Thinking the Impossible: Counterfactual Conditionals, Impossible Cases, and Thought Experiments

by

Poonam Dohutia

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirement for the degree of

Doctor of Philosophy

in

Philosophy

Waterloo, Ontario, Canada, 2008

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

In this thesis I present an account of the formal semantics of counterfactuals that systematically deals with impossible antecedents. This, in turn, allows us to gain a richer understanding of what makes certain thought experiments informative in spite of the impossibility of the situations they consider.

In Chapter II, I argue that there are major shortcomings in the leading theories of counterfactuals. The leading theories of counterfactuals (based on classical two-valued logic) are unable to account for counterfactuals with impossible antecedents. In such accounts, everything and anything follows from an impossible antecedent.

In Chapter III, I examine some crucial notions such as conceivability, imaginability, and possibility. Herein I argue that there is a distinction to be made between the notions of conceiving and imagining. Conceivability, it turns out, is a sufficient condition for being a case. Recent literature on the semantics for relevance logic have made some use of the notion of a "state", which differs from a world in that contradictions are true in some states; what is not done in that literature is to clarify how the notion of a state differs from an arbitrary collection of claims. I use the notion of a case as a (modal) tool to analyze counterfactuals with impossible antecedents, one for which, unlike the notion of states, it is clear why arbitrary collections of claims do not count.

In Chapter IV, I propose a new account of counterfactuals. This involves modifying existing possible worlds accounts of counterfactuals by replacing possible worlds by the "cases" identified in Chapter III. This theory discerns counterfactuals such as: "If Dave squared the circle, he would be more famous than Gödel" which seems true, from others like: "If Dave squared the circle, the sun would explode", which seems false.

In Chapter V I discuss one of the main pay offs of having an account of counterfactuals that deals systematically with counterfactuals with impossible antecedents. To apply the new account of counterfactual to thought experiments, first we have to transform the thought experiment in question into a series of counterfactuals. I show how this is to be done, in Chapter V. There are two advantages of such an account when we apply it to thought experiments: First, for thought experiments with impossible scenarios, our new account can explain how such thought experiments can still be informative. Secondly, for thought experiments like the Chinese Room, where it is not clear whether there is a subtle impossibility in the scenario or not, this new account with its continuous treatment of possible and impossible cases makes clear why the debate about such thought experiments looks the way it does. The crucial question is not whether there is such an impossibility, but what is the "nearest" situation in which there is a Chinese Room (whether it is impossible or not) and what we would say there (about the intentionality of the room). On traditional accounts, it becomes paramount to deal with the possibility question, because if it is an impossible scenario the lessons we learn are very different from the ones we learn if it is possible. There are no available theories of thought experiments that account for thought experiments with impossible/incomplete scenarios. With the new account of counterfactual and by applying it to thought experiments we over come this difficulty.

Acknowledgments

My curiosity towards possible worlds stemmed from my early years in undergrad when I first read Kripke's *Naming and Necessity*. Around the same time I was fascinated, as well as bothered by, results of some thought experiments in the personal identity literature. I had been also thinking about truth conditions for counterfactuals like "if my parents had not married, I would not have existed". Four years later, it was after a chat with the late Graham Solomon that I realized that the two could be married to resolve some challenges that counterfactuals pose. Without Graham this project would not have begun. For this and also for feeding my need for Philosophy of Language, I am indebted to Graham. Thus initiated, this thesis became approachable and finally took its shape under Dave DeVidi's supervision.

I owe thanks to many people in the Departments of Philosophy in the universities of Delhi and Waterloo, who have contributed immensely to my philosophical development: Prof. Manju Saxena for taking me under her wings; Professors Arindam Chakrabarti, Nirmalangshu Mukherjee, P. Jetli, Shashi Bharadwaj who made attending classes worthwhile in 35+ Celsius heat, sometimes without electricity; a colleague Nilanjan Bhowmik for bettering me philosophically.

My gratitude to my thesis supervisor Dave DeVidi cannot be overstated. I am indebted to him for making formal logic approachable and sheer fun, for his insight, patience, advise, moral support, and putting up with reading numerous drafts because English is my fourth language. Without Dave I would not have acquired the necessary skills to approach the semantics of relevance logic, counterfactuals, and possible worlds.

I owe very special thanks to Tim Kenyon for his insight and tremendous support. He has been instrumental in this thesis seeing the light at the end of the tunnel.

Some people without whom I could not have made Waterloo home (and probably would have flown back to India) deserve a mention. Prof. Jim Van Evra whose moral support from the time I arrived at Waterloo carried me through. Prof. Angus Kerr-Lawson and his wife Marge provided much needed moral support in a place where I didn't know a soul. Their kindness and generosity will stay with me. My special thanks to Prof. Rolf George for making sure that I was not alone in my first Christmas in Canada and for holding my hand so that I would not slip when I was pregnant. Prof. Bill Abbott, whose kindness, sense of humour, and understanding has been a constant source of inspiration. Linda Daniel and her husband Danny, who invited me to many meals and supported me through rough times. I am especially indebted to a very kind friend Debbie Dietrich. She has always been there to help whether it had to do with figuring out administrative details for funding or a much needed chat and check with reality. Without Debbie, I am not sure how I would have figured out my way through the maze.

Maitri Baruah, Sangeeta Goswami, Shibani Phukan, Anisa Yasmin, Jacinte Jean, Hanan Al-Khalaf, Paola Ansieta, Ignacio Fernandez are some close friends in India and in Canada who have been there for better or worse and have been my source of sustenance, encouragement, and much needed moral support.

Without my family in India I would not have been where I am today. It was my parents Pradip and Bindu Dohutia who instilled in me the value for education, knowledge, hard work and perseverance. I owe it to them for giving me the courage to dream, and for supporting me in every way they could. My sister Bipi and my brother Rajesh who supported me in so many ways that there is not enough room to express my gratitude.

Without my mother-in-law Angela Davis completing this thesis would have proved much more challenging, if not impossible. She has been unconditionally supportive in my getting this Ph.D. I am

thankful to Helen McLachlen and Jim Campbell, who have given invaluable support, a sense of security, and advice.

I owe very special thanks to my two kids Darshanaa and Millan for keeping it real . Their hugs and smiles at the end of day is what kept me going . I find it hard to believe that Darshanaa at the age of 8 has given me valuable advice in finishing this thesis.

Last but not least, my partner in crime and husband Jason Davis who has made this journey so much more fun than I could imagine. I cannot thank him enough for helping me tirelessly in my pursuit. Whether it is challenging me philosophically, explaining logic to me, or proof reading, he has put up with me through everything so far. Counterfactuals that are worth considering: 'If my parents had not met…', 'If I had not come to Canada….', 'If I had not met Jason…'.

Dedication

In memory of my grandfather the Late Raghunath Dohutia—my first teacher, an inspirational philosopher.

Table of Contents

Chapter I: Thought Experiments and Counterfactuals: Undiscovered Links

Counterfactual reasoning is important in many areas of philosophy, science and everyday life, and such reasoning often deals with impossibilities of various kinds. However, existing theories of counterfactuals do not offer a satisfactory account of why some of these counterfactuals are true and others false. In this thesis I develop a theory of counterfactuals that systematically handles counterfactuals with impossible antecedents. Such an account is important in various ways. In the thesis I pay attention to one example of the new theory's philosophical pay offs: having a semantics that allows us to understand how counterfactuals with impossible antecedents work allows us to understand how thought experiments work.

It is common to reason about what would be the case assuming certain impossibilities, e.g., if there were only finitely many primes, Gödel's theorem would not apply to most formal languages. So such counterfactuals are crucial to reasoning. Having an account that explains how these counterfactuals with impossible antecedents work helps us understand why is it useful to reason with such impossibilities in the first place.

There is a large and interesting literature on thought experiments, but as I show in Chapter V, the current book length treatments do not give us a grasp on how thought experiments with impossible scenarios can be useful or informative. Since, (as I show, and as is probably unsurprising when you think about it) thought experiments and counterfactuals are closely related, the theory of counterfactuals with impossible antecedents offers the promise of a better understanding of thought experiments with impossible scenarios.

Thought experiments and counterfactuals are two tools widely used in Philosophy. Both these devices invite the audience to construct a contrary-to-fact scenario, and in so doing, adduce acceptance

or rejection of claims about the real world. Thought experiments and counterfactuals are well known for the philosophical analyses they demand. The many analyses of thought experiments and those of counterfactuals have yielded enormous bodies of literature. While Philosophers are known to analyze the tools they use, the fact that sorting out a crucial class of counterfactuals that have implications for the thought experiments has gone unnoticed.

I take this as an opportunity to develop an account of counterfactuals that will also shed new light on thought experiments. The leading theories of counterfactuals (based on classical two-valued logic) are unable to account for counterfactuals with impossible antecedents. In such accounts, everything and anything follows from an impossible antecedent. Hence such counterfactuals are brushed aside. In Relevance Logic, in contrast, there are some important semantic tools that one can use to analyze impossibilities and inconsistencies, in a more fine-grained and less arbitrary way. Relevance Logics are non-classical logics that require, for the conditional P → Q to be true, that there be some kind of a content-related "connection between P and Q". In other words, for P → Q to be true, P must be relevant to Q in some sort of way.

What we have from classical accounts is a method of determining when a conditional is true based on classical implication. One way or another, the traditional accounts wind up saying that P □→Q (we shall use this symbolization throughout the thesis to stand for the claim "If P were true, Q would be true") is true when P ∪ S entails Q, where S is some suitably chosen set of statements. (The accounts differ in what they include in S—all the statements true in a possible world that is *nearly actual* in some sense, or a set of background presuppositions or whatever. For further details on this see Chapter II.) But if P is incoherent, then, classically, P □→ Q is true for any Q. Consider the

following two examples: “If Dave were to square the circle, he would be more famous than Gödel”, and “If Dave were to square the circle, horses would fly”. Prima facie, the former seems plausibly true, but the latter seems false. Under standard accounts of counterfactuals that depend on a classical notion of logical consequence, both of these counterfactuals come out true. But intuitively only the former seems true. So how do we make sense of counterfactuals of this kind? This is the main question that drives this thesis, for reasons that will presently become clear.

Clearly, to make progress we need to allow that it is possible for P to be incoherent and yet P $\Box\!\!\rightarrow$Q not come out true for some Q. The recent semantics using “states” in Relevance Logic allows for this. The crucial thing about “states” is that they allow for A & ¬A to hold in a state, without forcing every other statement to follow. I will use this feature of “states” and develop a new semantic tool called “cases” to fill the gaps in classical accounts of counterfactuals. This more comprehensive account of counterfactuals—one that can handle counterfactuals with impossible antecedents—is the core of my project. Apart from its inherent interest as a semantics, though, we can also use it to show that thought experiments with impossible/unimaginable conditions are not useless, and to sketch an explanation for why this is so.

This chapter is divided into 6 sections. In section 1, I put in place some basic characterization of counterfactuals. Section 2, is a brief discussion of the basic characterization of thought experiments. In section 3, I provide a list of thought experiments discussed and referred to in this thesis. In sections 4 and 5, I address the questions of why counterfactuals and thought experiments (respectively) are of interest to philosophers in particular. In section 6, I discuss the previously identified links between counterfactuals and thought experiments. In this section I also discuss the various philosophical payoffs that should arise from an improved account of counterfactuals.

Section 1. Counterfactuals

Counterfactuals have been investigated by many thinkers, yielding various accounts. What constitutes a counterfactual is much debated. I will use for the remainder of the thesis the term "counterfactual" as a shorthand for "counterfactual conditional". Consider the statement "Poonam is seven feet tall". This is obviously a *counterfactual statement*. However this is different from the *counterfactual conditional* "If Poonam had been seven feet tall, she would have been the tallest in her family". As a first approximation we can take counterfactuals to be subjunctive conditionals with false antecedents.

Of course, not all conditionals are subjunctive. Besides subjunctive conditionals there are those that are known as indicative conditionals. In the current literature, there is a debate as to whether these two types of conditionals are fundamentally distinct or whether subjunctive conditional is just a past tense form of the indicative. As their names suggest, there is, at least prima facie, a distinction to be made between subjunctive and indicative conditionals. Consider the following pair of conditionals due to Ernest Adams (1970):

(a) "If Oswald didn't kill Kennedy, someone else did" and

(b) "If Oswald hadn't killed Kennedy, someone else would have":

The former is in indicative and the latter is in subjunctive mood. One can easily accept (a) while rejecting (b): we all know that *somebody* killed Kennedy, so we know that if it was not Oswald it was someone else. But there are many people who think that (b) is false, because they think that Oswald acted alone in Kennedy's assassination.

The literature is replete with examples of such pairs. Consider another such pair:

"If Shakespeare did not write *Hamlet*, then some aristocrat did"

"If Shakespeare had not written *Hamlet*, then some aristocrat would have".

