regret* over what could have been [Loomes and Sugden, 1982]. Moreover, like myself, and unlike Kahneman and Tversky, they do not explain this preference by some sort of mistake on the behalf of the decision maker (see last fn.). However, their explanation of the Allais preference is less general than mine, in that it cannot account for e.g. the preference for tossing the coin in the kidney example. Their 'regret theory' is design to deal with cases of *rejoice and regret*, where what actually occurs is more or less desirable than what could have been; in other cases their theory does not differ from EU theory. In example 6, however, what could have been is neither more nor less desirable than what actually occurs (we assumed that the health care official is indifferent between which patient actually receives the kidney). Their theory is, therefore, no different from EU theory when it comes to dealing with such cases.

Here is a more precise way of making the above point. Lommes and Sugden's *modified utility* is given by

$$m_{ij}^{k} = c_{ij} + R(c_{ij} - c_{kj}) \tag{5.10}$$

where c is a 'choiceless' utility function (representing the utility an agent derives from an outcome without having chosen it), i refers to the act chosen, j to the actual state of the world, and k the act that could have been chosen. R is what they call a 'regret-rejoice' function, that measures how much regret or rejoice the agent feels when the value of $c_{ij} - c_{kj}$ respectively negative or positive. When $c_{ij} - c_{kj} = 0$, however, $R(c_{ij} - c_{kj}) = 0$. Hence, 'modified utility' does not differ from standard Savage or von Neumann-Morgenstern style utilities when there is no difference in the value of what is and what could have been.

Moreover, both Loomes and Sugden's regret theory and Kahneman and Tversky's prospect theory moreover allow for violations of transitivity (which says that if A is preferred to B and B to C then A should preferred to C). As we saw in last chapter, however, the theory that I develop to account for Allais' preference (and Diamond's preference) does not allow for violations of transitivity. Since I take transitivity to be a requirement of rationality,

[9] As this talk of *overweighting* certain outcomes shows, Kahneman and Tversky do not, (unlike e.g. Allais) try to *rationalise* Allais' preference. In fact, they claim that departures from the standard axioms of decision theory, like those often exhibited in the Allais choice situation, "must lead to normatively unacceptable consequences" (ibid, 277).

I think that unlike mine, neither regret theory nor prospect theory should be thought of as a theory of *rational* choice.

The closest treatment of 'counterfactual desirability' to mine is perhaps Broome's [Broome, 1991]. In some sense, he offers a unified treatment of what I call 'counterfactual desirability', and hence his theory deserves a special treatment. In the Allais example, Broome adds 'disappointment' to the outcome of ending up with nothing after having given up the risk-free prospect; and he adds 'treated unfairly' to the outcome where a patient does not get a kidney without having lost in a lottery. Thus, if we translate Broome's solution to the propositional language I am working with, then Broome's solution is to add to the factual propositions in question some *other* factual propositions, i.e. that the agent feels disappointment or that the patient is treated unfairly.

The primary benefit of my solution over Broome's (as I further discuss in next chapter), is that it explicitly models the desirabilistic relationships between actual and counterfactual outcomes that seem to be at the heart of Diamond's and Allais' preferences. Another weakness of Broome's treatment, relative to mine, is that formally it is not at all clear how the factual propositions Broome adds to the relevant outcomes are related to the other factual propositions involved. In the Allais paradox, for instance, he simply adds (as a primitive) the proposition that the agent feels disappointment to the proposition that the agent gets nothing after having turned down the risk-free prospect.[10] Hence, if we now let A represent the proposition that Allais chooses the risk-free prospect, B the proposition that Allais wins nothing, and C the proposition that Allais feels disappointed, Broome makes Allais' preference compatible with expected utility theory by pointing out that $Des(\neg B \wedge C)$ may be different from $Des(\neg B)$. But that means that formally, C need not be related to A and B.[11]

Yet another disadvantage of Broome's treatment, which I will discuss further in next chapter, is that his treatment does not really explain why Allais' and Diamond's preferences have generated these well-known paradoxes for decision theory. Broome adds regret and fairness as primitive properties to the outcome space. I, on the other hand, enrich the set of prospects to include counterfactual prospects, such that the regret and fairness *emerges* as a relationship between factual and counterfactual propositions. As I explain in next chapter, decision theorists (and economists) have traditionally been unwilling to accept that counterfactual outcomes can affect the value of actual outcomes. Since fairness and regret in the examples I have been discussing fundamentally concern such a relationship between what is and what could have been, I think that my solution explains why Diamond's and Allais' preferences have historically generated these problems for decision theory.

5.6 Concluding remarks

In this chapter I showed how measures for the three ways in which a conditional can be (un)desirable can be derived from Richard Jeffrey's general desirability measure. I take it that the most important lesson of this chapter, is that unlike factual (and sometimes

[10] That is, this is how we would describe what Broome does in a propositional framework like Jeffrey's.

[11] Broome himself might be happy with that, since perhaps he thinks that there need not be any systematic relationships between the facts, the counterfacts and an agents regrets. (I thank James Joyce for pointing this out.) Nevertheless, many people would, I think, claim that the type of regret that we are considering is irrational (or at least unreasonable) if, for instance, the agent believes that she could not have done otherwise, or believes that the actual outcome is the best possible outcome that she could reasonably expect. To state and discuss conditions like these, we need to be explicit about what the relationship is between the added variable and the other outcomes in our model.

probabilistic) desirability of conditionals, counterfactual desirability is not determined by the desirability of factual propositions. Focusing on the kidney allocation example, I will use this observation in next chapter to construct what I call *Modal Consequentialism*, and suggest that it satisfies some entrenched moral intuitions better than *Non-Modal Consequentialism* (such as classical utilitarianism and welfare economics).

Chapter 6

Fairness and Counterfactuals

6.1 Introduction

Consequentialists hold that the moral value of an alternative is determined by its consequences. This position allows for a variety of different views, for instance depending on how narrowly we define *consequences*, and the way in which the values of different consequences are combined when evaluating the overall value of an alternative. This chapter explores two views within this broad consequentialist school. One view, which I call *non-modal consequentialism* (NMC), claims that the moral value of an alternative is determined by its *non-modal consequences* and that there should be no interaction between consequences in different states of the world. The second view, which I call *modal consequentialism* (MC), claims that the moral value of an alternative is determined by both its *modal and non-modal* consequences and that consequences in different states of the world can interact (in a sense explained below).[1,2]

I will use the kidney example, which should be familiar by now, to explore the difference between modal and non-modal consequentialism. The main features of the example to keep in mind, for the present purposes, is that we are imagining a situation where a hospital has only a single kidney but two patients, Ann and Bob, who are in equal need of the kidney, have equal rights to treatment, etc. (More generally, we assume that in every respect that you find relevant for the decision of who should receive the kidney, Ann and Bob's situation is exactly symmetric.) To make the example even more clear, let's now imagine that Ann and Bob do not know that there is one kidney but two needing patients, nor will they know why they got the kidney if they do, or why they didn't if they don't.

According to what I call the *Fair Chance View* (FCV), we should toss a fair coin, or hold some other lottery that gives each patient a 0.5 chance of winning, to decide whether Ann or Bob receives the kidney. Below I show that unlike modal consequentialism, non-modal consequentialism is not consistent with the FCV. But existing versions of modal consequentialism do not, I contend, respect the intuition behind the FCV either. The main aim of this chapter is to show that the multidimensional decision theory developed in

[1] This distinction has, to my knowledge, not yet received any attention. Phillip Pettit has recently argued that some goods, such as love and friendship, make *modal demands*, in that they should persist through changes or after events that will (in all likelihood) never actualise. (Pettit discussed this in his 2011 Uehiro Lecture at the University of Oxford.) Similarly, my discussion establishes that fairness often makes modal demands, in that its requirements concern what happens not only in the actual world but also in merely possible worlds. But as far as I am aware, Pettit does not discuss the aforementioned distinction within consequentialism.

[2] A related distinction is that between *ex post* and *ex ante* consequentialism; the former judging the moral value of alternatives by their *actual* consequences, the latter by their *expected* consequences. As we shall see, there are situations where ex post and ex ante consequentialists may be in agreement as to what action to perform, but modal and non-modal consequentialists disagree.

chapter 4 is a modal-consequentialist theory that does better in this regard.

The next section defines the Fair Chance View and the two versions of consequentialism more precisely. In section 6.3, I use Leonard Savage's [Savage, 1972] classical decision theory to show the contradiction between the FCV and non-modal consequentialism, and remind the reader why Separability could be seen as the main culprit. Section 6.4 briefly discusses a modal-consequentialist theory that does not satisfy Separability. Although this theory can be made consistent with a preference for tossing a coin in situations like the one described above, it does not, I argue, satisfy the intuition behind the FCV. John Broome [Broome, 1991] has famously shown that we can, given the right description of alternatives (in particular their consequences), make Separability compatible with the preference for tossing a coin. As I explain in section 6.5, the resulting theory is modal, but nevertheless violates the intuition behind the FCV. Finally, in section 6.6 I try to provide further motivation for the decision theory developed in chapter 4, by showing that it is a version of modal consequentialism that satisfies the intuitions behind the FCV. In section 6.7 I discuss some advantages the new theory has over other versions of modal consequentialism.

6.2 The FCV and Two Forms of Consequentialism

Most people seem to have the intuition that in circumstances like those described in the example discussed above, we should hold a lottery to decide how to distribute the good in question. To justify this intuition from a consequentialist point of view, we need to show that the *consequences*[3] of holding the lottery are better than the consequences of giving the kidney to either Ann or Bob without holding such a lottery. There may be many different consequentialist justifications of the discussed intuition. But the one I will focus on is the following: A consequence (or situation) where Ann has received the kidney as a result of a lottery is (strictly) morally better than a consequence where Ann has received the kidney without 'winning' it in a lottery, *because* in the former case Bob *had a chance*. I take this common justification to follow from the more general *Fair Chance View*:

Fair Chance View (FCV)**.** *Suppose n individuals are in equal need of an indivisible good $m < n$ of which we are about to distribute, and that the individuals are identical in every other respect that is morally relevant to the decision of who should receive a good. Then a situation (or consequence) where m of these individuals receive the good but all n individuals* had an equal chance *of receiving the good is (strictly) morally better than a situation where m of these individuals receive the good and it is not true that all n individuals had an equal chance of receiving the good.*

Giving people a (or an equal) chance of getting a good, in situations like the one under discussion, is valuable in and of itself, according to the FCV as I understand it, rather than merely instrumentally valuable. (I will assume also that on this view, a situation where any individual in question has received the good in question is fair just in case a lottery was used to determine who was to receive a good.) I will not attempt to make a normative assessment of the view, nor address the many and deep philosophical issues surrounding it. For instance, I will set aside questions about whether the view requires that we distribute equally *objective* chances of receiving the good, or whether an equal distribution of *subjective* probabilities suffices (and if so, subjective probabilities of *whom*). Instead, I will try to capture this common view a bit more formally. The main thing to notice, for the present purposes, is the relation between chance and *counterfactuals*. What does it mean to say that even though

[3] In what follows, I will talk of 'consequences', 'outcomes' and 'situations' interchangeably.

Ann actually got the kidney, Bob had *a chance* of receiving the kidney? It means that things *could* (in some meaningful sense) have turned out differently, and if they had, Bob would have received the kidney.[4] Using the possible world framework for counterfactuals, we can express this by saying that a situation (in world w) where Ann has received the kidney is made morally better by the 'existence' of a possible world w' that *only* differs from w in that, firstly, a lottery turns out differently from the way it does in w, and, secondly, Bob receives the kidney.

As already indicated, one claim to be defended in this chapter is that unlike modal consequentialism, non-modal consequentialism is incompatible with the Fair Chance View. Before defending this claim, let me define the two views a bit more precisely. Above I said that according to consequentialism, the moral value of an alternative is determined by its consequences. But this description may be somewhat misleading. To be more precise, let us say, using the terminology developed in [Broome, 1991], that according to consequentialism in its most general form, the value of an alternative is determined by how it distributes consequences across *locations*. In Broome's view, there are three *dimensions* of these locations to consider: different states of the world, different people, and different times. (As we will see, we should add *counterfactual* dimensions to this list.) To keep things simple, I will assume that each location in both the dimension of time and people is valued equally (as is the case according to most forms of *utilitarianism*). Hence, for the present purposes, I define consequentialism, in its most general form, as the claim that the value of an alternative is determined by how it distributes consequences across different states of the world.

This characterisation is compatible with different forms of consequentialism, for instance depending on how consequences in different states of the world are weighted in the calculation of the overall value of an alternative.[5] But more importantly for the present purposes, consequentialism, thus characterised, comes in different forms depending on, firstly, whether we allow that modal properties matter for the moral value of the consequence in each (or some) state; and, secondly, whether we allow for the possibility that, when evaluating the overall value of an alternative, the contributions that consequences in some states make depends on consequences in some other states.

What I call 'non-modal consequentialism' is a strict version of consequentialism, in that it neither allows that modal properties can be of moral importance nor for interactions between consequences in different states of the world:[6]

Non-Modal Consequentialism. *The moral value of an alternative is determined by how it distributes non-modal consequences across different states of the world, and consequences in different mutually incompatible states of the world make independent contributions to the overall value of an alternative.*

From a general consequentialist theory (as described above) we get non-modal consequen-

[4] The claim that Bob had an *equal* chance can just as naturally be captured in terms of counterfactuals. For it simply means that the counterfactual outcome where Bob receives the kidney was just as likely (when the lottery took place) as the actual outcome of the lottery.

[5] According to *ex ante* consequentialists, for instance, the consequence in each state of the world is weighted by the probability of that state being actual. Accord to *ex post* consequentialists, however, consequences in all states except the actual one get weighted by 0.

[6] Where exactly to draw the line between the modal and non-modal properties of our world is a difficult question, which I will not directly address. But I take it that chances, counterfactuals and dispositions are clearly part of the modal properties of the world, whereas the "spatiotemporal arrangement of local qualities" ([Lewis, 1994]: 474) is part of the non-modal properties of the world. In any case, I think that solving the problem of how to demarcate the modal from the non-modal is more important for those who claim that only one of these types of properties is of moral relevance, than for those (like myself) who admit that *both* modal and non-modal properties are of importance.

tialism when we add the following principles:

> **First Principle of NMC.** The moral value of a consequence in a particular state of the world is fully determined by the *non-modal* properties of that consequence. Thus in each state of the world, only the actual consequence in that state is of moral significance.
>
> **Second Principle of NMC.** If alternative A has different consequences depending on whether state of the world s_1, s_2, s_3, etc., turns out to be actual (where s_1, s_2, s_3, etc., are mutually incompatible), then for any of these s_i, the value that the consequence in s_i contributes to the overall moral value of A is independent of the consequence in any $s_j \neq s_i$.

I will call a consequentialist theory *modal* if it violates either the first or the second principle of NMC. Here is what the two principles have in common, which justifies calling a theory that violates either of these modal. If a theory does not satisfy the first principle, then the *value* of consequences in one state of the world may depend on what occurs in other states of the world, but if a theory does not satisfy the second principle, then the *contribution* that a consequence in one state of the world makes towards the overall value of an alternative may depend what occurs in other states of the world. Violations of both principles thus imply that there is some sort of value dependency between what occurs in different, mutually incompatible states of the world.

The first principle of NMC is a version of Actualism, which I discussed in chapter 4, but now stated as a moral principle for consequentialist ethics.

The second principle of NMC is related to *Separability*, that was discussed in chapter 4. Separability is usually discussed as a property of *preferences* – or, in a moral context, a property of what we might call 'betterness judgements'.[7] Let's remind ourselves what Separability requires (in a single-dimensional framework). We can represent each alternative A by an n-tuple, e.g. $A = \langle a_1, a_2, ...a_n \rangle$, where the a_is are the possible consequences of A. Now take the alternative A and create two new alternatives: A_b created by replacing a_i in the original alternative with b and A_c created by replacing a_i with c. Do the same for alternative D: create D_b by replacing d_i in the original alternative with b and D_c by replacing d_i with c. A betterness judgement (or preference) is separable just in case for any manoeuvre like the one just described, A_b is better than (or preferred to) A_c if and only if D_b is better than (or preferred to) D_c.

Decision theorists typically start with an ordering of alternatives, or a set of properties of orderings, and then show what kind of value functions can represent such an ordering (or an ordering with those properties). But it can be useful to start with a property of a valuation and see what ordering properties it implies. Let us suppose that when a non-modal consequentialist orders a set of alternatives according to 'betterness', she first finds out the moral value of each alternative, in accordance with the two principles of NMC, and then orders the alternatives according to moral value. (Moreover, let us restrict our attention to alternatives where both probabilistic and causal independence holds between the alternatives and the states of the world.) Then since moral value, according to her,

[7] A 'betterness judgement', as I am using the term, is an overall comparative judgement. Hence, it is (formally) very much like preference. But to emphasise that the judgement in question may be objective (if some version of moral realism is true) I will often talk about betterness judgments rather than preferences. When talking about the requirements of decision in general, I will however talk about preferences.

satisfies the second principle of NMC, her betterness judgement satisfies Separability and Actualism.