Jonathan Bennett points out that although in examples like the above typically the indicative conditional is shown as the acceptable one while the subjunctive is not, this does not always have to be the case. One can also come up with examples in which the subjunctive is acceptable and the indicative unacceptable. Consider Bennett's own example: "The Department wants the eulogy for our honorary graduand to be written by either Alston or Bennett, and we two decide that he will do it. Asked later who wrote the eulogy, I reply that I think Alston did, adding 'If not, I don't know who wrote it.' I would reject 'If he didn't, I did' because I know that I didn't; but I may well accept 'If he hadn't, I would have'" (Bennett, 2003, p. 8).

The point is that, on the face of it, these two kinds of conditionals seem to say different things. Of course, what accounts for these differences, and the degree of the relationship between the two sorts of conditionals, is itself a subject of philosophical dispute. Below I will briefly sketch Robert Stalnaker's and David Lewis's respective views on this. In this context I will also sketch a recent debate on the subject between Bennett and Dorothy Edgington.

3Stalnaker offers a view in which the two kinds of conditionals are more or less similar. In his account indicative and subjunctive conditionals receive structurally similar semantic interpretation. According to Stalnaker, in each kind of conditional (where A is the antecedent and C is the consequent), the speaker says that C is true at an A-world picked out by a certain 'selection function'. Exactly what the selection function is will vary according to the context, though it will always involve some sort of 'most similar' notion. He introduces a 'context set' of worlds, from which the selection function should find its value, if possible. As Bennett explains, " 'the context set' is Stalnaker's name for the set of worlds at each of which all those taken-for-granted propositions are true" (Bennett, 2003, p. 358). The difference between indicative and subjunctive conditionals, then, is that with subjunctive conditionals the selection function may need to go outside the context set to pick its value. So, "[I]f

there are no signals to the contrary, we take it that someone who asserts a conditional is using a selection function that picks an A-world belonging to the context set, that is, a world that is not agreed on all hands to be out of the running as a candidate for actuality. When a speaker uses 'would', however, he thereby signals that he regards himself as free to reach outside the context set, selecting a world that nobody in his vicinity thinks might be actual" (Bennett, 2003, p. 358). Thus in case of a subjunctive conditional the speaker "signals to the contrary" that the selection function is to pick a world outside the context set.

Lewis on the contrary offers a view according to which there are some important structural differences between the two kinds of conditionals. In his book *Counterfactuals*, Lewis provides the following argument for distinguishing between indicative and subjunctive conditionals (what he calls counterfactuals):

> As Ernest Adams has observed, the first conditional below is probably true, but the second may very well be false…
>
> if Oswald did not kill Kennedy, then someone else did
>
> if Oswald had not killed Kennedy, the someone else would have
>
> Therefore, there really are two different sorts of conditional; not a single conditional that can appear as indicative or as counterfactual depending on speaker's opinion about the truth of the antecedent (Lewis, 1973, p. 3)

Lewis's argument is directed against the view (which he alludes to in the above quotation) that indicative and subjunctive conditionals do not really differ in their structure, but only with regard to the truth-value of the antecedent. On this view indicative conditionals do not express any commitment concerning the value of the antecedent, whereas subjunctive conditionals express a commitment to the

falsehood of the antecedent. However, according to Lewis, since there are examples of conditionals (like the ones above) which are the same, save that one is indicative and one is subjunctive, for which the truth conditions are clearly different, there must be more to the difference between the two kinds of conditionals than their view of the truth value of the antecedent.[1]

Bennett also draws a sharp distinction between indicative conditionals and subjunctive conditionals. This has not always been Bennett's view, as he himself admits in *A Philosophical Guide to Conditionals* (Bennett, 2003, p. 13). He had held what is known as the *relocation thesis*, according to which, e.g., the conditional *If you go swimming today, your cold will get worse* (indicative), uttered now, is acceptable to us now if, and only if, the conditional uttered tomorrow, *If you had gone swimming yesterday, your cold would have gotten worse* (subjunctive), is acceptable to us tomorrow (Bennett, 1988).

Edgington holds a thesis similar to the relocation thesis that Bennett calls the *correspondence thesis* (CT), which is the following[2]:

> (CT) For any A and B, if A □→ B is the right thing to think at a certain time, then at some earlier time A → B was the right thing to think. (Bennett, 2003, p. 366)

According to Edgington the two sorts of conditionals, i.e., indicative and subjunctive, are actually fundamentally the same, with counterfactuals merely being the past-tense version of indicatives. One consequence of this is that she rejects both the idea that indicatives have anything to do with the material conditionals and that they are truth functional, as do many other authors working on conditionals these days. In her words:

> That there is not a huge difference between them is shown by examples like the following: "Don't go in there", I say, "If you go in you will get hurt". You look sceptical but stay

[1] I have relied on the following source for this discussion: Fogelin, R. J: 1998, pp.286-289.

[2] Bennett uses '>' for the subjunctive conditional symbol '□→'.

> outside, when there is large crash as the roof collapses. "You see", I say, "if you had gone in you would have got hurt. I told you so" (Edgington, 2006).

Let's consider the following two familiar conditionals:

DD: If (it is the case that) Oswald didn't kill Kennedy, (it is the case that) no one else did.

HW: If it had been the case that Oswald didn't kill Kennedy, it would have been the case that no one else did.

Edgington in "On Conditionals", in a section 'One theory or two?' [on the matter of two sorts of conditionals] discloses that her answer is "one", because the two types of conditionals differ only in tense. If Edgington's thesis about tense is right, it defeats the argument that there must be two conditional connectives. The argument that there must be two conditional connectives is the following: the same two propositions, N [No one else killed Kennedy] and O [Oswald did not kill Kennedy] conditionally connected, express HW and DD. We accept HW and reject DD. So the conditional connective in HW does not mean the same as that in DD.

> On the above [tense] analysis, we do have the same two propositions, conditionally connected in the same way, but connecting different times. Accepting something as probable now, and accepting the same thing as having been probable then, are mutually independent judgments. Thus, for example, the truth of HW comes from the acceptability at an earlier time, for any well-informed person, of 'If Oswald doesn't kill Kennedy, no one else will'; this is indicative, and obviously differs only in tense from DD.

As Bennett makes clear in his book, he is no longer persuaded by such considerations. To understand his reasons, it is useful to begin by noting that he suggests that the correspondence thesis (CT) is no more plausible than the suppositional account of conditionals, because that account is the

only one that would warrant accepting CT.

The suppositional account of indicatives holds that a "conditional judgement involves two propositions, which play different roles. One is the content of a supposition. The other is the content of a judgement made under that supposition. They do not combine to yield a single proposition which is judged to be likely to be true just when the second is judged likely to be true on the supposition of the first" (Edgington, 2006). Thus, Phillips explains, a conditional like, 'If today is Wednesday, then tomorrow is Thursday,' is a conditional assertion of the consequent. If, in fact, it is Wednesday, then this remark is equivalent in force to an assertion of, 'Tomorrow is Thursday.' If today is not Wednesday, no proposition is asserted. According to the suppositional view, conditionals are not true or false. They are evaluated in terms of their conditional probabilities. Notice that conditional assertion and assertion of a conditional probability are two separate things[3].

The extension of the suppositional view to subjunctives relies on the correspondence thesis (CT):

(CT) For any A and B, if A $\Box\!\!\rightarrow$ B is the right thing to think at a certain time, then at some earlier time A $\rightarrow$ B was the right thing to think. (Bennett, 2003, p. 366)

Note the lingering ambiguity in what is meant by "right" here. It could either mean "true" or "warranted".

Consider the following two examples:

(1) Tweedledee and Tweedledum toss a coin and, whilst it is in mid-air, Tweedledee calls heads. The coin lands tails, and Tweedledee loses. It seems that he is right in saying: 'If I had bet tails, I would have won'.

(2) You cancel your booking for a flight which subsequently crashes due to unexpected circumstances.

[3] Also see Ian Phillips's paper "Morgenbesser Cases (MC) and Closet Determinism" for a superb summary of the debate between Bennett and Edgington.

Your relief seems well expressed by the conditional: 'If I had caught that plane, I would probably be dead' (Edgington, 2003) [4].

The above examples point out that while some counterfactuals seem intuitively true, their truth can only be established on the basis of hindsight. As Ian Phillips points out, this feature has significant implications when it comes to theorizing about counterfactuals. For example, for Bennett, (2) poses a problem for extending a suppositional account of the indicative conditional to the subjunctive (Bennett, 2003, pp. 366-369).

Bennett rejects (CT). The reason is that cases like (2) show the existence of counterfactuals that we intuitively regard as correct but for which there seems to be no previously right indicative: 'would haves' without any previously acceptable 'will'. For whilst the coin is in mid-air or the plane on the ground it would be unacceptable to say, 'If I bet tails, I will win,' let alone, 'If I catch the plane, I will die.'

According to Edgington, however, there is (in fact) a previously correct indicative even though it would then have been irrational to endorse it. She thinks that with 'the benefit of hindsight' we would judge even a fortune-teller right if they had said: 'If Poonam boards the plane, she will not live'. After all, on hearing about the plane crash Poonam might exclaim: 'My God, the fortune-teller was right!' (Edgington, 2003, p. 22)

Bennett questions Edgington's ability to make this move. Notice that Edgington says that the fortune-teller 'was right' or 'vindicated' rather than that they spoke the truth. This is because conditionals do not have any truth value on her account. Indicative conditionals instead have objective conditional probabilities. Given this, can we say that the fortune-teller was right insofar as Poonam's death did (in fact) have an objectively high probability conditional on Poonam catching the plane? It

[4] Sydney Morgenbesser cites a similar case in Slote (1978), p. 27. See bibliography for reference.

seems we cannot, for even in hindsight the plane crash was then objectively extremely unlikely. Hence, Bennett complains:

> It is not clear to me what the probabilities are in the light of which the indicatives are judged to be 'right' in hindsight. … [F]or the fortune-teller's conditional to be 'vindicated', room must… be found in the story for a nearly 100 per cent probability of the plane's crashing given that ...[Poonam] was on it. I cannot find… anything allowing us to say that the predicter's conditional probability for the plane's crashing given my being on it was, though not 'justified at the time', correct, right, vindicated (Bennett, 2003, p. 367).

Responding to this Edgington says:

> Lucky guesses are sometimes right… The value to be assigned to the hindsightful counterfactual trumps the most rational value to be assigned to the forward-looking indicative. The chance that C given A, beforehand provides the best available opinion on whether C if A, but it can be overturned by subsequent events, not predictable in advance (Edgington, 2003, p 23).

As Bennett points out, this is not merely hind-sight but hind-rightness-making. In cases like this, the idea of the forward indicative's being 'right' depends on the idea of subjunctive's being right—the explanatory direction runs from subjunctive to indicative, not the other way. The suppositional theory hoped to explain counterfactuals in terms of indicatives. But the above explanation has reversed the order of priorities. It is our independent grasp on subjunctives that seems to determine the rightness or wrongness of the indicatives. Thus, whether or not we hold on to (CT), we remain in need of an independent theory of subjunctives (Bennett, 2003, p. 368).

So much for the views on whether the two kinds of conditionals should be treated as the same or distinct. Without pretending that I have solved the issue, for the purpose of this thesis I side with Bennett's current view. There is something to be said about the distinction between indicative and subjunctive conditionals. I will treat them as distinct and deal with subjunctive conditionals only. As I mentioned earlier I will take counterfactuals as subjunctive conditionals with false antecedents. For the

purpose of this thesis, treating subjunctive and indicative conditionals as separate provides for some much needed simplicity. As I have mentioned before, the main focus of this thesis is counterfactuals with impossible antecedents. In order to deal with conditionals like "If Dave were to square the circle, then the sun would explode", it is hard to see why we would need to have determined whether it is reasonable to accept CT or a similar thesis.

Section 2. Thought Experiments

Thought experiments have held a central role in philosophical inquiry since at least Descartes. They have been investigated extensively by various philosophers yielding numerous accounts. While most philosophers probably think they have at least a reasonable idea of what a thought experiment is, an exact, satisfactory, characterization of the notion is not so easy. In a recent book devoted to thought experiments and their analyses, *Thought Experiment: On the Powers and Limits of Imaginary Cases*, Tamar Gendler provides the following characterization of thought experiments which is quite robust (Gendler, 2000, p. 21):

(1) An imaginary scenario is described.

(2) An argument is offered that attempts to establish the correct evaluation of the scenario.

(3) This evaluation of the imagined scenario is then taken to reveal something about cases beyond the scenario.

For now, suffice it to say that Gendler's characterization does a good job of describing roughly what a thought experiment is. For some classic examples of thought experiments see section 2 of this chapter.

In surveying the literature on thought experiments, I found that one could divide the philosophers with regard to their position on thought experiments, roughly into three broad categories.

First there are those who see thought experiments as being no more, and no more useful, than arguments. Philosophers in this category such as John Norton, maintain that any good scientific thought experiment can be transformed into a non-thought-experimental argument without loss of any demonstrative force. Enthusiasts (like James Brown) of thought experiments, on the other hand, stretch the use of thought experiments much further; he seems to think that a certain class of thought experiments (viz., what he calls *platonic* thought experiments) have an ability to reveal platonic reality. What makes him an enthusiast is that in his view some thought experiments reveal things about platonic facts that are inaccessible by any other means. The third category of philosophers are the moderates. While they believe that thought experiments are a distinct category (and not arguments in disguise, for example) and are useful, they do not make any claims about thought experiments revealing knowledge about platonic or any other realm that is otherwise inaccessible. Both the enthusiasts and the moderates maintain that this tool, when employed, asks the audience to imagine and work through some relevant implications of the hypothetical world. In so doing it can highlight major flaws or contradictions in one's own and others' theories or beliefs. Thought experiments have a surprising ability to reveal previously unknown tensions between explicit, conscious beliefs and implicit, unconscious ones. In many cases it seems that it is in this surprising ability that their persuasive force lies.

Before going on to discuss counterfactuals, it will be useful to gather in one place brief descriptions of several of the thought experiments that I refer to frequently in the remainder of the thesis.

Section 3. Some Paradigmatic Thought Experiments

In this section I will outline five thought experiments that I will consider and refer to

throughout the thesis. I will devote a little more time and space to Derek Parfit's fission and P. F. Strawson's auditory world since they prove to be crucial to my thesis. Strawson's auditory world is non-standard in its formulation and thus, unlike the rest it has proven challenging to provide a succinct version of this thought experiment.

Galileo's Falling Body

Imagine that a heavy and a light body are strapped together and dropped from a significant height. What would the Aristotelian expect to be the natural speed of their combination, given that the Aristotelian thinks that heavier bodies fall at a faster rate than lighter bodies? On the one hand, the lighter body should slow down the heavier one while the heavier body speeds up the lighter one, so their combination should fall with a speed that lies between the natural speeds of its components. (That is, if the heavy body falls at a rate of 8, and the light body at a rate of 4, then their combination should fall at a rate between 4 and 8 (cf. Galileo 1638/1989, p. 107).) On the other hand, since the weight of the two bodies combined is greater than the weight of the heavy body alone, their combination should fall with a natural speed greater than that of the heavy body. (That is, if the heavy body falls at a rate of 8 and the light body with a rate of 4, their combination should fall at a rate greater than 8.) But then the combined body is predicted to fall both more quickly and more slowly than the heavy body alone (cf. Galileo 1638/1989, pp. 107-108). The way out of this paradox is to assume that the natural speed with which a body falls is independent of its weight (Gendler, 2000, pp. 40-41).

Newton's Bucket

Newton proposes this thought experiment for his postulation of absolute space. We are to imagine away all the rest of the material universe, except a bucket of water suspended by a twisted rope. The bucket goes through three distinct successive states:

State 1: at the instant the bucket is released, there is no relative motion between the water and the bucket. And the surface of water is level,

State 2: shortly after that bucket is released the water and the bucket are in relative motion, i.e., motion with respect to one another. The water is still flat in state 2.

State 3: we reach this state after some time has passed. The water and the bucket are at relative rest, i.e., at rest with respect to one another. But the water is not level; its surface is concave at this stage.

Now, how do we account for the difference between state 1 and state 3? We cannot explain it by appealing to relative motion since there is no relative motion at either state. Newton's answer is the following: in state 1 water and bucket are at absolute rest i.e., at rest with respect to absolute space. And in state 3 the water and the bucket are in absolute motion, i.e., in motion with respect to absolute space. The difference in absolute motion explains the observed difference in water level. We should therefore, accept absolute space (Brown, 1991, pp. 9-10).

Ship of Theseus

I have borrowed this version from Gendler (Gendler, 2000, p. 68). There was once a thirty-oared ship that belonged to Theseus, which, during the years, went through gradual repair. In the years, one by one, each of its original planks was replaced with a new plank of the same size, shape, and material, and the old planks were gathered in a barn on the shore. Eventually none of the original planks remained a piece of the original vessel. However, its appearance remained unchanged. One fine afternoon, Theseus collected the planks from the barn, and nailed them together in a form identical to that of the original ship. Suddenly Theseus was presented with both practical and metaphysical difficulties: there were two ships before him. Did one or both require re-christening? Which one was the fine ship of Theseus? Was either one of them identical with the ship he had commissioned some

years back?

Chinese Room

John Searle asks his reader imagine that someone who knows only English is put in a room. She receives sheets of papers with shapes on them through a slot. The person consults a huge rule book linking shapes to other shapes. Symbols come in through the slot and after consulting the rule book the person inside the room puts out symbols linked to the symbol coming in and puts it out through another slot. Unbeknownst to the person inside the room the symbols coming in are encoded questions in Chinese and the symbols going out are answers to these questions. The thought experiment is intended to show that computers merely use syntactic rules to manipulate symbol strings, but have no understanding of meaning or semantics.

Fission

Fission can be reconstructed in the following way: Imagine triplet brothers who were in an accident. In this accident the body of one (call him Brainy) is fatally injured, while the brains of the other two brothers are totally destroyed. Brainy's brain is intact while the other two brothers bodies are in relatively good shape. Brainy is so constituted that the physical bases for his psychological characteristics happen to be realized in duplicate, one complete set in each lobe. Doctors operating on him after the accident divide his brain in half and transplant the two hemispheres into the bodies of the two brothers. There are two scenarios to be considered:

(1) Single-transfer case: only the left transplant is successful, and the right transplant is destroyed. The resulting individual (call him Lefty) has all of Brainy's memories and psychological states and a body almost indistinguishable from that of Brainy's before the accident. From this Parfit concludes: In the single transfer case, Lefty is Brainy.

(2) Double-transfer case: in this scenario both transplants are successful. Each of the resulting individuals (call them Lefty and Righty) has all of Brainy's memories and psychological states and a body almost indistinguishable from that of Brainy's before the accident.

Now, according to Parfit, Lefty and Righty are not the same person (since they occupy distinct spatial locations, undergo different experiences and so on). If Lefty and Righty are different people, and Brainy is a single person, then Lefty and Righty cannot both be identical to Brainy. So in the double transfer case Lefty is not Brainy.

The problem that fission presents is the following: we have a process which, if it happens a certain way (in the single-transfer-case), would result in continued existence of an entity over time (where Lefty is Brainy). If the same process happened in another way (in the double-transfer-case) it would result in the creation of two new entities (where Lefty is not Brainy) (Gendler, 2000, pp. 120-125). As Parfit puts it, "How could a double success be a failure?"

Now, to the extent that the process is intrinsically the same in both cases, how could the rationality of one's attitude towards one's continuer depend on whether the process ends up being entity creating or entity preserving? Parfit concludes that what makes one's prudential concern for oneself tomorrow rational is not the fact that oneself-tomorrow will be the identical to oneself-today, but only that they will be connected by the right sort of relation of psychological continuity and connectedness. So strict numerical identity is not what matters (Gendler, 2000, pp. 146-147).

I should also mention that there is another version of the "fission" thought experiment that is sometimes discussed, and that I will occasionally mention. If we are to imagine Brainy *literally* undergoing fission, dividing, amoeba-like, into two identical Brainy-successors before our very eyes.

Auditory World

Strawson, in *Individuals,* introduces this thought experiment in an attempt to examine the thesis that space is a prerequisite for objectivity. He invites the reader to imagine a being (called Hero, following Gareth Evans) who resides in a world that is completely auditory. Is it conceivable that such a subject be able to grasp the concept of objective particulars? Thus he asks: ‘‘Could a being whose experience was purely auditory have a conceptual scheme which provided for objective particulars?’’ (Strawson, 1959, p. 66). According to Strawson, auditory perception is what is needed for the identification of a sound particular, and the auditory experience of continuity and discontinuity is what is needed for distinguishing sound particulars. But for the re-identification of sound particulars spatial criteria are needed. For re-identifiability, in the auditory world, we need an analogue to spatial dimensions in our conceptual scheme. For an analogue of space in the auditory universe, Strawson introduces the universal master sound (M-sound). Hero distinguishes between different “locations” of a particular sound or sound sequence against the background of the pitch M-sound. Thus the sound world is conceived of as containing many particulars, unheard at any moment, but perhaps audible at other positions than the one occupied at the moment that it is not being heard. This provides for re-identifiability of sound particulars.

According to Strawson, the re-identifiability of particulars in the auditory world is at least a necessary condition for non-solipsistic consciousness. Is it also a sufficient condition? According to Strawson it is, for the concept of a re-identifiable particular entails the concept of a particular's existing while unobserved, and thus the distinction between being observed and being unobserved. But this distinction is based on the idea of an *observer*. In the auditory world, Hero could not make this distinction without having the idea of himself as an observer. In order for Hero to be able to make the distinction between observed and unobserved entities, this distinction should be based on something

purely auditory. He cannot make the distinction based on the idea of an observer, since the idea of an observer is not purely auditory. Being a member of the auditory world, he is just a sound. What we need to figure out is what conditions need to be satisfied in order for Hero to be more—to be a subject of his experience. Thus, the question is: what are the conditions requiring fulfillment for a non-solipsistic consciousness. Or how are the conditions of subjective/objective experience fulfilled? From this question we can ask the more general question: where do I get the idea of myself as a subject that has experiences of things that are other than myself? Strawson runs another experiment to retain the distinction between observed and unobserved entities in the auditory world and goes on a bit further with the thought experiment.

Eventually trying to reproduce the general features of the ordinary world in the auditory world, he concludes that the fantasy of producing as close an analogy " besides being tedious, would be difficult, to elaborate. For it is too little clear exactly what general features to reproduce, and why. It might be better, at this point to abandon the auditory world…" (Strawson, 1959, p. 85).

Section 4. Why Counterfactuals Are of Interest to Philosophers

Counterfactuals are interesting for various reasons. Ordinarily, counterfactuals are an undeniable part of the process of our knowledge acquisition. People often contemplate hypothetically considering "what if" situations, and in so doing, extrapolate information about themselves or their states. One considers the following kinds of situations, e.g., "What if I had not applied to Canadian graduate schools", "What if I had been a millionaire…". The list can go on. One can see that the step from the "what if" form to the subjunctive counterfactual form is small. One can just as easily say "Had I been a millionaire…" or "Had I not applied to a Canadian Graduate school…". In theorizing in science, political analysis, economic analysis and our day-to-day lives, we constantly make use of

counterfactuals.

Philosophically, counterfactuals are interesting from an epistemological and semantic standpoint, inter alia. There are two important questions about counterfactuals that I will ask. The first is the epistemological question, “How could I know whether a counterfactual is true?” and the second is the semantic question “What are the truth conditions for a counterfactual conditional?”.

First let us consider the epistemological question: How can we know when counterfactuals are true, if any of them are true? For some counterfactuals, at least, if they are true, they are contingently so. It seems that the only way to discover contingencies is by looking at the world around us. But how are we to tell what *would* have happened, if something *had* been the case (that in fact was not the case)? Observations, after all, seem at best a tool for finding out what *is* the case. The fact that counterfactuals do not seem to be about the actual—rather that they seem to be about possible states of affairs other than the actual ones (setting aside the ones with impossible antecedents for now), yet they tell us something about the real world (at least the interesting or useful ones do)—makes them epistemologically rather mysterious. It is this epistemological mystique that draws philosophers' initial attention to them.

Another important question for philosophers is how are we to distinguish counterfactuals that are true from the ones that are false? In other words, what are the truth conditions for a counterfactual conditional? Consider the following counterfactual: “Had Bush not been the president of the U.S., the Iraq war *would not* have taken place”. While Bush was the president, the above sentence seems true or in any case likely, and, if it is, it tells us something about the real world. Of course, this sentence comes out true if counterfactuals are understood simply as the familiar material conditionals of classical logic, but so do sentences like "Had Bush not been the president, Iraq *would* have been invaded," which seems false if the first is really true. It is an interesting fact about counterfactuals that they are non-

truth functional. That is to say that the truth value of a counterfactual does not depend merely on the truth value of its components. So, a truth functional analysis of counterfactuals proves unsatisfactory.

How are we to distinguish true counterfactuals from false ones? In order to distinguish true counterfactuals from false counterfactuals we have to analyze them a certain way. Much of the philosophical work involved in the investigation of counterfactuals comes down to sorting out suitable truth conditions for them so the different truth conditions of counterfactuals like the above can be explained. We find numerous accounts by various philosophers that try to do just that. We will consider various analyses proposed by philosophers such as Goodman, Chisholm, Stalnaker, and Lewis, in Chapter II.

Section 5. Why Thought Experiments Are of Interest to Philosophers

What about thought experiments? As a philosopher it is hard not be irked when a first year student's reaction to, say, John Locke's discussion of the consciousness of a prince waking up to in the body of a pauper is something like: "Thought experiments? Why waste your time talking about things that never happen?" In response I like to point them to the following passage from Roy Sorensen:

> At this point [after having warped their minds with brain-in-a-vat scenario] many students feel they got their money's worth. They leave class like they leave an absorbing matinée---a little disoriented, a little preoccupied, depressurizing to everyday reality. But there is the occasional objection, "So what? What do those brains [in-a-vat] have to do with anything? They are just *hypothetical*." One could respond with correct but arcane allusions to counterfactuals and the general relevance of possible worlds. Tu quoque is also tempting. Doesn't any sensible student heed *possibilia* when crossing the street or practicing birth control? But the best response is to find partners in crime: scientists conduct thought experiments, so why pick on the philosophers? (Sorensen, 1992, p. 8)

At this point the first year student can still come back and say: "So what? The fact that scientists use thought experiment is hardly a reason why they are of any interest to philosophers". Gendler's words point us towards answering this question:

> Thinking about imaginary cases can help us learn new things about the world. This simple fact is both a commonplace and a puzzle. It is a commonplace because it is undeniable that imaginary cases play a central role in our investigation of the world—in legal reasoning, in linguistic theorizing, in philosophical inquiry, in scientific exploration, and in ordinary conversation. And it is a puzzle because it is *prima facie* surprising that thinking about what there isn't and how things aren't should help us to learn about what there is and how things are (Gendler, 2000, p. 1).