One might wonder whether the distinction in consequentialist ethics that I have been making is really important. Surely *everyone* accepts that modal properties are morally important, someone might say. So what is the point of discussing this distinction between modal and non-modal consequentialism?[8] On a closer look, it is apparent that not everyone accepts the import of modal properties. According to classical utilitarianism, for instance, one should always choose the act that maximises the total amount of pleasure over pain: 'the greatest happiness for the greatest number', as it is often put. Of course, people might feel (psychological) pleasure and pain due to what could have been. But contrary to what e.g. the FCV claims, the truth of a counterfactual is in itself of no moral importance according to classical utilitarianism; all that matters is how people feel about their actual situation.

Economists and decision theorists have traditionally also been very reluctant to accept that what could have been matters for the (rational) evaluation of actual outcomes. In a classical defence of the 'Independence Axiom', found in one form or other in most decision theories (and implied by Separability as previously defined), Nobel laureate Paul Samuelson argues for formal Separability as captured by the second principle of NMC, based on an intuition like the one expressed by the first principle of NMC. In chapter 4, I explained Samuelson's reasoning (and cited Broome's approval of it). The upshot is that Samuelson and Broome think that because there should be no desirabilistic dependencies between mutually incompatible outcomes (as the first principle of NMC states), an evaluation or ordering of alternatives should satisfy Separability (as the second principle of NMC states).

A similar attitude is suggested by Leonard Savage's reaction to the Allais Paradox. Rather than explaining away the paradox by making attitudes to non-modal properties of gambles part of the description of the outcomes, as many people have done since (albeit usually by calling them something like 'global' rather than 'modal' properties of gambles), he insisted that the common 'Allais preference' is simply irrational ([Savage, 1972]: 102-103). But that arguably implies that it is irrational to care about the relationship between actual outcomes and what could have been (as I have already argued). Hence, since Savage wanted to formulate a theory of *rational* choice, he must have been taking himself to be formulating a theory that was non-modal.[9]

Finally, it might be worth mentioning that certain consequentialist theories should be classified as 'modal' for reasons that have nothing to do with what they say about the fairness of lotteries. These include John C. Harsanyi's [Harsanyi, 1977] and Richard Arneson's [Arneson, 1990] views. Both authors claim (roughly) that e.g. social policies should maximally satisfy people's *hypothetical* preferences, i.e. those preferences that people *would* have in ideal circumstances. What is true in a counterfactual world therefore makes a difference to the moral value of outcomes and alternatives in the actual world. Although I will focus on the fairness of lotteries in this chapter, it should become clear that the framework developed in chapter 4 (and discuss again in section 6.6) also provides a formal model in which to state and explore claims made by theories like Harsanyi's and Arneson's.

[8] I thank an anonymous referee for *Economics and Philosophy* for pressing me on this issue.

[9] Peter Hammond also famously defines consequentialism in a way that, in effect, makes what I am calling 'non-modal consequentialism' a non-consequentialist theory (see e.g. [Hammond, 1987], [Hammond, 1988]).

6.3 Non-Modal Consequentialism vs. Fair Chance

I will focus on Savage's version of decision theory to show why a non-modal consequentialist theory is incompatible with the Fair Chance View.[10] Recall that according to Savage's theory, the value of an alternative A, denoted by $U(A)$, is given by:[11]

$$EU(L) = \sum_{s_i \epsilon \mathbf{S}} u(A(s_i)).P(s_i)$$

where $\mathbf{S}$ is a partition of the possible states, $A(s_i)$ is the consequence of A if s_i happens to be the actual state of the world, u a utility measure on (maximally specific) consequences, and P a probability measure on states.

Savage's equation satisfies the second principle of non-modal consequentialism, i.e. the Separability property. In decision theoretic jargon, Savage's utility function is *additively separable*: the value of each alternative is a weighted *sum* of the values of each of its possible consequences. Hence, the value that $u(s_i(A))$ contributes towards the overall value of A is independent of any $s_j(A)$. But, crucially, for Savage's theory to be an appropriate formalisation of non-modal consequentialism, we need to assume that each $s_i(A)$ is a *non-modal* consequence.

To relate the above characterisation of non-modal consequentialism back to the example of Ann, Bob and the kidney, let L be a lottery that gives Ann and Bob an equal chance of receiving the kidney (depending for instance on whether a fair coin lands heads up or tails up), and A (B) the alternative of giving the kidney to Ann (Bob) without holding a lottery. Then the Fair Chance View implies the following betterness judgement, which I will refer to as the *Fair Chance Judgement* (FCJ): $A \prec L$, $B \prec L$ (where we now interpret '.. $\prec$ ...' as '... is worse than ...'). Then given that Savage's theory is an appropriate formalisation of NMC, the latter is only compatible with the FCV if the following inequalities can simultaneously be satisfied:

$$\sum_{s_i \in \mathcal{S}} u(s_i(A)).P(s_i) < \sum_{s_i \in \mathcal{S}} u(s_i(L)).P(s_i) \tag{6.1}$$

$$\sum_{s_i \in \mathcal{S}} u(s_i(B)).P(s_i) < \sum_{s_i \in \mathcal{S}} u(s_i(L)).P(s_i) \tag{6.2}$$

where each $s_i(\alpha)$ is a *non-modal* consequence.

Now let ANN (BOB) represent the *consequence* where Ann (Bob) receives the kidney. Then A (B) is certain to have ANN (BOB) as consequence, but L will either result in ANN or BOB. Thus we can represent A (B) by ANN (BOB), and L by the n-tuple $\langle ANN, BOB \rangle$. There do not seem to be any modal properties built into the description of these consequences, so non-modal consequentialists should be happy with this representation of the three alternatives. But now the non-modal consequentialist runs into trouble, as we saw in chapter 4. For according to NMC, the value of the three alternatives is then given by: $U(A) = u(ANN)$, $U(B) = u(BOB)$ and $U(L) = u(ANN)0.5 + u(BOB)0.5$. But obviously, $u(ANN)0.5 + u(BOB)0.5$ can never be greater than $u(ANN)$ and *also* greater than $u(BOB)$.

The above tension clearly has something to do with Separability; in particular, the

[10] As I hope will become evident, my argument does not depend on any features special to Savage's version of decision theory. In particular, dropping the assumption of Savage's that states of the world are causally and/or probabilistically independent of alternatives (or *acts*, as Savage calls them) does not affect the argument.

[11] I assume here that the possible consequences of each alternative are at most countably infinite. Otherwise we would take the integral rather than the sum.

additively separable form of Savage's utility function. If the above is the right description of the alternatives, then if the value of an alternative is a probability weighted sum of the values of its possible consequences, then the value of L can never be greater than the value of both A and B. In next section I discuss an attempt to make the FCV compatible with consequentialism by dropping Separability, thus violating the second principle of NMC. But the above tension can also be attributed to the way in which the alternatives have been described. In sections 4 and 5 I discuss two attempts – one old and one novel – to make the FCV compatible with consequentialism by describing the alternatives (and their consequences) in a way that violates the first principle of NMC. (The new attempt also violates the second principle of NMC, as we have seen.) Perhaps unsurprisingly, I will argue that only my new solution – the one that I discussed in chapter 4 – succeeds in making consequentialism compatible with the intuition behind the FCV.

6.4 Consequentialism Without Separability

Suppose we calculate the values of the three alternatives as follows:

$$U(A) = u(ANN).r(P(\top)),$$
$$U(B) = u(BOB).r(P(\top)),$$
$$U(L) = u(ANN).r(P(S)) + u(BOB).r(P(\neg S))$$

where $\top$ is a tautology and r a *risk function* of a *risk seeking* agent – i.e. an agent who prefers a gamble with an expected value of x to a risk-free alternative the value of whose consequence is x – and $\{S, \neg S\}$ is a partition of the possible states of the world into two equiprobable events.[12] Then it might be possible to represent the judgment that L is better than both A and B as maximising risk-weighted utility; if, for instance, we allow for the possibility that $u(ANN).r(P(S)) + u(BOB).r(P(\neg S))$ is greater than *both* $u(ANN).r(P(\top))$ and $u(BOB).r(P(\top))$. And given how I have characterised consequentialism, i.e. as the view that the value of an alternative is determined by how it distributes consequences across states of the world, risk-weighted utility theory is a consequentialist theory.

This solution satisfies the first principle of NMC. For the consequences of the lottery, thus described (*ANN* and *BOB*), do not contain modal properties. But this solution is still incompatible with non-modal consequentialism, since it violates the second principle of NMC. To see this, notice that alternative A can be reformulated as the 'lottery' that has *ANN* as consequence in both the S-states and $\neg S$-states. Hence, for the risk-weighted solution to work, r has to allow for the possibility that $u(ANN).r(P(S)) + u(BOB).r(P(\neg S))$ is greater than both $u(ANN).r(P(S)) + u(ANN).r(P(\neg S))$ and $u(BOB).r(P(S)) + u(BOB).r(P(\neg S))$. But for that to be possible, we need to assume that the function r is sensitive to *global* features of alternatives; i.e. that r is such that the contribution that $u(ANN)$ makes towards the overall value of an alternative partly depends on what other consequences the alternative can have. Hence, this solution violates the second principle of NMC.

Should *modal* consequentialists be happy with this solution? There are, in my view, two related reasons for seeking an alternative way of making consequentialism compatible with the FCV. Firstly, it seems to me that the Fair Chance View is an example of a more general phenomenon where what could have been is important for the evaluation of the

[12] This could be seen as a variant of Lara Buchak's *Risk-Weighted Expected Utility Theory* [Buchak, 2013], albeit with some important differences. (For instance, the way in which she defines r means that it is always the case that $r(P(\top)) = P(\top) = 1$, and that the expected utility of a lottery can never exceed the utilities of all of its possible prizes, contrary to what I am assuming here.)

desirability of what actually occurs (as I have already explained). Unlike the above solution, the one I developed in chapter 4 (and discuss again in section 6.6) explicitly models this relationship between what is and what could have been. Secondly, the above solution suggests that accepting the FCV has something to do with being risk seeking. More precisely, this way of making consequentialism compatible with the FCV suggests that the reason consequentialism as formalised by Savage seems incompatible with the FCV, is that that formalisation places certain restrictions on attitudes towards risk. But those who accept my first objection will agree that the problem with Savage's framework, from the perspective of the FCV, is not so much the theory's restriction on risk attitudes, but rather its insensitivity to the desirabilistic relationships between what is and what could have been.

6.5 Broome's Redescription Strategy

Contrary to my suggestion in section 6.3, many people will undoubtedly have the intuition that the consequence where Ann (Bob) receives the kidney as a result of the lottery L is *not* the same consequence as Ann (Bob) receiving the kidney as a result of the risk-free alternative A (B). The former consequence is *fair* whereas the latter is not, which must mean that these are not the same consequences. Hence, it seems, the consequences of A, B and L were not properly described in last section. Similarly to what Broome ([Broome, 1991]: ch. 5) suggests, we should perhaps write the fairness directly into the description of the outcomes of the lottery; such that, for instance, L has *ANN*&*Fair*, *BOB*&*Fair* as its two possible consequences, but A (B) has *ANN* (*BOB*) as the only possible consequence. And then the trouble we saw in section 6.3 disappears, since there are many (additively separable) functions EU that simultaneously satisfy:

$$EU(A) = u(ANN) < EU(L) = u(ANN\&Fair).0.5 + u(BOB\&Fair).0.5 \tag{6.3}$$

$$EU(B) = u(BOB) < EU(L) = u(ANN\&Fair).0.5 + u(BOB\&Fair).0.5 \tag{6.4}$$

Broome's solution thus makes a preference for tossing the coin in the example discussed at that start of the paper compatible with consequentialism without giving up Separability (i.e. without violating the second principle of NMC). But the solution clearly violates the first principle of NMC. Assuming that the FCV is part of our conception of fairness, then when we start refining our consequence set to include consequences that have 'fairness' written into their description, for examples like the one I have been discussing, we are in effect creating dependencies between consequences in different mutually incompatible states of the world. The consequence *ANN*&*Fair* for instance implicitly refers to what could have been, in the sense that a *necessary* condition for *ANN*&*Fair* to be a possible consequence of some alternative C, is that *BOB*&*Fair* is *also* (at least considered to be) a possible consequence of C.[13] For given the FCV, C can only result in *ANN*&*Fair* if C is some sort of lottery or random choice mechanism that has *BOB*&*Fair* as consequence in some state of the world. Hence, given that the moral value of the consequence of the lottery is, on Broome's suggestion, partly a function of this modal property, his suggestion violates the first principle of NMC.[14]

[13] That is, assuming that C is the initial choice of who should receive the kidney, rather than for instance the act of giving the kidney to Ann after she has won it in a lottery.

[14] I should emphasise that violating the first principle of NMC does not necessarily entail a departure from orthodox decision theory. (I thank Jim Joyce for pressing me on this issue.) Many economists and philosophers with welfarist inclinations seem to typically adhere to the first principle of NMC in their *application* of expected utility theory; i.e., in the way in which they conceptualise consequences. But formally, expected utility theory does

The implicit reference to what could have been is precisely the reason Broome's preferred description of the consequences runs into troubles with (what he calls) the *Rectangular Field Assumption* (RFA). RFA is a technical assumption of many of the traditional decision theoretic representation theorems, such as Savage's, needed (given the other assumptions of these theorems) to construct from an agent's preferences a value function that is unique up to positive affine transformation. Recall that an alternative can be represented by an n-tuple of consequences, e.g. $A = \langle a_1, a_2, ..., a_n \rangle$, where a_i is the consequence of (choosing) alternative A if state of the world s_i happens to be actual. Given any set of alternatives that an agent's preferences are defined over, each state has associated with it a set of possible outcomes. Call that set for the i-th state C_i. Then take the product of all of these sets: $C_1 \times C_2 \times ... \times C_n$. Now the RFA states that *any* n-tuple created by picking consequences from some or all of the sets in the product is an alternative in the agent's preference ordering. That is, if we go through the sets, from C_1 to C_n, and pick arbitrary consequences from some or all of these sets, then the resulting n-tuple of consequences is an alternative in the agent's preference ordering.

Going back to our example, we could for instance pick two consequences: *ANN* and *ANN*&*Fair*. The resulting alternative would then be the ordered pair $\langle ANN, ANN\&Fair \rangle$. This is an alternative where Ann gets the kidney in any state of the world; but, in addition, if some state turns out to be actual, then Ann gets the kidney fairly! But that is, of course, conceptually impossible on the Fair Chance View. For given this conception of fairness, either Ann or Bob can only receive the kidney fairly if some random mechanism was used to determine who was to receive the good. But whenever such a random choice mechanism is used, it will *not* be the case that the same patient receives the kidney in all states of the world. In other words, an outcome is fair only if it is not true that that same patient receives the kidney in all states of the world. Another way to put this, is that given Broome's description of consequences, the RFA requires that it be possible that a lottery that is unfair results in a consequence that is fair. It seems clear that this requirement goes again the intuition behind the FCV. Hence, it is not at all clear that Broome's 're-description strategy' makes the FCV compatible with Savage's consequentialist framework.[15]

I should emphasise that the tension with the Rectangular Field Assumption is not the main reason I think we should seek alternative ways of making the Fair Chance View compatible with consequentialism. Not all decision theoretic representation theorems require the RFA,[16] and there are well known problems with the assumption that have nothing to do with fairness. The main problem I have with Broome's solutions, is rather that unlike the solution suggested in next section, Broome's fails to make explicit the desirabilistic dependency between actual and counterfactual outcomes that seems at the heart of the FCV.

not require adherence to the principle. (As we will se below, however, violation of the first principle of NMC is in tension with – but not, strictly speaking, inconsistent with – a technical assumption of *some* expected utility theories.)

[15]It may be worth stating the RFA in the terminology of Savage's framework. Here, the assumption is that any function from the state space $\mathcal{S}$ we are working with to the consequence set C, is an *act* in the agent's preference ordering. Assume that H and T (standing for e.g. *coin comes up heads* and *coin comes comes up tails*) are two events that partition the state space. Then if C contains both *ANN* and *ANN*&*Fair*, the function

$$f^*(s_i) = \begin{cases} ANN & \text{if } s_i \in T, \\ ANN\&Fair & \text{if } s_i \in H. \end{cases}$$

should be an act in the agent's preference ordering. This is an act that has the consequence that Ann receives the kidney in any state of the world, and moreover receives it fairly if a state if H happens to be actual.

[16]In particular, the Bolker-Jeffrey theorem for Jeffrey's decision theory does not contain the Rectangular Field Assumption. Broome's solution can easily be reformulated for Jeffrey's framework. Moreover, the representation theorem for expected utility theory that we prove in [Bradley and Stefánsson, 2015] does not rely on the assumption.

Examining the tension between Broome's solution and the RFA nevertheless does serve an important role in the present argument, since it illustrates the difficulty in dropping the first principle of non-modal consequentialism while holding on to Savage's framework. The reason the RFA creates trouble for Broome is that some consequences as Broome describes them refer, as we have seen, to what occurs in states of the world in which they themselves do not occur, and thus cannot be combined with any arbitrary consequence as the RFA requires. In other words, the tension with the RFA stems from Broome's violation of the first principle of NMC.