An important philosophical issue that arises when considering Gender's point is: *How* could thought experiments work? In trying to answer this question philosophers offer different analyses. For John Norton, there is nothing special about how thought experiments work. Every scientific thought experiment is replaceable by an argument. So a good thought experiment is a good argument in disguise. James Brown offers a very different analysis; although some thought experiments are replaceable by arguments, there is a special class of thought experiments (that he calls *platonic* thought experiments) that work by revealing facts about platonic realm. Gendler offers an account in which thought experiments lead us to new knowledge, helping us reconfigure old data in a new light. Thus one can see that thought experiments puzzle philosophers as much as counterfactuals do.

Section 6. Links between Counterfactuals and Thought Experiments

Both Gendler's and Sorensen's comments in the previous section help us see the connection between counterfactuals and thought experiments. It is precisely the puzzling feature of thought experiments that Gendler talks about that also make counterfactuals interesting: that in spite of being about merely possible states of affairs counterfactuals seem to tell us something about the actual world. Sorensen seems to take it for granted that one can allude to counterfactuals and possible worlds, in order to explain the usefulness of thought experiments.

As I mentioned in the beginning one important philosophical pay off of developing a comprehensive theory of counterfactual that deals with counterfactuals with impossible antecedents is

that with such a theory we can account for thought experiments with impossible scenarios. For this purpose, we start by analyzing thought experiments in terms of counterfactuals.

It is an underlying assumption made by many theorists, including Gendler, that the prevalent uses of thought experiments in Philosophy, particularly in areas of Metaphysics and Philosophy of Mind, are seriously limited. These thought experiments are limited, according to Gendler, because of some significant inherent flaws in them. These flaws are a result of using a thought experiment to derive unwarranted conclusions. These conclusions are not warranted by the underlying theory at hand, in the context of which the thought experiment aims to generate new knowledge. As I shall argue in Chapter V, Gendler's account suffers the following limitations: (a) it fails to account for the usefulness of thought experiments that involve imagining impossible scenarios, (b) it fails to recognize that there is not always an underlying theory in the context of which a thought experiment tries to generate new knowledge, and (c) imaginability cannot be a formal criterion for characterizing thought experiments.

It is well recognized that counterfactual sentences are used in the articulation of thought experiments; most thought experiments can be expressed in terms of a series of counterfactuals where, in each case, the antecedent (though unnaturally long) recapitulates the imaginary scenario described and the consequents depict possible outcomes of the thought experiment. By developing a new semantics of counterfactuals with impossible antecedents and analyzing thought experiments with such an account, I may be able to overcome some shortcomings in Gendler's account. I may be in a position to shows how some thought experiments, like Strawson's auditory world, can be informative. Strawson puts forth this thought experiment in the second chapter of *Individuals* to probe the thesis that space is a prerequisite for objectivity. Strawson's objective in this thought experiment is to better understand the concept of objectivity in the actual world by construing a parallel but very different conceptual scheme and trying to reproduce the features of the actual world that are required for

objectivity, in that world. He invites the reader to imagine a completely auditory world with a being residing in that world. Could this being have the concept of objective particulars, as separate entities from himself? Strawson asks. Although Strawson succeeds in reproducing some of the features of the ordinary world in the auditory world, he eventually abandons the project as he finds that the task, besides being tedious, is also very difficult, as it is not clear exactly what general features to reproduce and why.

In this thesis I shall argue that there are major shortcomings in the leading theories of counterfactuals. Within the book length accounts of thought experiments, there are no available theories that account for thought experiments with impossible/incomplete scenarios. At best such thought experiments are discounted as having little use for generating new knowledge. Similarly, as we already noted the leading theories of counterfactuals (based on classical two-valued logic) are unable to account for counterfactuals with impossible antecedents. In such accounts, everything and anything follows from an impossible antecedent. A more comprehensive account of counterfactuals—one that can handle counterfactuals with impossible antecedents is the core of my project. Apart from its inherent interest as a semantics, though, we can also use it to show that thought experiments with impossible/unimaginable conditions are not useless, and to sketch an explanation for why this is so.

So within the counterfactual literature there is no solution for counterfactuals with impossible antecedents and within the thought experiment literature there is no good story about thought experiments that involve impossible/incomplete scenarios. Is this a mere coincidence—absence of a way to account for impossibilities? Although I do not have an answer to this question, it suggests a strategy. Perhaps it will help us figure out something crucial if we try extending a sound analysis of counterfactuals to thought experiments. We need a new theory of counterfactuals—one that will account for counterfactuals with impossible antecedents. If we can construct such a theory it may be

helpful to analyze thought experiments with impossible/incomplete scenarios in terms of a counterfactual, and to figure out their truth conditions.

In this thesis I will show how with the help of Relevance Logic and Lewis-Stalnaker style semantics we can construct a theory that will be able to handle counterfactuals with impossible antecedents. Such an account has the following two advantages: First, with the help of such an account we will be able to distinguish between counterfactuals such as "If Dave squared the circle, he would be more famous than Gödel" which seems true, and "If Dave squared the circle, the sun would explode", which seems false. Secondly, by translating thought experiments into counterfactuals we will be able to account for the usefulness of thought experiments that are incomplete/impossible.

The remainder of this thesis is divided into four chapters. Chapter II will begin the necessary preliminary conceptual spadework by briefly surveying the available accounts of counterfactuals. As we will see, theories of counterfactuals can be broadly divided into two categories viz., meta-linguistic theories and theories that depend on possible world semantics. Under the first category, I will be considering in particular the works of Goodman, Chisholm, and Kvart. And under the possible world theories I will be discussing the theories due to Lewis and Stalnaker. I will indicate some of the crucial advantages and disadvantages of each account.

Much of Chapter III is dedicated to isolating the notion of a "case". Conceivability it turns out is a sufficient condition for a set of sentences to be a case. Cases, which are strategically defined in Chapter III, are used a modal tool for analyzing counterfactuals with impossible antecedents. In Chapter III, I begin by examining the notions of conceivability and imaginability, and argue that there is a distinction to be made between the two. In this context I also examine the question: is conceivability a guide to possibility? There is a natural tendency to assume that conceivability entails possibility. I argue that conceivability is not a guide to possibility although "full and coherent describability" is. But

this is far from a reason to regard conceivability as philosophically unimportant. While there are impossibilities that are conceivable, not all impossibilities (and in particular, not every contradictory set of sentences) are conceivable, and those that are conceivable are the "cases".

In Chapter IV, we will provide an account that offers a systematic analysis of counterfactuals with impossible antecedents. It is, in essence, a modification of the possible worlds accounts described in Chapter II, under which a counterfactual is true if its consequent is true at the "closest" world in which the antecedent is true. However, "possible worlds" will be replaced with "cases". And there remains philosophical work to do, because the existing analyses of "closeness" make sense for possible worlds, but do not tell us much about how to tell which impossible cases are closer than any others to actuality. This topic will occupy much of that chapter.

In Chapter V, I will show that with this new semantics, thought experiments with impossible/unimaginable conditions need not be discounted. With the help of four crucial thought experiments I will show how the new theory of counterfactuals allows for some nice representation of crucial philosophical issues in hand.

Chapter II: Counterfactuals: An Overview

The main purpose of this chapter is to survey the main body of literature on counterfactuals. Needless to say, there are numerous accounts of counterfactuals. The various analyses proposed for counterfactual statements can be divided into two broad categories: metalinguistic or syntactic analyses in the pre-Kripke era and the possible world approach in the post-Kripke era. In this chapter, I will discuss several accounts of counterfactuals; specifically the works of Chisholm, Goodman, Kvart, Stalnaker, and Lewis. This chapter naturally divides into two main parts. Part 1 is a discussion of syntactic/metalinguistic accounts of counterfactuals. In part 2, possible world accounts are discussed. In part 1 sections 1 and 2, I will discuss Chisholm's, and Goodman's respective accounts. In Part 2, I address the following questions: Why possible worlds? Why do Lewis and Stalnaker make the shift form a Chisholm/Goodman type syntactic approach to the possible world approach? In part 2 section 1, I will briefly introduce the basic semantics for possible worlds. Sections 2 and 3 of part 2 are discussions of Stalnaker's and Lewis's particular accounts of counterfactuals respectively. In section 4, I consider the fundamental differences between these two accounts. Section 5 is a brief discussion of Kvart's syntactic account of counterfactuals. Notice that although Kvart offers a syntactic account of counterfactuals he falls into the post-Kripke era. Section 6 is a brief introduction to the generalized Lewis-Stalnaker style analysis that I adopt in a later chapter to analyze counterfactuals with impossible antecedents.

Part 1. Syntactic/Metalinguistic Theories of Counterfactuals

Chisholm's and Goodman's accounts, which fall into the pre-Kripke era, are examples of the syntactic approach to counterfactuals, while Stalnaker and Lewis are examples of the (by now

orthodox) possible worlds approach, in the post-Kripke era.

Also known as cotenability theories of conditionals, the basic idea behind Chisholm's and Goodman's accounts is that a conditional is assertable if its antecedent, together with its cotenable premises, entails its consequent. So, $p \rightarrow q$ is true if q follows by law from p together with a set Γ of true sentences [such that for any $r \in \Gamma$ it is not the case that $p \rightarrow \neg r$]. Thus the *truth conditions* of conditionals are evaluated on the basis of whether or not an argument exists from the antecedent and suitable cotenable premises to the conditional's conclusion.

At the time Chisholm and Goodman produced their accounts of counterfactuals, it was clear that these conditionals were part of a cluster of inter-related notions that are central to our understanding of science. For instance, *dispositional properties are central* to many sorts of explanations—we explain the dissolving of salt when placed in water by appeal to its *solubility*, the breaking of glass when struck by its *fragility*, and the logical behaviourists reduced mental states to dispositional properties to behave. But it is clear that standard accounts of indicative conditionals won't easily represent dispositional terms—for example, "If this salt is put in water it will dissolve" is true if the salt is never in fact put in water. So if we use this as our account of what it means to say that salt is soluble then, if my shoe never finds its way into water, this definition would count it as soluble as well. The difference between the two seems to be that when we take these conditionals in *subjunctive* form, the salt conditional is true, the shoe conditional not. So subjunctives seem likely to be a key to explaining dispositional terms. Similarly, and famously, a key difference between l*aws of nature* and mere *accidental generalizations* is that laws "support counterfactuals". A law like "metals expand when heated" implies that if this iron bar is heated it will expand. "All the coins in my pocket are dimes" is only accidentally true, so "If this quarter was in my pocket, it would be a dime" is not implied.

Thus a clearer understanding of counterfactuals promises a clearer understanding of many things. However, the prospects of understanding counterfactuals independently of these other notions did not seem promising. Edgington describes the guiding idea for Chisholm and Goodman as follows:

> Counterfactuals appeared to be connected not only with dispositional properties but with laws of nature. Laws, it seemed, have counterfactual implications, accidentally true generalizations don't. If we understood counterfactuals, this might illuminate the notion of law. And conversely. Leaving the problem "What is a law?" for another day, perhaps counterfactuals can be explained as law-governed conditionals (Edgington, 1995, p. 247).

Before discussing the particular accounts, it is worth mentioning here that although it can be safely said that most syntactic accounts like Chisholm's and Goodman's fall in the pre-Kripke era, we see an interesting revival of such accounts in the post-Kripke era as well. Kvart's account is such an example. We will discuss Kvart's account after Stalnaker's and Lewis's in part 2 of this chapter. Now let us look at Chisholm's and Goodman's accounts respectively.

Section 1. Chisholm

Chisholm derives his account from F.P. Ramsey who comments:

> In general we can say with Mill that 'if *p* then *q'* means that *q* is inferable from *p*, that is, of course, from *p* together with certain facts and laws not stated but in some way indicated by the context. This means $p \supset q$ from these facts and laws… If two people are arguing about '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* (Chisholm, 1949, p. 489).

Thus the counterfactual "If *H* were to be the case, *W* would be the case" (call this *C*), could be thought of as another way of saying that the indicative statement 'H' together with certain previous information, entails '*W*'. But, Chisholm points out, it is important to spell out what "previous information" refers to, since, the *meaning* of the conditional should not be confused with the *grounds* on the basis of which it is asserted. Two people may have completely different stocks of knowledge and could affirm *C* on

extremely divergent grounds. But when each of them affirms *C* it must be assumed that she is saying exactly the same thing. They may supplement *H* with additional information, in order to deduce *W*. So the *meaning* must allow that they may supplement *H* with *W*. But since this statement (with the same meaning) can be asserted by anyone, this supplementary additional information added to *H*, need not be a statement expressing any particular item in either of their stores of knowledge, nor indeed need it express any knowledge at all. In asserting a subjunctive conditional we say something more general —"that there is *some* true statement which, taken with *H*, entails *W*" (Chisholm, 1949, p. 490). This amounts to what Kvart calls **Chisholm's Formula** (CF) (Kvart, 1986, p. 3)

CF: $(\exists\, p)$ [*p* is true & $(p \,\&\, H \rightarrow W)$]

Although CF might appear to be the most plausible way of translating counterfactuals, it faces the following two problems, considered by Chisholm. The first one has to do with the with the trivialization of the formula when we use certain substituends for *p*. Chisholm considers a variety of cases that might have this result.

The first case has to do with universal conditionals with an antecedent (but not a consequent) which contains free variables that are vacuous. For instance, let:

H = "Poonam won scholarship number 21423 from SSHRC".

It follows from H that:

p = "there is an x such that Poonam won x from SSHRC".

Now, since Poonam did not win a scholarship from SSHRC,

p* = "for all x (if x is a SSHRC scholarship Poonam has won, then dogs can fly)"

is vacuously true, because there is no x that has the property attributed x in the antecedent of the conditional.

But, of course, in classical logic p* is equivalent to[5]:

q = "((there is an x such that x is a scholarship Poonam won x from SSHRC) → dogs can fly)"

So, since the universal conditional is vacuously true, so is the existential conditional sentence.

Recall, H ("Poonam won scholarship number 21423 from SSHRC") implies p ("there is an x such that Poonam won x from SSHRC"). And H & p* implies W ("dogs can fly"). Therefore, H implies W.

Thus "if Poonam won scholarship number 21423 from SSHRC then dogs can fly" comes out true in Chisholm's account. If this process is carried out many false counterfactuals including "if H then ¬W" come out true. But obviously both "if H then W" and "if H then ¬W" cannot be true.

Chisholm's suggestion to avoid this problem is "to insure that it [CF] contain no universal statement whose antecedent determines an empty class and no material conditional (or material implication) whose antecedent which is asserted merely on the ground that it is false (or its consequent true). Every universal conditional included in *p* must have 'existential import', that is, every universal conditional must have conjoined with it a statement asserting that there are members of the class determined by the antecedent" (Chisholm, 1949, p. 491).

CF faces another threat of trivialization on the following count, even if we disallow vacuous universal conditionals from being included in *p*. The problem can arise in the following way: Take a conditional of the form "if *Fa*, then *Ga*", where *a* denotes. To take Chisholm's own example, suppose Fx = x sees the play and Gx = x enjoys the play. Now let the following statement be a substituend for *p* in **CF** (*x*) [*x* = *a* → (*Fx* → *Gx*)]. The problem of vacuous universal conditionals (i.e., a universal conditional with an empty antecedent) does not arise since '*a*' denotes. But since, *Fa* is false, insofar as it is the antecedent of a counterfactual, the proposed substituend for *p* is true, and together with *Fa*, it

[5] It is a fact of classical predicate logic that the universal conditional "for all x (x → W)" (where x occurs free in the antecedent and not in the consequent) is equivalent to "(∃x) (…x…) → W".

clearly implies *Ga*, thus making the counterfactual true under Chisholm's amended formula. But obviously, many counterfactuals of the form are false. To avoid such difficulty we have to add a further restriction on our formula; "Let us say: *p* includes no universal conditional whose consequent includes any two functions which are logically equivalent to '*x* sees the play' and to '*x* does not enjoy the play': i.e., any consequent must exclude either functions logically equivalent to "*x* sees the play" or functions logically equivalent to "*x* does not enjoy the play" (Chisholm, 1949, p. 491). However, this restriction is not enough to preclude trivialization either. There is a third case where the formula might be trivialized—where the consequent of the formula is true. We should add a further restriction viz., that the indicative version of the consequent does not entail *p*. If this restriction is imposed then the problem of trivialization can be avoided.

An easy way to show how this problem arises in our original example "if H were the case, W would be the case", is to choose W for p in CF. This will always make such a counterfactual come out true, in Chisholm's account. The restriction that W does not entail p neatly rules out substituting W for p, since W obviously entails W. Chisholm however, does not think that all these restrictions are sufficient to overcome these difficulties. He writes "…the above restrictions as they stand are really not sufficient to exclude types of cases for which they are designed….With a little ingenuity, these restrictions may be evaded and,…in order to deal with the cases hitherto considered, our formula must be one of extraordinary complexity" (Chisholm, 1949, p. 492).

The second problem has to do with how to distinguish accidental generalizations, which do not warrant the inference of certain counterfactuals from non-accidental, law-like statements. For example, suppose two men were to sit on a park bench, quite independently of each other and that each of them were Irish, as it happened. We could then say: "(*x*) (if *x* is on….park bench at…time, *x* is Irish)". Our formula then, in a particular case, could entail "If Ivan were to be on…park bench at…time, Ivan would

be Irish". But this is not warranted, i.e., that counterfactual is actually false. Again, consider Chisholm's own example of a small community where each of the lawyers happens to have three children. We may say: "(*x*), if *x* is a lawyer in…community in 2002, *x* has three children". But we should not want to say of Jones who we know not to be a lawyer, that if Jones were to practice law, then there he too would have had three children. The difficulty is that our universal conditionals about the park bench and the lawyers describe what are, in some sense, "accidents" or "coincidences". How are we to distinguish such "accidental" conditionals from statements such as "all men are mortal", "all wolves are ferocious", etc., which describe "non-accidental" connections, and which do not give rise to similar problems when substituted for p in CF? Our formula must exclude such "accidental" universal conditionals. But the only obvious way to exclude these is to say that unlike the "non-accidental" ones, they do not warrant the inference of certain counterfactuals.

Chisholm suggests two alternatives for excluding "accidental" conditionals. They are (1) supply the qualifications CF lacks to handle these cases or (2) accept the counterfactuals as expressing some irreducible connection between the antecedent and the consequent, and thus reject or alter radically the extensional logic.

Kvart argues that no modification of the sort Chisholm offers is likely to save his account, and so we need a change of theory rather than putting mere patches on the original. His reasoning is as follows: Take for example, "if *A* were the case, *B* would be the case". Now one can take $A \rightarrow B$ for *p*, since *A* is false, $A \rightarrow B$ is true thus rendering the formula CF, in this case $(A \ \& \ A \rightarrow B) \rightarrow B$ true. Under the proposed criterion, every counterfactual comes out true. We are back to square one, for nothing is more troubling than the fact that under a proposed criterion every counterfactual comes out true (or false) for it is well known that there are false as well true counterfactuals. "A natural way to overcome this difficulty would be to require ***B* to not entail *p***. This move would, of course, do away

with all the difficulties Chisholm specified: it would rule out taking p as '(x) [$x = a \rightarrow (Fx \rightarrow Gx)$] in the second case [since $Ga \rightarrow Fa \rightarrow Ga$]" (Kvart, 1986, p. 6). It would also rule out vacuous universal generalizations in the first case; and it would rule out taking *p* as *B*, in the third case.

> However, this move is insufficient, since we can specify, for every counterfactual "If *A* were true, *B* would be true", a statement *M* that does not follow from *B*, and take *p* as $A \rightarrow (M \,\&\, B)$. In this case *p* is still true since (*A* is false), but in general *B* does not entail *p*; and since $A \,\&\, [A \rightarrow (M \,\&\, B)] \rightarrow B$, Chisholm's criterion is satisfied, and is thereby trivialized. Almost every counterfactual can be made to come out true this way….The use of this type of substituend for p allows for across-the-board trivialization, not just counterexamples for particular cases. (Kvart, 1986, pp. 5-6)

One more attempt might be made to save Chisholm's proposal, by putting the blame on A and requiring that ¬A does not entail p. We can still trivialize the criterion by selecting a true statement N such that ¬A does not entail p and B does not entail N, and take p as N& (A $\rightarrow B$). In this case p is certainly true since A is false and N is true. Chisholm's restriction on CF is satisfied since neither $\neg A \rightarrow [N \,\&\, (A \rightarrow B)]$ nor $B \rightarrow [N \,\&\, (A \rightarrow B)]$ (in other words, neither ¬A, nor B entails p) and yet [A & N & (A $\rightarrow B$)] $\rightarrow B$. Thus again, the criterion is trivialized. Chisholm's strategy for adding restrictions to avoid trivialization does not look promising. In principle, every trivialization procedure of the above kind may be avoided by adding more restrictions but there is no telling how complicated the amended formula will end up and where we can stop with these restrictions. In general, trivialization procedures are more serious than just falsification of a formula since trivialization produces a host of counterfactuals for the formula and thereby shows that the problem is not limited to an isolated case only.

Section 2. Goodman

Goodman's approach to the problem of counterfactuals is quite similar to Chisholm's. Recall that Chisholm's formula was: $(\exists p)$ [*p* is true & $(p \,\&\, A \rightarrow B)$]. Goodman adopts the same formula with

the additional requirements: (1) (p & A) must be self compatible; and (2) the entailment must proceed via some relevant laws. According to Goodman, a counterfactual is true if and only if "for some set S of true sentences, A & S be self compatible and lead by law to the consequent" [note, Goodman writes "." for "&"] (Goodman, 1965, p. 11) [We will call this CFG 1].

This has the following form:

CFG 2 : $(\exists s)$ [S & (A & S $\rightarrow^{L}$ C) & R (S,A)]

where "C" stands for the consequent, "$\rightarrow^{L}$" means inferability by law and R (S, A) is a constraint on S, which in CFG 1 was the constraint that A & S be self compatible.

Thus, as Kvart points out, the line of argument against Chisholm's formula applies equally well to Goodman's—A choice of A $\rightarrow$ *B* as S fulfills the requirements of leading to B by law (albeit vacuously) and of being compatible with A (except in the case in which A is incompatible with B, which is a relatively uninteresting and minor subcase) (Kvart, 1986, p. 9).

Thus almost every counterfactual will come out true under formulation CFG 1. Goodman, in the face of this trivialization, proposes the following criterion: CFG 3: "A counterfactual is true if and only if there is some set S of true statements such that A & S is self-compatible and leads by law to the consequent, while there is no such set S′ such that A & S′ is self compatible and leads by law to the negation of the consequent" (Goodman, 1965, pp. 11-12). This has the following form:

CFG 4: $(\exists s)$ [S & (A & S $\rightarrow^{L}$ C) & R (S,A)] & - $(\exists S')$ [S′ & (A & S′$\rightarrow^{L}$ ¬C) & R (S′, A)]

As Goodman points out this step faces the problem that S′ can be taken as {¬C} [C here stands for the consequent], where ¬C is compatible with A. This would make the second conjunct false and trivialize the new addition. Notice that this trivializes in the opposite direction, making all counterfactuals false instead of making them all true. Goodman, re-formulates his criterion in the following way : (we will call this CFG 5)

CFG 5 "…a counterfactual is true if and only if there is some set S of true sentences such that S is compatible with C and with ¬C and such that A & S is self compatible and leads by law to C; while there is no set S′ compatible with C and with ¬C and such that A & S′ is self compatible and leads by law to ¬C" (Goodman, 1965, p. 13).

This formulation is not free from trivialization either. However, for the sake of brevity we will not go into the further details of the threat this formulation faces. Goodman observes that the last formulation faces difficulty of different sort: A further restriction on the set is required on the set S. It cannot include statements that in Goodman's terminology are "not cotenable" with A (i.e., the statements that *would not* be true if A were true). Goodman is immediately faced with circularity for cotenability is defined in terms of counterfactuals, whereas now the truth conditions for the counterfactuals are defined in terms of cotenability. Goodman recognizes this problem himself and says the following: "Though unwilling to accept this conclusion, I do not at present see any way of meeting this difficulty" (Goodman, 1965, pp. 16-17).

As we can see, the two main metalinguistic accounts of counterfactuals face severe problems —Chisholm's account faces threats of trivialization and Goodman's account faces circularity. Such results may be a reason for departure from this syntactic approach.

Part 2: Why Possible Worlds?

Before we go on to discuss Lewis's and Stalnaker's particular possible world approaches for analyzing counterfactuals we need to say something about the shift from the syntactic to the possible worlds approach. Why was the possible world approach so successful in convincing almost everybody to give up the syntactic approach? What advantages does it have over those approaches?

In Dorothy Edgington's words:

> With Saul Kripke's semantics for modal logic (1963) came the revival of the philosopher's dream, a possible world. It is a promising tool for the elucidation of non-truth-functional sentential connectives. It is certainly useful in the formulation and clarification of modal thought. And it is natural to turn to it for an elucidation of conditionals, which, on the face of it, are about possible situations (Edgington, 1995, p. 250).