6.6 A New Version of Modal Consequentialism

I now finally turn to formulating a version of modal consequentialism that satisfies the intuition behind the Fair Chance View better than the theories already discussed. The framework is the same as the one I developed in chapter 4, but here I will focus on its application to the notion of fair chance.

Recall that according to Jeffrey's theory, the desirability of any particular proposition, A, is a weighted average of the desirabilities of the different (mutually exclusive) ways in which the proposition can be true, where the weight on each way w_i that proposition A can be true is given by $P(w_i \mid \mathrm{A})$. Hence, as we have already seen, one way of formulating Jeffrey's desirability measure is as follows:

$$Des(p) = \sum_{w_i \in \mathrm{A}} P(w_i \mid \mathrm{A}).Des(w_i) \tag{6.5}$$

Jeffrey's evaluation of propositions is clearly consequentialist, given how I have characterised consequentialism. The possible ways in which a proposition can come true can be understood as the possible *consequences* of the proposition coming true, and we can interpret that which determines the way in which a proposition comes true (if it comes true) as a 'state' of our world. Representing the Fair Chance Judgement as maximising the value of a Jeffrey-desirability function is, therefore, one way of showing that the Fair Chance View is compatible with consequentialism. However, someone who evaluates propositions according to Jeffrey's equation will not always satisfy Separability (as I explained in chapter 4).

Although Jeffrey's theory violates Separability, the theory as Jeffrey himself interpreted it – i.e. as a version of *non-modal* consequentialism (where *Des* is defined over only factual propositions) – nevertheless runs into the same problem we have seen with Savage's: there is no pair of desirability and probability functions relative to which the Fair Chance Judgement can be represented as maximising desirability. Recall however that I mentioned in chapter 4 that it could be argued that the consequences of the lottery should be formulated as $ANN \wedge \mathrm{L}$, in which case there are pairs of desirability and probability functions relative to which the Fair Chance Judgement can be represented as maximising desirability. But $ANN \wedge \mathrm{L}$ is not a non-modal consequence: in effect, this description of the consequence has built into it that the consequence in question had 0.5 chance of occurring (since L is the proposition that either *ANN* or *BOB* will occur will equal chance).

Making Jeffrey's theory compatible with the Fair Chance Judgement by including the lottery in the description of the consequences moreover suffers, in my view, from the same problem as Broome's suggestion, namely that it does not make explicit the importance of

counterfactuals for the Fair Chance View. As we saw in chapter 4, we can, however, represent the FCJ (which I then called Diamond's preference) as maximising desirability once we have extended Jeffrey's decision theory to counterfactuals. And representing the judgement with that framework does make explicit the importance of counterfactuals. I will not go through all the details again. However, I will provide a slightly different argument from the one presented in chapter 4, to show that the extended theory can represent the FCJ. The reason for going through this again is that I will now show that the representation satisfies certain moral intuitions that I ignored in chapter 4.

Let us now focus on a model with only four worlds, w_1 to w_4, and suppose this is how worlds and sentence match up:

	B	$\neg B$
A	w_1	w_2
$\neg A$	w_3	w_4

Now let A express the proposition (A) that the coin comes up heads and $\neg A$ the proposition that the coin comes up tails. Let B express the proposition (B) that Ann receives the kidney and $\neg B$ the proposition that Bob receives the kidney. (Recall from chapter 4, page 87, the simplifying assumptions already made.)

To represent the Fair Chance Judgement in a multidimensional model, we need to work with two suppositions: the supposition that A and the supposition that $\neg A$. But actually, we only need to consider two dimensions at a time, since the FCJ only orders alternatives according to what is actually true and what would be true under a *contrary-to-factual* supposition. That is, the judgement for instance is that a situation where the coin comes up heads and Ann receives the kidney, is made better or worse depending on whether Ann also receives the kidney under the (contrary-to-factual) supposition that the coin comes up tails. But it says nothing about whether the desirability of this situation depends on what is true under the (matter-of-factual) 'supposition' that the coin comes up heads. Hence, for a situation where A is true at the actual world, we need not represent the 'counter'-world under the supposition that A; and similarly for the situation where $\neg A$ is true. So although semantically we are now working with a three-dimensional model, we only need to worry about two dimensions at a time.

Let θ denote the world we need not consider at each time. If A is actually true, then θ is the 'counter'-world under the supposition that A (which is the second world as I am setting up the n-tuples) but if $\neg A$ is actually true, then θ is the 'counter'-world under the supposition that $\neg P$ (which is the third world as I am setting up the n-tuples). The following is thus the space of possible situations[17] we need (call this whole space of situations $\mathcal{W}$):

Actuality	Possible Situations	
w_1	$\langle w_1, \theta, w_3 \rangle$	$\langle w_1, \theta, w_4 \rangle$
w_2	$\langle w_2, \theta, w_3 \rangle$	$\langle w_2, \theta, w_4 \rangle$
w_3	$\langle w_3, w_1, \theta \rangle$	$\langle w_3, w_2, \theta \rangle$
w_4	$\langle w_4, w_1, \theta \rangle$	$\langle w_4, w_2, \theta \rangle$

I will assume that it makes no difference, according to the FCV, whether Ann or Bob actually receives the kidney, provided that one of them does, since by assumption their situation is symmetric in all ways that are relevant for the decision of who should receive the kidney.

[17] To emphasise to consequentialist nature of the theory, I will in this section call each n-tuple a *situation*.

Hence, the fair situations are equally good when Ann actually receives the kidney as when Bob actually receives the kidney and similarly for the unfair situations. The FCJ thus orders the situations into two equivalence classes, where every situation from the 'Good' class is better than every situation from the 'Bad' class; but any two situations within the same class are equally good:

Bad	Good
$\langle w_1, \theta, w_3\rangle$	$\langle w_1, \theta, w_4\rangle$
$\langle w_2, \theta, w_4\rangle$	$\langle w_2, \theta, w_3\rangle$
$\langle w_3, w_1, \theta\rangle$	$\langle w_3, w_2, \theta\rangle$
$\langle w_4, w_2, \theta\rangle$	$\langle w_4, w_1, \theta\rangle$

What is common to the ('bad') situations in the left column is that the person who actually receives the kidney would also have received it had the coin come up differently. In the situations in the right column, however, whoever actually receives the kidney would not have received it had the coin landed differently.

The Fair Chance Judgement can now be formulated as follows:[18]

$$\begin{aligned}\langle w_1, \theta, w_3\rangle \sim \langle w_2, \theta, w_4\rangle \sim \langle w_3, w_1, \theta\rangle \sim \langle w_4, w_2, \theta\rangle \\ \prec \langle w_1, \theta, w_4\rangle \sim \langle w_2, \theta, w_3\rangle \sim \langle w_3, w_2, \theta\rangle \sim \langle w_4, w_1, \theta\rangle \quad (6.7)\end{aligned}$$

and the FCJ can be represented by a function V that satisfies:

$$\begin{aligned}V(\langle w_1, \theta, w_3\rangle) = V(\langle w_2, \theta, w_4\rangle) = V(\langle w_3, w_1, \theta\rangle) = V(\langle w_4, w_2, \theta\rangle) \\ < V(\langle w_1, \theta, w_4\rangle) = V(\langle w_2, \theta, w_3\rangle) = V(\langle w_3, w_2, \theta\rangle) = V(\langle w_4, w_1, \theta\rangle) \quad (6.8)\end{aligned}$$

There will certainly be many functions satisfying 6.8 (and thus representing 6.7): any *ordinal* utility function, defined over a set of world-triples, can represent this ordering. So to some extent, we have reached the goal of making the FCV compatible with (modal) consequentialism. For we have found a way of showing that there is a function whose assignment of values to situations corresponds to whether the FCV deems the situation fair. And we do not have to worry about clashes with the Rectangular Field Assumption, since the assumption is not needed for the existence of such a function.