So, with Kripke's modal logic, philosophers now had the tools to analyze possibilities with formal rigour—an option pre-Kripke philosophers did not have, although it was realized that the problem of counterfactuals spans beyond syntax. Hence Chisholm's comment:

> Like Russell in his theory of descriptions, we want to find a new way of saying something —in this case, in order to assure ourselves that we *can* restate what we ordinarily express in subjunctive conditionals. The problem is epistemological and metaphysical, as well as logical and linguistic; we want to know what it is, if anything, that we have to assume about the universe if we are to claim validity for our counter-factual knowledge (Chisholm, 1949, p. 486).

Chisholm (and likewise Goodman) essentially tried to unpack a conditional in terms of its syntactical properties and determine what we have to assume about the world in making a claim of this sort. So analyzed, these statements seemed to be connected with dispositional properties and laws of nature. However, Chisholm and Goodman did not have the tools available to them to develop logical systems in which they could show how these counterfactuals behaved. Hence, the logical side of the analysis remained un(der) developed.

Now the question that naturally comes up is: what is it about the possible worlds approach that makes it easier to formulate a plausible account of the meaning of counterfactuals than it is on a syntactic approach? The syntactic accounts of Chisholm and Goodman considered in Part 1 involve saying that A → B is true if B "follows from" A along with some extra assumptions. The job then was to figure out what are the legitimate candidates to serve as the extra assumptions. There was a lot of effort devoted to figuring out what these extra assumptions are. Proponents of these approaches argue

that the "surface logical form" of subjunctive conditionals leaves out this extra detail, and try to specify what the true logical form of the statements is. The possible worlds accounts take the surface logical form more seriously: For instance, a sentence like "If A were the case, then B", we assume that it is really a statement about alternative ways things might be, in particular ways in which A might be true. The benefit is that there is no need for generating a class of statements that, along with A, makes B derivable; the cost is assuming that these "alternative ways" are something we need to take with some degree of metaphysical seriousness. Also, we must recognize that there are many alternative ways in which A might be true, and only in some will B also be true. So which ones(s) do the job of determining the truth value of the conditional needs to be figured out. That is, we need to take on the job of picking out "nearby" worlds in which if A is true, B is true also.

So much for the shift from metalinguistic accounts to possible world accounts of counterfactuals. Before considering particular accounts, it will be handy to have some details about possible world semantics.

Section 1. Possible World Semantics

Stalnaker's and Lewis' accounts of the truth conditions of counterfactuals make use of Kripke's possible worlds semantics. In the usual Kripke semantics for modal logic, one begins by defining a **frame**. A frame is a set of worlds[6], and a relation R, defined on the set of worlds, called the (alethic) **accessibility** relation. An **interpretation** for a frame is a (logically consistent) truth-value assignment to the sentences of the language in each world. In the analysis of counterfactuals, one often identifies a world with the set of all sentences true within it, and restricts attention to one particular frame under some particular interpretation. We assume that there are as many sentences as there are facts about a world, and so these sentences are expected to describe the world completely. Usually the

[6] For a detailed discussion about what these worlds are see Bennett, 2003, pp. 152-158

number of worlds is assumed to be the cardinality of the power set of the set of sentences. Some subset of these other worlds is the set of possible worlds, with respect to a given world w: these are the worlds w has alethic access to. In addition to the usual logical connectives, '$\Box\rightarrow$' is introduced as the subjunctive conditional, and it is to be read as, 'if it were the case that _, then it would be the case that _'.

Section 2. Stalnaker

Stalnaker, in his analysis of counterfactuals, relies on a function (which he calls the selection function) that takes as inputs a base world *i* and an antecedent A and returns the nearest possible world to *i* in which A is true**.**

In his paper "A Theory of Conditionals", Stalnaker identifies three problems associated with counterfactuals. These three problems are the following:

(1) The *logical problem* of conditionals: this involves "the task of describing the formal properties of the *conditional function*: a function usually represented in English by the words "if… then", taking ordered pairs of propositions into propositions" (Stalnaker, 1968, p. 41).

(2) The *pragmatic problem:* this "derives from the belief…, that the formal properties of the conditional function, together with all of the *facts*, may not be sufficient for determining the truth value of a counterfactual; that is, different values of conditional statements may be consistent with a single valuation of all non-conditional statements. The task set by the problem is to find and defend criteria for choosing among these different valuations" (Stalnaker, 1968, p. 41).

(3) The *epistemological problem:* this "is based on the fact that many counterfactuals seem to be synthetic, and contingent, statements about unrealized possibilities. But contingent statements must be capable of confirmation by empirical evidence, and the investigator can gather evidence only in the

actual world. How are conditionals which are both empirical and contrary-to-fact possible at all? How do we learn about possible worlds, and where are the facts (or counter-facts) which make counterfactuals true?" (Stalnaker, 1968, p. 42). Such concerns, as Stalnaker points out, have led philosophers to analyze counterfactual conditionals in non-conditional terms. Steering away from analyzing conditional statements in terms of non-conditional statements, Stalnaker focuses on the logical problem of conditionals. And as we will see he uses modal logic to analyze counterfactuals.

We get an idea of what counterfactuals are for Stalnaker from his comment in devising the semantical rule for the selection function:

> The interpretation [of the selection function] shows conditional logic to be an extension of modal logic. Modal logic provides a way of talking about what is true in the actual world, in all possible worlds, or in at least one, unspecified world. The addition of the selection function to the semantics and the conditional connective to the object language of modal logic provides a way of talking also about what is true in *particular* non-actual possible situations. This is what counterfactuals are: statements about particular counterfactual worlds (Stalnaker, 1968, p. 46).

Stalnaker's proposal for evaluating counterfactuals is based on Ramsey's suggestion about how to decide whether or not to believe in a conditional statement. According to Ramsey, if one has no opinion about the truth of the antecedent of a conditional, one should do the following thought experiment:

> Add the antecedent, hypothetically, to one's stock of beliefs or knowledge, and consider whether or not the consequent is true. One's belief about the conditional should be the same as one's hypothetical belief, under this condition, about the consequent.

Ramsey's solution deals only with situations in which one has no opinion about the truth value of the antecedent. In a situation in which we believe the antecedent to be true, but we want to assess the conditional as a whole since we are not sure whether the conditional is true, we do not need to make any changes to our stock of beliefs. However, if one already believes in the truth of the antecedent,

should the opinion about the conditional be different in this case? What about a situation in which you know or believe the antecedent to be false? One cannot simply add it to one's stock of beliefs without introducing a contradiction.

One way to think about Stalnaker's proposal is as a modernization of Ramsey's proposal that handles such cases. For situations where the antecedent of the conditional is believed to be true or false, Stalnaker suggest the following:

> First, add the antecedent (hypothetically) to your stock of beliefs; second, make whatever adjustments are required to maintain consistency (without modifying the hypothetical belief in the antecedent); finally, consider, whether or not the consequent is then true (Stalnaker, 1968, p. 44).

Stalnaker, after considering the belief conditions, goes on to consider the truth conditions of a counterfactual. He says: "Now that we have found an answer to the question, 'How do we decide whether or not we believe a conditional statement?' the problem is to make the transition from belief conditions to truth conditions; that is, to find a set of truth conditions for statements having conditional form which explains why we use the method we do use to evaluate them" (Stalnaker, 1968, p. 33). Kripke's possible world semantics are used since that "is just what we need to make this transition, since a possible world is the ontological analogue of a stock of hypothetical beliefs" (Stalnaker, 1968, p. 33). In addition to possible worlds, Stalnaker introduces in his semantical apparatus:

> *a selection function* (s-function hereafter), which takes a proposition and a possible world as arguments and a possible world as its value. The s-function selects, for each antecedent A, a particular possible world in which A is true. The *assertion* which the conditional makes, then, is that the consequent is true in the world selected. A conditional is true in the actual world when its consequent is true in the selected world (Stalnaker, 1968, p. 45).

The possible world that is taken as an argument of the selection function is called a base world

and the value of the s-function is called the selected world. Thus where f (A, α) = β, A is the antecedent, α is the *base world*, and β is the *selected world*.

Stalnaker proposes the following semantic rule for a conditional (using '>', as the conditional connective):

- A > B is true in α if B is true in f (A, α)

- A > B is false in α if B is false in f (A, α)

Stalnaker goes on to say that the world selected cannot be just any world. He introduces the following conditions on the s-function, using λ to refer to the (unique) impossible world, i.e., the world in which all statements are both true and false:

- For all antecedents A and base world α, A must be true in f (A, α).

- For all antecedents A and base world α, f (A, α) = λ only if there is no world possible with respect to α in which A is true.

The first condition requires that the antecedent be true in the selected world and ensures that tautologies like "If snow is white, then snow is white" are true. The second condition requires that the absurd world be selected only in the case that the antecedent is impossible. Note that according to the fourth condition, this is a specific sort of impossibility. That is, the antecedent is impossible with respect to α.

Stalnaker's account also requires that the selected world differ minimally from the actual

world. This according to Stalnaker implies

> First, that there are no differences between the actual world and the selected world except those that are required, implicitly or explicitly, by the antecedent. Further, it means that among the alternate ways of making the required changes, one must choose one that does the least violence to the correct description and explanation of the actual world (Stalnaker, 1968, p. 46).

These he admits are vague conditions which depend largely on pragmatic considerations for their application. However, they suggest that the selection is based on an ordering of possible worlds with respect to their similarities to the base world. He introduces two more formal constraints on the s-function:

- For all base worlds α and all antecedents A, if A is true in α, then f (A, α) = α.

- For all base worlds α and all antecedents B and B', if B is true in f (B', α) and B' is true in f (B', α), then f (B, α) = f (B', α). (Stalnaker, 1968, p. 47)

These two conditions together ensure that the s-function establishes a total ordering of all selected worlds with respect to each possible world, with the base world preceding all others in the ordering. Stalnaker admits that the conditions on the selection function are "far from sufficient to determine the function uniquely…" (Stalnaker, 1968, p. 47).

One of the main problems in Stalnaker's account is that the s-function picks out one unique possible world even in cases where there is more than one "closest" possible world. What does Stalnaker do to rectify this? We will come to the answer to this question in the next section.

Section 3. Lewis

Lewis begins his book *Counterfactuals* by saying:

> '*if kangaroos had no tails*, *they would topple over'* seems to me to mean something like this: in any possible state of affairs in which kangaroos have no tails, and which resembles our actual state of affairs as much as kangaroos having no tails permits it to, the kangaroos topple over. (Lewis, 1973, p. 1)

Lewis uses the conditional operator '□→' for statements of the form:

'if it were the case that…, then it would be the case that…'

and '◊→' for statements of the form:

'if it were the case that…, then it might be the case that…'

Lewis writes:

> The right general analysis of counterfactuals, in my opinion, is one based on comparative similarity of possible worlds. Roughly, a counterfactual is true if every world that makes the antecedent true without gratuitous departure from actuality is a world that also makes the consequent true…A counterfactual "if it were that A, then it would be that C" is (non-vacuously) true if and only if some (accessible) world where both A and C are true is more similar to our actual world, overall, than is any world where A is true but C is false (Lewis, 1979, pp. 464-65).

Lewis stipulates that if it is not the case that if A were to be true, B might be true, then A ◊→¬B (read 'if A were the case, not B might be the case'). In the event of a 'tie', where worlds where B is true are equally similar to worlds where B is false[7], neither A □→B, nor A □→¬B is true. In such a case, given A, B could be either true or false.

How are these other worlds related to each other, and especially to the base world? The first thing to notice about these worlds, for Lewis, is that they are ordered by similarity with respect to each world. For any two worlds *a* and *b*, *a* is at least as similar to base world *w* as *b* is (not exclusive) or vice versa.

[7] There are two ways that a tie could happen. If there was no closest world to w with A and B true and no closest world to w with A true and B false, then there would be a tie. The other way is if there are closest worlds with A and B, and A and not B respectively true, but these two worlds are equally similar to w.

The relation that Lewis presumes holds on these worlds he calls a ***weak*** *ordering* or a ***(****total****)*** *preordering*. In a footnote he clarifies:

> 'weak' because, unlike a *strong* (or *linear*) *ordering*, ties are permitted: two different things can stand in the relation to each other, and thus be tied in the ordering. 'Preordering' because if we take equivalence classes under the relation of being thus tied, the induced ordering of the equivalence classes is a strong ordering… (Lewis, 1973, p. 48).

He uses an axiomatic definition of "comparative similarity" using the notation '$j \leq_i k$'. This is to be read as the world *j* is at least as similar to the world *i* as the world k is. He also uses the notation $j <_i k$ (defined as it is not the case that $k \leq_i j$). We read this as j is more similar to i than k is (Lewis, 1973, p.48).

The six axioms with which he defines the relation $\leq_i$ are (for any base world i, a set S_i of worlds, regarded as the set of worlds accessible from i, and any worlds a, b and c):

(1) $\leq_i$ is transitive, i.e., if $a \leq_i b$, and $b \leq_i c$ then $a \leq_i c$.

(2) $\leq_i$ is strongly connected, i.e., $a \leq_i b$ or $b \leq_i a$.

(3) $\leq_i$ is reflexive, i.e., $a \leq_i a$. In other words the world i is *self-accessible*

(4) The world *i* is *strictly*$\leq_i$ *-minimal* i.e., for any world a, different from *i,* $i < a$.