I have however not yet shown that there is a *Jeffrey* desirability function that represents 6.7. But we can do so by construction. Let V and P assign values to the basic situations (i.e. the ordered triples) in $\mathcal{W}$. The functions extend to any *proposition* α, i.e. to any set of situations, according to the following rules (recall that α could be factual or conditional):

$$V(\alpha) = \sum_{\langle w_i, w_j, w_k\rangle \in \alpha} V(\langle w_i, w_j, w_k\rangle).P(\langle w_i, w_j, w_k\rangle \mid \alpha) \quad (6.9)$$

[18] Each triple in 6.7 is a (maximally specific) proposition, and in fact a conjunction of a factual and a counterfactual proposition. $\langle w_4, w_2\rangle$ for instance is the proposition that $\neg A \wedge \neg B \wedge A \mathrel{\Box\!\!\rightarrow} \neg B$, $\langle w_4, w_1\rangle$ the proposition that $\neg A \wedge \neg B \wedge A \mathrel{\Box\!\!\rightarrow} B$, etc. Hence, 6.7 is equivalent to:

$$\begin{aligned}(A \wedge B \wedge \neg A \mathrel{\Box\!\!\rightarrow} B) \sim (A \wedge \neg B \wedge \neg A \mathrel{\Box\!\!\rightarrow} \neg B) \sim (\neg A \wedge B \wedge A \mathrel{\Box\!\!\rightarrow} B) \sim (\neg A \wedge \neg B \wedge A \mathrel{\Box\!\!\rightarrow} \neg B) \\ \prec [(A \wedge B \wedge \neg A \mathrel{\Box\!\!\rightarrow} \neg B) \sim (A \wedge \neg B \wedge \neg A \mathrel{\Box\!\!\rightarrow} B) \sim (\neg A \wedge B \wedge A \mathrel{\Box\!\!\rightarrow} \neg B) \sim (\neg A \wedge \neg B \wedge A \mathrel{\Box\!\!\rightarrow} B)] \quad (6.6)\end{aligned}$$

$$P(\alpha) = \sum_{\langle w_i, w_j, w_k \rangle \in \alpha} P(\langle w_i, w_j, w_k \rangle) \tag{6.10}$$

Then by construction, V is a Jeffrey desirability function: the desirability of a proposition, according to this function, is a weighted average of the desirabilities of the different ways in which the propositions can come true (i.e. the different situations compatible with the proposition), where the weights are given by the appropriate conditional probabilities.

Now let's see whether this function can represent the Fair Chance Judgement, as formulated in 6.7. Recall the two equivalence classes of propositions (sets of n-tuples) induced by the FCJ. Let us call the 'good' equivalence class G and the 'bad' equivalence $\neg G$. We can stipulate that:

1. $\forall \langle w_i, w_j, w_k \rangle \in G: \quad V(\langle w_i, w_j, w_k \rangle) = 1$
2. $\forall \langle w_l, w_m, w_n \rangle \in \neg G: \quad V(\langle w_l, w_m, w_n \rangle) = -1$

Then it is clear that V represents the FCJ, as formulated in 6.7: for any two basic situations, $\langle w_i, w_j, w_k \rangle$ and $\langle w_l, w_m, w_n \rangle$, we have: if $\langle w_i, w_j, w_k \rangle \sim \langle w_l, w_m, w_n \rangle$ according to FCJ, then $V(\langle w_i, w_j, w_k \rangle)$ and $V(\langle w_l, w_m, w_n \rangle)$ are both either -1 or 1; but if $\langle w_i, w_j, w_k \rangle \prec \langle w_l, w_m, w_n \rangle$ according to FCJ, then $V(\langle w_i, w_j, w_k \rangle) = -1 \langle V(\langle w_l, w_m, w_n \rangle) = 1$. So we have constructed a Jeffrey desirability function that represents the FCJ.

G and $\neg G$ are also propositions (sets of n-tuples of worlds), and by the above stipulation: $V(G) = 1$ and $V(\neg G) = -1$. But for any arbitrary proposition α:

$$V(\alpha) = V(G).P(G \mid \alpha) + V(\neg G).P(\neg G \mid \alpha) = P(G \mid \alpha) - P(\neg G \mid \alpha) \tag{6.11}$$

Up to now I have been focusing only on propositions that are either completely fair or completely unfair; i.e. propositions that are either subsets of G or $\neg G$ but do not overlap the two sets. And I have formulated the FCJ as only having something to say about propositions that are either completely fair or unfair. In that respect, what I have done so far in this chapter adds nothing to what was already accomplished in chapter 4. Notice, however, that we can easily construct propositions that overlap the two sets G and $\neg G$. Let's call such propositions 'mixed'. $m = \{\langle w_1, \theta, w_3 \rangle, \langle w_1, \theta, w_4 \rangle\}$ is a mixed proposition, for instance, since the first of its elements is an unfair situation but the second is fair. Intuitively, a proposition like m is a *biased* lottery. In $\langle w_1, \theta, w_3 \rangle$ Ann gets the kidney no matter how the coin lands. So since this situation is possible given m, but no situation that gives Bob a greater chance than Ann is possible given m, m is biased towards Ann.

Now take any two mixed propositions m_1 and m_2: suppose each is a set of two triples, one of which is an element of G but the other of $\neg G$. According to V, as I have constructed it, $V(m_1) \leq V(m_2)$ just in case $P(G \mid m_1) \leq P(G \mid m_2)$. It is natural to think of $P(G \mid m_i)$ as measuring how unbiased m_i is: if $P(G \mid m_i) = 1$ then m_i is not at all biased but if $P(G \mid m_i) = 0$ then m_i is maximally biased. So the more (less) biased a proposition is, the lower (higher) value V assigns to it. Although I have said nothing about how the Fair Chance View judges biased lotteries, this is exactly what we should want.[19] For any mixed propositions m_1 and m_2, the former should be better than the latter, on this view, if and only if it is less biased; but if they are equally biased, then they should be equally good (or bad). So the function I have constructed does not only capture the intuition that a situation where Ann (Bob) receives the kidney is fair only if Bob (Ann) had a chance; it also captures the intuition that situations

[19] In fact, this needs to hold for the FCJ to satisfy continuity.

where they both had an *equal* chance at receiving the kidney are more fair than situations where their chances were unequal.

6.7 Implications of the New Solution

The framework developed in chapter 4, and again in last section, has the advantage over the other modal consequentialist theories I have been discussing that it explicitly models the desirabilistic relationship between what is and what could have been, which seems at the heart of the Fair Chance View. In addition, it has the following advantage over Broome's suggestion for how to make consequentialism compatible with the preference for tossing the coin. Broome's solution consists in adding a primitive fairness property to the outcome space. My suggested solution, however, consists in enriching decision theory to include counterfactual prospects in such a way that the fairness property *emerges* as a relationship between actual and counterfactual outcomes. Thus, I contend, my solution better explains what orthodox decision theory lacks when it comes to capturing common intuitions about fair distribution of chances, such as the intuition underlying the FCV.

Broome might of course complain that I have myself added some primitive variable to decision theory, namely the (counterfactual) supposition operator. But those who are already motivated by the Fair Chance View, or more generally recognise that what could have been is often important for the evaluation of actual outcomes, hopefully agree that this extra complexity is more than offset by the benefit of being able to formally represent the desirabilistic dependency between facts and counterfactuals.

Using the multidimensional framework to represent the Fair Chance Judgement has the interesting implication that the extra value generated by the truth of the relevant counterfactual does not supervene on the non-modal facts. The table representing the 'goodness partition', for instance, has each actual world in both the 'Good' and the 'Bad' column. The situations $\langle w_1, \theta, w_3 \rangle$ and $\langle w_1, \theta, w_4 \rangle$ for instance share all non-modal facts and differ only in what *would* be true if $\neg A$ were. But the latter is fair whereas the former is not, which suggests that fairness does not supervene on non-modal facts. More generally, if the multidimensional semantics, formulated as Bradley suggests (i.e. without the added supervenience constraints that I discussed in chapter 3), is the correct semantics for counterfactuals, then counterfactuals don't supervene on non-modal facts.[20] But then if the moral value of a situation partly depends on what counterfactuals are true, as the FCV states, then the moral value of a situation does not supervene on its non-modal facts (contrary to what Broome claims [Broome, 1991]: 114-115).

The non-modal consequentialist will without a doubt point out that the failure of fairness to supervene on non-modal facts is merely an artefact of our model. And she might argue as follows. This failure of counterfactuals to supervene on non-modal facts, according to the multidimensional semantics, is a reason for looking for a different semantics for counterfactuals, if we want to insist that fairness is partly determined by what counterfactuals are true. According to the best known semantics for counterfactuals, i.e. the Stalnaker-Lewis semantics ([Stalnaker, 1968], [Lewis, 1986a]), counterfactuals *do* supervene on (and are implied by) factual propositions.[21] Let f_1 be the set of factual propositions that (on this semantics)

[20] The same is true given Hannes Leitgeb's recent semantics for counterfactuals ([Leitgeb, 2012a], [Leitgeb, 2012b]).

[21] Assuming that both the possible worlds that serve as truth makes for counterfactuals and the relevant similarity relation is entirely determined by the facts of the actual world, as Stalnaker and Lewis seem to do.

imply the counterfactual $\neg A \mathrel{\Box\!\!\rightarrow} B$ and f_2 the set of factual propositions that imply the counterfactual $\neg A \mathrel{\Box\!\!\rightarrow} \neg B$. What really explains our judgement that $V(A \wedge B \wedge \neg A \mathrel{\Box\!\!\rightarrow} \neg B)$ can be different from $V(A \wedge B \wedge \neg A \mathrel{\Box\!\!\rightarrow} B)$, the non-modal consequentialist might argue, is the fact that for us, $V(A \wedge B \wedge f_2)$ might be different from $V(A \wedge B \wedge f_1)$. To put the point less abstractly, although the truth of a particular counterfactual is *one* difference between a situation where Ann receives the kidney fairly and one where Ann receives the kidney unfairly, what *really* makes the moral difference is the set of factual propositions that *implies* the relevant counterfactual.

I do not want to argue against the view that counterfactuals and other modal facts supervene on non-modal facts. However, there are various arrangements of non-modal facts that can make true the particular modal facts we are interested in. For instance, there are various ways of making it true that Ann and Bob have an equal chance of receiving the kidney. Many of these arrangements of non-modal facts are equally good, from a moral perspective. And what makes them morally good, is the fact that they all entail the relevant modal fact; in the case we are considering, the fact that Ann and Bob have an equal chance. In other words, the *reason* all these different arrangements of non-modal facts are morally good is that they entail this particular modal fact.

Moreover, and more generally, we should make a distinction between facts that *carry* value and facts on which the carriers of value supervene.[22] Even if it is true, as Humeans claim (and I tend to agree), that all facts supervene on the non-modal facts, that does not, by itself, mean that these non-modal facts are the carriers of value. Perhaps the following analogy will help. Every intrinsic (as opposed to relational) property of a painting is determined by how the atoms that make up the painting are arranged. Hence, the aesthetic qualities of the painting supervene on this arrangement of atoms. These aesthetic qualities, most people think, carry some value, over and above that which is carried by the atoms that make up the painting. Similarly, whether or not Bob could have received the kidney may be determined by the non-modal facts of our world. But that does not mean that this particular counterfactual carries no value over and above these non-modal facts.

6.8 Concluding Remarks

Decision theorists and economists have historically ignored the role counterfactuals often play when intuitively rational people evaluate the desirability actual outcomes or states of affairs. As I have explained in this thesis, this has made their decision theory inconsistent with seemingly rational preferences. Moreover, I hope to have shown that given a plausible theory of counterfactuals, we can extend Richard Jeffrey's theory to counterfactual prospects, in a way that makes it possible to accommodate the aforementioned preferences.

[22] I thank Wlodek Rabinowicz for suggesting this wording.

Bibliography

[Adams, 1975] Adams, E. W. (1975). *The Logic of Conditionals*. D. Reidel Publishing Company.

[Allais, 1953] Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomesde l'ecole Americaine. *Econometrica*, 21(4):503–546.

[Allais, 1979] Allais, M. (1979). The foundations of a positive theory of choice involving risk and a criticism of the postulates and axioms of the american school. In Allais, M. and Hagen, O., editors, *Expected Utility Theory and the Allais Paradox: Contemporary Discussions of Decisions under Uncertainty with Allais' Rejoinder*.

[Arneson, 1990] Arneson, R. J. (1990). Liberalism, distributive subjectivism, and equal opportunity for welfare. *Philosophy and Public Affairs*, 19(2):158–194.

[Bennett, 2003] Bennett, J. (2003). *A Philosophical Guide to Conditionals*. Clarendon Press.

[Bolker, 1966] Bolker, E. D. (1966). Functions resembling quotients of measures. *Transactions of the American Mathematical Society*, 124(2):292–312.

[Bovens and Rabinowicz, ms] Bovens, L. and Rabinowicz, W. (ms.). The anatomy of regret: Taxonomy and measurement. Unpublished manuscript.

[Bradley, 1998] Bradley, R. (1998). A representation theorem for a decision theory with conditionals. *Synthese*, 116(2):187–229.

[Bradley, 1999] Bradley, R. (1999). Conditional desirability. *Theory and Decision*, 47(1):23–55.

[Bradley, 2000] Bradley, R. (2000). A preservation condition for conditionals. *Analysis*, 60(3):219–222.

[Bradley, 2002] Bradley, R. (2002). Indicative conditionals. *Erkenntnis*, 56(3):345–378.

[Bradley, 2007a] Bradley, R. (2007a). A defence of the ramsey test. *Mind*, 116(461):1–21.

[Bradley, 2007b] Bradley, R. (2007b). A unified Bayesian decision theory. *Theory and Decision*, 63(3):233–263.

[Bradley, 2011] Bradley, R. (2011). Conditionals and supposition-based reasoning. *Topoi*, 30(1):39–45.

[Bradley, 2012] Bradley, R. (2012). Multidimensional possible-world semantics for conditionals. *Philosophical Review*, 121(4):539–571.

[Bradley, ms] Bradley, R. (ms). *Decision Theory with a Human Face*.

[Bradley and List, 2009] Bradley, R. and List, C. (2009). Desire-as-belief revisited. *Analysis*, 69(1):31–37.

[Bradley and Stefánsson, 2015] Bradley, R. and Stefánsson, H. O. (2015). Counterfactual desirability. *British Journal for the Philosophy of Science*.

[Broome, 1991] Broome, J. (1991). *Weighing Goods*. Basil Blackwell.

[Buchak, 2013] Buchak, L. (2013). *Risk and Rationality*. Oxford University Press.

[Cartwright, 1979] Cartwright, N. (1979). Causal laws and effective strategies. *Noûs*, 12(4):419–437.

[Chalmers and Hájek, 2007] Chalmers, D. J. and Hájek, A. (2007). Ramsey + Moore = God. *Analysis*, 67(2):170–172.

[Cross, 2009] Cross, C. B. (2009). The problem of counterfactual conditionals. *Erkenntnis*, 70(2):173–188.

[Diamond, 1967] Diamond, P. (1967). Cardinal welfare, individualistic ethics, and interpersonal comparison of utility: Comment. *Journal of Political Economy*, 75(5):765–766.

[Edgington, 1995] Edgington, D. (1995). On conditionals. *Mind*, 104(414):113–128.

[Ferguson, 2013] Ferguson, B. (2013). *The Paradox of Exploitation: A New Solution*. PhD thesis, London School of Economics and Political Science.

[Fitelson and Hájek, ms] Fitelson, B. and Hájek, A. (ms.). Declarations of independence. Unpublished manuscript.

[Foley, 1992] Foley, R. (1992). The epistemology of belief and the epistemology of degrees of belief. *Americal Philosophical Quarterley*, 29(2):111–121.

[Frigg and Hoefer, 2010] Frigg, R. and Hoefer, C. (2010). Determinism and chance from a Humean perspective. In Dennis Dieks, Wenceslao Gonzalez, S. H. M. W. F. S. and Uebel, T., editors, *The Present Situation in the Philosophy of Science*, pages 351–271. Springer.

[Frigg and Hoefer, ta] Frigg, R. and Hoefer, C. (ta.). The best Humean system for statistical mechanics. *Erkenntnis*. Forthcoming.

[Gibbard, 1981] Gibbard, A. (1981). Two recent theories of conditionals. In Harper, W. L., Stalnaker, R., and Pearce, G., editors, *Ifs: Conditionals, Belief, Decision, Chance, and Time*. D. Reidel Publishing Company.

[Gibbard and Harper, 1981] Gibbard, A. and Harper, W. L. (1981). Counterfactuals and two kinds of expected utility theory. In Harper, W. L., Stalnaker, R., and Pearce, G., editors, *Ifs: Conditionals, Belief, Decision, Chance, and Time*. D. Reidel Publishing Company.

[Gundersen, 2004] Gundersen, L. B. (2004). Outline of a new semantics for counterfactuals. *Pacific Philosophical Quarterly*, 85(1):1–20.

[Hájek, 2003] Hájek, A. (2003). What conditional probability could not be. *Synthese 137*, 137(3):273–323.

[Hájek, ms] Hájek, A. (ms). *Most Counterfactuals are False*.

[Hájek and Hall, 1994] Hájek, A. and Hall, N. (1994). The hypothesis of the conditional construal of conditional probability. In *Probability and Conditionals: Belief Revision and Rational Decision*. Cambridge University Press.

[Hájek and Pettit, 2004] Hájek, A. and Pettit, P. (2004). Desire beyond belief. *Australasian Journal of Philosophy*, 82(1):77–92.

[Hammond, 1987] Hammond, P. (1987). Consequentialism and and the Independence axiom. In Munier, B., editor, *Risk, Decision and Rationality*. Springer.

[Hammond, 1988] Hammond, P. (1988). Consequentialist foundations for expected utility theory. *Theory and Decision*, 25(1):25–78.