(5) Inaccessible worlds are $\leq_i$ *maximal, i.e.,* if *a* is inaccessible then $b \leq_i a$.

(6) Accessible worlds are more similar to *i* than inaccessible worlds. If *a* belongs to S_i and b doesn't, then $a <_i b$

What makes this relation a weak ordering is that it need not be anti-symmetric—i.e., it is possible that $a \leq_i b$ and $b \leq_i a$ and yet $a \neq b$. Condition 1 and 3 are obviously required, since he uses "$\leq_i$" to explain the comparative similarity between worlds. 3 implies that a world that is not self-accessible cannot have the "$\leq_i$" associated with it. 2 is a curious assumption that says that for every two worlds

one of them is at least as similar to *i* as the other. So the similarity relation is presumed to be decidable. This also has the implication that two worlds cannot be incomparable. Condition 4 says that *i* is most similar to itself, and is the unique world to be most similar in this way. 5 implies that inaccessible worlds are least similar to *i*. This condition implies that all inaccessible worlds are equally dissimilar to *i*. Condition 6 states that if there are two worlds, one accessible to *i* and the other inaccessible to it, then the accessible world is more similar to *i* than the inaccessible world.

Section 4. Differences between Lewis's and Stalnaker's Approach

As we can see, there are differences in formulations between Lewis's and Stalnaker's accounts. Lewis devised a triadic relation whose arguments are the propositions A, the actual world α, and a set of worlds. Stalnaker, on the other hand, devised what he calls a selection function going from the pair {A, α} to a unique world—the A-world closest to α.

As Bennett pointed out:

> The differences between relations and functions does not matter here: we could rewrite Stalnaker's theory in terms of a triadic relation [like Lewis's], or Lewis's in terms of a function. What matters is the difference between getting from {A, α} to a class of worlds and getting from {A, α} to a single world (Bennett, 2003, p. 179).

Lewis allows ties in the relative similarity between worlds, as well as no limit on how closely worlds can resemble each other. For Stalnaker, however, there is always exactly one closest similar world. Stalnaker's theory can be seen as a special case of Lewis's that does not make these allowances.

As Edgington points out, besides these differences, there is also a difference in their goals:

> Stalnaker's project is less ambitious. He does not expect there to be an informative analysis of "A-world which differs minimally from the actual world" which could be specified independently of the judgments about what would have been true if A were true. Lewis seeks a genuine analysis of counterfactuals in terms which do not presuppose them (Edgington, 1995, p. 251).

The following are the differences between Stalnaker's and Lewis's accounts with regards to the assumed ordering of the possible worlds: Unlike Stalnaker, Lewis makes the assumption that worlds inaccessible according to base world *i* are worlds that are equally dissimilar to *i* (according to condition 6, in part 2 section 2). On the other hand, the two assumptions that Stalnaker makes and Lewis does not are called the *limit* assumption and *uniqueness* assumption. The limit assumption is the assumption that for every possible world *i* and non-empty proposition *A*, there is at least one *A*-world minimally different from *i*. In other words, there are most similar A-worlds. The uniqueness assumption is the assumption that for every world *i* and proposition *A*, there is at most one *A*-world minimally different from *i* (Stalnaker, 1978, pp. 88-89). The uniqueness assumption rules out ties in similarity. It says that no distinct possible worlds are ever equally similar to any given base world. But according to Stalnaker only the limit assumption, is "reasonable to make" in practice (Stalnaker, 1984 ,p. 133). In contrast, Lewis allows ties in the relative similarity between worlds, as well as no limit to how closely worlds can resemble each other.

These contrasting views are a consequence of their position on whether or not the principle of 'conditional excluded middle' (CEM), which states (A → C) v (A→ ¬ C) for all A and C, holds. This is explained succinctly by Bennett as follows:

> If there is always a unique closest A-world, then at that world C is either true or false; if true then so is A → C [Bennett uses ">" instead of "→"]; if false then A → ¬ C is true. Thus, one of those must be true, and so we have CEM. And if there *are* ties for closest, CEM is *not* universally true. Suppose w1 and w2 are A-worlds that are tied for closest to α. Since they are distinct worlds, there must be some proposition C that is true at w1 and false at w2. So the conditional A → C is false, because w2 is a closest A-world and C is false at it; and A → ¬C is also false, because w1 is a closest A-world and ¬C is false at it. So (A → C) v (A → ¬C) is false in this case because each disjunct is false. (Bennett, 2003, p. 183)

According to Stalnaker, CEM holds relative to a given selection function f. In other words,

CEM does not hold in case where f has not been determined. Notice that the argument against CEM where there is a tie for closest worlds assumes that if C is false at some closest A-world then A → C is false. Stalnaker does not agree with this assumption. According to him, if C is true in some closest A-worlds but not all, A → C is indeterminate—neither true nor false. In such a case the selection function does not select a world.

Thus on Stalnaker's theory the following are neither true nor false:

'if Bizet and Verdi were compatriots, Bizet would be Italian'

'if Bizet and Verdi were compatriots, Bizet would not be Italian'

On Lewis's theory however, they are both false.

As we mentioned in the previous section, one of the main problems with Stalnaker's account, which he recognizes himself, is the following: A world that the s-function picks out must be as close as possible to the actual world. Thus the selection function picks out one unique possible world even in cases where there is more than one "closest" possible world. Consider the following example:

"If my mother had three children,—the first male, and she had more boys than girls, then the second born would have been a male".

Of course this is false. Actually my mother has more daughters than sons. Consider the following two possible worlds: in one the third child is a boy and in two the second child is a boy. In this case it does not make any sense to believe that world one is closer to the actual world than two. But the s-function would have to either select a world of the first kind or the second. If world one is selected, then the following counterfactual will come out true:

"If my mother had given birth to more boys than girls, then the third born would have been a male". And if world two is selected then the original counterfactual will come out true. But one is no better a candidate for truth than two.

Talking about the uniqueness assumption, Stalnaker says:

> It [the uniqueness assumption] says that no distinct possible worlds are ever equally similar to any given possible world. That is, without a doubt, a grossly implausible assumption to make about the kind of similarity relation we use to interpret conditionals, and it is an assumption which the abstract semantic theory I want to defend does make. But like many idealized assumptions made in abstract semantic theory, it may be relaxed in the application of the theory (Stalnaker, 1978, p. 89).

He goes on to say: "To reconcile the determinacy of abstract semantic theory with the indeterminacy of realistic application, we need a general theory of vagueness" (Stalnaker, 1978, p. 89). And the theory of vagueness he uses is the theory of supervaluation developed by Van Fraassen.

Stalnaker's defense, in a nutshell,[8] is the following: it would be highly implausible to rule out ties with regard to similarity between worlds. He does this by adding a structure of supervaluation to the semantics. This supervaluation feeds on s-functions that may differ in how they decide such ties. Such supervaluation may give a truth value to neither of the two counterfactuals, since they would be decided differently by different s-functions. Yet, he holds that conditional excluded middle ($A \rightarrow B \vee A \rightarrow \neg B$) would still be valid, even though in some cases neither $A \rightarrow B$ nor $A \rightarrow \neg B$ would be true. (Kvart, 1986, p.16)

In Stalnaker's words:

> Using the method of supervaluations, we may acknowledge, without modifying the abstract semantic theory of conditionals, that the selection functions that are actually used in making and interpreting counterfactual conditional statements correspond to orderings of possible worlds that admit ties and incomparabilities (Stalnaker, 1978, p. 90).

Stalnaker points out that in using the method of supervaluations, "we are not resorting to an *ad hoc* device to save the theory, since the method of supervaluations, or some account of semantic indeterminacy, is necessary anyway to account for pervasive semantic underdetermination in natural

[8] For a detailed account see Stalnaker, 1981 pp 87-104.

language. Whatever theory of conditionals one favors, one must admit that vagueness is particularly prevalent in the use of conditional sentences" (Stalnaker, 1978, p. 90).

In the case where, for some world, there is no single closest world because of the vagueness of language, Stalnaker unlike Lewis, accepts that if the sentence is vague it is neither true nor false. This analysis better explains linguistic behaviour than Lewis'. Lewis himself admits as much, writing about

the claim that (A □ →B) & (A □ → ¬ B) & (A □ →(B v ¬B)), "… I must admit, it does sound like a contradiction. Stalnaker's theory does, and mine does not, respect the opinion of any ordinary language speaker who cares to insist that it is a contradiction." (Lewis, 1973, p. 80).

Lewis and Stalnaker also differ in their ontological commitments about possible worlds. Lewis is a thoroughgoing realist about possible worlds. Bennett calls him an extreme realist about possible worlds. Other worlds, for Lewis, as different from ours as you please, are as real as ours and actual to their inhabitants as ours is to us. On Lewis' account these possible worlds are metaphysically on a par with the actual world, which is in no way metaphysically special. The other possible worlds are as metaphysically real to the inhabitant of those worlds, as the actual world is to us—the inhabitants of this world. To say that there are no blue swans in this world is to say that in this world in which I live there are no blue swans. But since it is possible that there are blue swans, in some other worlds there are blue swans and an inhabitant of such a world could truly say: "In the actual world there are swans that are blue." In this view 'actual world' means the world that the person in question inhabits.

Bennett offers the following response to Lewis' Extreme Realism which probably expresses the feelings of many:

> Like most philosophers, I cannot believe that corresponding to each different position my right foot could have at this moment there are countless worlds, each of them a real cosmos; though I admit to having no *basis* to my incredulity (Bennett, 2003, p. 153).

But there are philosophical objections that extend beyond mere incredulity. Like Bennett I am persuaded by Robert Adams' moral objection against extreme realism[9].

> I see a child about to wander onto a busy street where she risks being hurt or killed. If I can save her, I should; I am morally required to. However, while I can affect what happens to this girl, I cannot influence the range of what is *possible* for her; the possibilities are laid out rigidly and immovably, and I can only affect which of them is actual. For most of us, that generates a moral imperative; making some possible harm non-actual—that is clearly worth doing! But it seems less so in the context of Lewis's extreme realism, according to which every possibility is real. By saving this child from being hurt, I do nothing to reduce the total amount of pain suffered; the pain that the child *could* have suffered *is* suffered by some child at some world. I merely ensure that the sufferer is not at my world, the one you and I call 'actual' (Bennett, 2003, p. 154).

Stalnaker claims to be a "modest realist" (Stalnaker, 1984, p. 169) about possible worlds. According to him, on this view, at least some counterfactuals are both irreducible and determinately true or false. Counterfactuals have propositional content (they assert true/false things about the actual world) and, he thinks, as long as the meaning of a counterfactual has some propositional content, he is a realist about possible worlds.

Nowadays, though, a philosopher's claim to "realism" is always open to scrutiny, and the starting place for investigation of whether a view deserves that label is usually the work of Michael Dummett. Is a view like Stalnaker's realist or not according to Dummett's analysis? Stalnaker himself tries to answer this question in the last few pages of his book *Inquiry*, with reference to Dummett's own discussion of counterfactuals in *Truth and Other Enigmas* (pp. 145-146). Certainly, at first blush, Stalnaker's view seems to run afoul of Dummett's famous *bivalence criterion* for realism: realism for a discourse implies that all suitably formulated statements in the discourse are determinately true or determinately false; and so by Dummett's lights Stalnaker ought to be counted as an anti-realist about

[9] For a detailed discussion of Lewis's rebuttal, see Bennett, 2003, pp. 153-155.

counterfactuals, to the extent that, as we have seen, he is of the view that some counterfactuals have no determinate truth values.

According to Stalnaker, Dummett provides the following three options for a counterfactual theorist. Either s/he is a naive realist, a reductionist, or an anti-realist. Dummett's argument, in a nutshell, is the following: counterfactuals are either bare or simple truths/falsities; that is, they are true in virtue of themselves and not because of other propositions, or they are true/false because of some other proposition, in terms of which the meaning of the sentence can be restated. An analysis of counterfactuals that asserts that they are "bare truths", Dummett calls "naïve". If the meaning can be restated in terms of some other proposition, then the analysis is "reductionist" (Stalnaker, 1984, pp. 160-169).

Stalnaker argues that this a false dichotomy due to Dummett's equivocation on the word 'bare'. Stalnaker writes, quoting Dummett:

> At one point, bare truth is defined in terms of reducibility. "A statement is barely true if it is true, but there is no class of statements not containing it or a trivial variant of it to which any class containing it can be reduced." But then it is said that "this amounts to holding that we cannot expect a non-trivial answer to the question 'In virtue of what is a statement … true when it is true?" (Stalnaker, 1984, p.161)

Stalnaker asserts that "there is no conflict between the semantic analysis, which says that a counterfactual is true by virtue of the truth of the consequent in some different possible situation, and the realist thesis that a counterfactual is true in virtue of some fact about the actual world." (Stalnaker, 1984, p.163) For, counterfactuals that are analyzed with respect to some possible world have some propositional content. Stalnaker concludes that some counterfactuals are bare truths/falsities and hence irreducible, but complex, and not true only in virtue of themselves.