[Harper, 1981] Harper, W. L. (1981). A sketch of some recent developments in the theory of conditionals. In Harper, W. L., Stalnaker, R., and Pearce, G., editors, *Ifs: Conditionals, Belief, Decision, Chance, and Time*. D. Reidel Publishing Company.

[Harsanyi, 1977] Harsanyi, J. C. (1982/1977). Morality and the theory of rational behaviour. In Sen, A. and Williams, B., editors, *Utilitarianism and Beyond*. Cambridge University Press.

[Hoefer, 2007] Hoefer, C. (2007). The third way on objective proabability: A sceptic's guide to objective chance. *Mind*, 116(463):449–496.

[Hume, 1740] Hume, D. (2000/1739-1740). *A Treatise of Human Nature*. Oxford University Press.

[Jackson, 1991] Jackson, F. (1991). *Conditionals*. Oxford University Press).

[Jeffrey, 1982] Jeffrey, R. (1982). The sure thing principle. *Philosophy of Science*, 2:719–730.

[Jeffrey, 1983] Jeffrey, R. (1990/1983). *The Logic of Decision*. The University of Chicago Press (paperback edition).

[Jeffrey, 1991] Jeffrey, R. (1991). Matter-of-fact conditionals. *Proceedings of the Aristotelian Society, Supplementary Volumes*, 65:161–183.

[Joyce, 1999] Joyce, J. M. (1999). *The Foundations of Causal Decision Theory*. Cambridge University Press.

[Joyce, 2000] Joyce, J. M. (2000). Why we still need the logic of decision. *Philosophy of Science*, 67:S1–S13.

[Joyce, 2002] Joyce, J. M. (2002). Levi on causal decision theory and the possibility of predicting one's own actions. *Philosophical Studies*, 110(1):69–102.

[Kahneman and Tversky, 1979] Kahneman, D. and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–292.

[Kaufmann, 2004] Kaufmann, S. (2004). Conditioning against the grain: Abduction and indicative conditionals. *Journal of Philosophical Logic*, 33(6):583–606.

[Kyburg, 1961] Kyburg, H. E. (1961). *Probability and the Logic of Rational Belief*. Wesleyan University Press.

[Leitgeb, 2012a] Leitgeb, H. (2012a). A probabilistic semantics for counterfactuals. Part A. *The Review of Symbolic Logic*, 5(1):26–84.

[Leitgeb, 2012b] Leitgeb, H. (2012b). A probabilistic semantics for counterfactuals. Part B. *The Review of Symbolic Logic*, 5(1):85–121.

[Levi, 2000] Levi, I. (2000). The Foundations of Causal Decision Theory by James M. Joyce. *The Journal of Philosophy*, 97(7):387–402.

[Lewis, 1976] Lewis, D. (1976). Probabilities of conditionals and conditional probabilities. *Philosophical Review*, 85(3):297–315.

[Lewis, 1980] Lewis, D. (1980). A subjectivist's guide to objective chance. In Jeffrey, R. C., editor, *Studies in Inductive Logic and Probability*. University of California Press.

[Lewis, 1981] Lewis, D. (1981). Causal decision theory. *Australasian Journal of Philosophy*, 59(1):5–30.

[Lewis, 1983] Lewis, D. (1983). New work for a theory of universals. *Australasian Journal of Philosophy*, 61(4):343–377.

[Lewis, 1986a] Lewis, D. (1986a). *Counterfactuals*. Blackwell (revised edition).

[Lewis, 1986b] Lewis, D. (1986b). Probabilities of conditionals and conditional probabilities II. *Philosophical Review*, 95(4):581–589.

[Lewis, 1987] Lewis, D. (1987). Introduction. In *Philosophical Papers, Vol. 2*. Oxford University Press.

[Lewis, 1988] Lewis, D. (1988). Desire as belief. *Mind*, 97(387):323–32.

[Lewis, 1994] Lewis, D. (1994). Humean supervenience debugged. *Mind*, 103(412):473–490.

[Lewis, 1996] Lewis, D. (1996). Desire as belief II. *Mind*, 105(418):303–313.

[Loewer, 2001] Loewer, B. (2001). Determinism and chance. *Studies in History and Philosophy of Science Part B 32*, 32(4):609–620.

[Loomes and Sugden, 1982] Loomes, G. and Sugden, R. (1982). Regret theory: An alternative theory of rational choice under risk. *The Economic Journal*, 92:805–824.

[Makinson, 1965] Makinson, D. C. (1965). The paradox of the preface. *Analysis*, 25(6):205–207.

[McGee, 1985] McGee, V. (1985). A counterexample to modus ponens. *The Journal of Philosophy*, 82(9):462–471.

[McGee, 2000] McGee, V. (2000). To tell the truth about conditionals. *Analysis*, 60(1):107–111.

[Nozick, 1969] Nozick, R. (1969). Newcomb's problem and two principles of choice. In Rescher, N., editor, *Essays in Honor of Carl G. Hempel*. Reidel.

[Pettit, 1991] Pettit, P. (1991). Decision theory and folk psychology. In Bacharach, M. and Hurley, S., editors, *Foundations of Decision Theory: Issues and Advances*. Basil Blackwell.

[Price, 1989] Price, H. (1989). Defending Desire-as-Belief. *Mind*, 98(389):119–127.

[Rabinowicz, 2002] Rabinowicz, W. (2002). Does practical deliberation crowd out self-prediction? *Erkenntnis*, 57(1):91–122.

[Ramsey, 1929] Ramsey, F. P. (1990/1929). General propositions and causality. In Mellor, D. H., editor, *Philosophical Papers*. Cambridge University Press.

[Samuelson, 1952] Samuelson, P. A. (1952). Probability, utility, and the independence axiom. *Econometrica*, 20(4):670–678.

[Savage, 1972] Savage, L. (1972). *The Foundations of Statistics*. Dover Publication (revised edition).

[Sen, 1985] Sen, A. (1985). Rationality and uncertainty. *Theory and Decision*, 18:109–127.

[Skyrms, 1981] Skyrms, B. (1981). The prior propensity account of subjunctive conditionals. In Harper, W. L., Stalnaker, R., and Pearce, G., editors, *Ifs: Conditionals, Belief, Decision, Chance, and Time*. D. Reidel Publishing Company.

[Skyrms, 1994] Skyrms, B. (1994). Adams conditionals. In Eells, E. and Skyrms, B., editors, *Probability and Conditionals: Belief Revision and Rational Decision*. Cambridge University Press.

[Smith, 1994] Smith, M. (1994). *The Moral Problem*. Wiley-Blackwel.

[Spohn, 1977] Spohn, W. (1977). Where Luce and Krantz do really generalize Savage's decision model. *Erkenntnis*, 11(1):113 – 34.

[Stalnaker, 1968] Stalnaker, R. (1968). A theory of conditionals. In Rescher, N., editor, *Studies in Logical Theory*. Blackwell.

[Stalnaker, 1970] Stalnaker, R. (1970). Probability and conditionals. *Philosophy of Science*, 64-80(1):23–42.

[Stalnaker, 1981] Stalnaker, R. (1981). A defence of the conditional excluded middle. In Harper, W. L., Stalnaker, R., and Pearce, G., editors, *Ifs: Conditionals, Belief, Decision, Chance, and Time*. D. Reidel Publishing Company.

[Stalnaker, 1984] Stalnaker, R. (1984). *Inquiry*. MIT Press.

[Stefánsson, 2014a] Stefánsson, H. O. (2014a). Desires, beliefs and conditional desirability. *Synthese*, 191(16):4019–4035.

[Stefánsson, 2014b] Stefánsson, H. O. (2014b). Fair chance and modal consequentialism. *Economics and Philosophy*, (forthcoming).

[Stefánsson, 2014c] Stefánsson, H. O. (2014c). Humean supervenience and multidimensional semantics. *Erkenntnis*, (forthcoming).

[Stefánsson, 2014d] Stefánsson, H. O. (2014d). A Lewisian trilemma. *Ratio*, 27(3):262–275.

[Stefánsson, 2014e] Stefánsson, H. O. (2014e). Review of Lara Buchak's Risk and Rationality. *Economics and Philosophy*, 30(2):252–260.

[van Fraassen, 1980] van Fraassen, B. (1980). Critical notice: Ellis, Rational Belief Systems. *Canadian Journal of Philosophy*, 10(3):497–511.

[von Neumann and Morgenstern, 1944] von Neumann, J. and Morgenstern, O. (2007/1944). *Games and Economic Behavior*. Princeton University Press.

[Williams, 2010] Williams, R. G. (2010). Defending conditional excluded middle. *Noûs*, 44(4):650–668.

[Williams, 2012] Williams, R. G. (2012). Counterfactual triviality: A Lewis-impossibility argument for counterfactuals. *Philosophy and Phenomenological Research*, 85(3):648–670.