As Stalnaker points out "Dummett's characterization of realism seems to leave no room for being a realist about some members of a given class of statements and not about others. Dummett seems to assume that if some counterfactuals fail to be barely true or false, then all must…" (Stalnaker, 1984, p. 165). So the question whether he has any anti-realist inklings is still not answered clearly. Stalnaker allows for truth-value gaps and what he calls "superficial" and "deep" sources of truth value gaps. Superficial indeterminacy occurs either when some statements like "Jack is tall" expresses a vague proposition or when "there is some indeterminacy in the relation between the sentence (as used in a given context) and the proposition it expresses" (Stalnaker, 1984, p. 166). But "the indeterminacy in counterfactuals seems deeper than this", Stalnaker asserts (Stalnaker, 1984, p. 165).

About deep indeterminacy, Stalnaker says that it is conceivable, in some particular cases, that with regard to the counterfactual antecedent there is indeterminacy in the relation between the sentence and the proposition it expresses in the actual world, but its semantic determination, via possible worlds, is straightforward. In such a case, we can regard the meaning of the antecedent of the counterfactual as determining a *space* of possible worlds and "…for each possible world in the space, the [antecedent] proposition is true or false. But when we ask, is the statement true in the actual world, we find the question has no answer for the following reason: there are no facts which determine which of several points (some on each side of the sharp line) is the actual world. Our conceptual space of possibilities, I am supposing, has cut things up too finely, making distinctions to which nothing really answers." (Stalnaker, 1984, p. 166).

Stalnaker admits that the distinction between deep and superficial semantic indeterminacy is clearer in the model than in its application. However its intuitive basis is evident from the following example Stalnaker offers:

Consider Tweedledee and Tweedledum, again. They had a coin but before it was tossed someone ran off with it. Having no other coin Tweedledee and Tweedledum argued about how it would have landed if it had been flipped. Tweedledee is convinced that it would have landed heads and Tweedledum is convinced that it would have landed tails. Neither has a reason for ─his conviction. They agree that the coin was normal and the toss was fair. There is little inclination in such a situation to say that one of them must be right. We may think its absurd to assume that the counterfactual "if the coin had been tossed it would have landed heads", is either barely true or barely false, but Tweedledee and Tweedledum obviously make this assumption. They not only make apparently conflicting statements, they have genuinely conflicting beliefs. Their conflicting beliefs may make contrasting actions rational, and may lead to different attitudes of other kinds. Maybe Tweedledee believes that because the coin would have landed heads, he will have an unlucky day, in which case he ought, given his beliefs, to be cautious. Perhaps Tweedledum, although he believes that the coin would have landed tails, wished that Tweedledee had been right.

In the above example, the counterfactual "if the coin had been tossed it would have landed heads", the question whether the statement is true in the actual world, remains unanswerable since there are no facts which determine which several points is the actual world. In such situations according to Stalnaker "[W]e may need a distinction between possible situations in which the coin would have landed heads and possible situations in which the coin would have landed tails in order to give an adequate description of attitudes of Tweedledee and Tweedledum, even if we don't need or want such a distinction in order to locate the actual world in a space of possibilities" (Stalnaker, 194„ p.167). Thus for purposes of psychological explanation we need the finer grained space of possibilities "to which nothing really [i.e. actually] answers."

Here, then, we find Stalnaker's case for realism in spite of the failure of bivalence for

counterfactual claims. Many counterfactuals are *barely true* in the sense of not being reducible to another class of statements, but *not* in the sense of failing to be *true in virtue* of something. They are *true* by virtue of the truth of the consequent in a different possible situation. But not all counterfactuals are barely true or false in this way, because the meaning of an antecedent may fail to pick out the relevant alternative (possible situations). Nevertheless, in either case the result is that possible worlds are indispensable. This indispensability is what Stalnaker points to when asked what makes him a realist about possible worlds. This is what seems to be implied when Stalnaker claims to be a modest realist. Thus we see that metaphysical consequence of the respective accounts of Stalnaker and Lewis, make them a modest realist and an extreme realist, respectively, about possible worlds.

Some critics of possible worlds maintain that the metaphysical baggage of having to take possible worlds somewhat seriously is an unwelcome price to pay. One such critic is Kvart. He developed a syntactic account of counterfactuals, in the post-Kripke era trying to avoid metaphysical consequences of possible world accounts.

Section 5. Kvart

Nowadays, in the post-Kripke era, almost everybody in the field advocates some or other sort of possible worlds account. Igal Kvart is an interesting exception to this. Kvart in his award winning book *A Theory of Counterfactuals* argues that the metaphysical price of possible worlds semantics is too high, and that a more sophisticated version of the earlier "syntactic" views can share the benefits of the possible worlds accounts.

We have seen that, generally speaking, a Chisholm-Goodman type of analysis, provides truth conditions for conditionals in terms of the following:

$p \rightarrow q$ is true if q follows by law from p together with a set Γ of true sentences such that for $r \in \Gamma$ it is

not the case that $p \rightarrow \neg r$.

Given that such accounts provide truth conditions for conditionals in terms of the truth conditions of other conditionals, it involves an infinite regress. Breaking free of this regress requires providing an independent characterization of Γ. Kvart makes a sophisticated attempt to do just that. He revives certain aspects of the pre-Kripkean approach because of worries about the metaphysical baggage that seems to come with the Lewis-Stalnaker account.

I will explain the basic idea in Kvart's account with the following example:
Suppose Poonam lived in India, and during her Master's in India she got a scholarship to study at a Canadian University. She came to Canada where subsequently she met Jason and made a family. While studying in Canada she got a job offer in India and realized that "Had I (Poonam) not met Jason and begun a family in Canada, I (Poonam) would have gone back to work in India".

How do we analyze such counterfactual statements? Kvart explains the truth of such counterfactuals, when they are true, as follows: For a counterfactual to be true the consequent must be inferable from the antecedent and some implicit premises. Now the question is what those implicit premises are. It seems obvious that these implicit premises must include some part of the world's history prior to the antecedent. But that does not exhaust the premises since no adequate analysis will make the statement in question come out true unless there is a premise to the effect that Poonam actually got a job offer from India. However, this event (Poonam getting a job offer) does not belong to the prior history of the world (the world in which the antecedent event was supposed to take place), but to the time after it. So, the question still remains—what are these implicit premises?

The important question to be asked according to Kvart is: what is special about the actual event of Poonam getting a job offer from an Indian University vis-à-vis the antecedent event of Poonam not meeting Jason? Kvart's proposal for analyzing counterfactual statements such as this one is the

following: the occurrence of the antecedent event (Poonam not meeting Jason) would not "endanger" or "would not put at risk" the occurrence of the job offer. "And it is the statements describing such events, whose occurrence is not "risked" by the antecedent event that we would want to retain as implicit premises out of the history of the world from the time to which the antecedent pertains onwards" (Kvart, 1986, pp. xi-xii).

What do these notions of "endangering" or "putting at risk" amount to? According to Kvart they imply that the antecedent event (of Poonam not meeting Jason) bears some negative causal relevance to such events (of Poonam getting a job offer). According to Kvart, we want to preserve all true statements pertaining to times later then the antecedent event that describes events to which the antecedent event bears no negative causal relevance. Not bearing negative causal relevance amounts to either causal irrelevance or pure positive causal relevance. In other words, we want to include in the implicit premises all true statements describing events that pertain to times later than the antecedent event and especially/particularly the ones that describe the events that are not denied or annihilated but may be positively affected by the antecedent event. Going back to our example—"Had Poonam not met Jason and had a family in Canada, she would have gone back to work in India", we want to preserve all events not endangered by the truth of the antecedent. And *that's* the reason why the statement about the job offer should be preserved, presuming it isn't endangered. In other words, we want to include in our implicit premises true statements such as, "Poonam gets a job offer", "Poonam didn't write nasty letters to her prospective employers in India", etc. (These are events to which the antecedent bears no negative causal relevance). Kvart adds an extra clause that we also want to preserve laws of nature.

This is how Kvart analyzes counterfactuals whose antecedents are compatible with the prior history of the world. He explains the notions of "causal relevance" and "purely positive causal

relevance" in terms of probabilistic analysis using a notion of objective conditional probability.

The requirement that besides including true statements (pertaining to time later than the antecedent event) that are either causally irrelevant or else purely positively causally relevant, among our implicit premises, we also want to preserve laws of nature, limits the scope of Kvart's account. Since in evaluating counterfactuals with impossible antecedents we may encounter physically impossible antecedents, and such antecedents would seem to have negative relevance for some or other laws. So, while I think Kvart's account would be worth deeper investigation if my goal were a general account of counterfactuals, through the remainder of the thesis I shall follow the mainstream by presuming that a possible worlds account is the state of the art, and so my proposed revisions will be revisions to that sort of account. This will leave an interesting question of whether similar revisions could be made to Kvart's account to achieve the same ends I achieve in Chapter IV and Chapter V, though I will not pursue that question in this thesis.

Section 6. A Generalized Lewis-Stalnaker Account

Differences in formulation and ontological commitments aside, I will for the remainder of this thesis adopt a generalized Lewis-Stalnaker style analysis and build on this generalized account, borrowing (and developing) some ideas from contemporary semantics for Relevance Logic, to construct a new theory of counterfactuals.

The generalized Lewis-Stalnaker account I will use is the following: In analyzing a conditional of the appropriate sort (i.e., P $\Box\!\!\rightarrow$ Q), one checks the *closest* possible worlds where the antecedent (P) is true, and if in all those possible worlds (where P is true), the consequent (Q) is true as well, the conditional in question is true.

Chapter III: Conceivability, Possibility and The Notion of A Case

In this chapter I investigate the notions of conceivability, imaginability, and possibility with a special purpose. The eventual goal is to isolate the notion of a case—a notion that will play a role in our new semantics for counterfactuals that is similar to the role played by possible worlds in Lewis-Stalnaker type accounts.

As we have seen, there are similar problems with analyses of thought experiments and counterfactuals, and there are similarities in the structures. There is a fairly straightforward way to think about a link between counterfactuals and thought experiments . Recall the three parts of the description of a thought experiment, described in section 2, Chapter I—scenario description, argument for X being the correct evaluation, and X being taken as the lesson (i.e., what the argument reveals about the real world). Now, consider a counterfactual with an antecedent that is a collection of sentences that describes the scenario, and consequent being the X. The argument then can be regarded as making a case that this particular counterfactual is true. So having an account of what makes counterfactual true or false gives us a starting point for explaining how thought experiments work, and perhaps even a way to evaluate the quality of the arguments.

However, if the thought experiment in question considers an impossible scenario, the counterfactual we end up with, using this process, is worthless. Since, as we have seen, the current accounts of counterfactual do not offer satisfactory explanations of such counterfactuals. So we need to develop an account of counterfactuals that is more comprehensive and is able to systematically handle counterfactuals with impossible antecedents. This chapter is a step in this direction. As we will see in this chapter, examining the notion of conceivability is a key to understanding the notion of case. Cases are crucial to our core project of developing a new semantics for counterfactuals.

There are two apparent advantages to expressing thought experiments in terms of counterfactuals, especially for incomplete thought experiments. First, in case of a thought experiment with an incomplete scenario, when we express it in terms of a counterfactual, the antecedent, of course, turns out incomplete. But as we know, all counterfactuals have incomplete antecedents. The literature on counterfactuals, as we have seen earlier in the thesis, includes some persuasive accounts about what determines the truth values of counterfactuals, in spite of their antecedents not completely describing the alternative possibilities relevant to determining the truth values. By expressing thought experiments in terms of counterfactuals we can overcome the problem that incompleteness poses for thought experiments. So, the main advantage of analyzing thought experiments in this way is that it makes it possible to exploit the same accounts of counterfactuals to explain the truth conditions of claims about thought experiments.

Another advantage of expressing thought experiments in terms of counterfactuals is that there has been a significant development in the theories of counterfactuals, especially in the post-Kripke era. So, we can look at various theories of counterfactuals to see if we can use them to analyze incomplete/unimaginable, yet useful thought experiments.

It is worth mentioning here that, of all the different kinds of thought experiments that are incomplete, my focus is on the kind of thought experiments that are incomplete because there is some inconsistency, contradiction, impossibility or other incoherence built into the specification of the experimental set up. What do we do with these special kinds of thought experiments? In Chapter IV, we will see how, by using semantics borrowed from Relevance Logic and a Lewis-Stalnaker style account of counterfactuals, and by building on the borrowed semantics, we can make sense of such incomplete and impossible scenarios. We will be in a position to see why incompleteness and impossibility do not always hamper a thought experiment's informativeness/usefulness. We will also

find out, by looking at theories of counterfactuals, how we can prevent everything and anything following from an impossibility. This comes in handy in explaining why from some important thought experiments such as Strawson's auditory world, everything does not follow. Notice that the auditory world is a special case not only in that it is incomplete, but in that it is also impossible. This is what is in store for Chapter IV.

This chapter is divided into seven sections. Section 1, includes the preliminaries that set the stage for what is to come. Section 2 is a brief historical discussion of the two notions of imaginability and conceivability. In section 3, we discuss different kinds of possibilities viz., epistemic, logical, metaphysical, and nomological possibilities and some key issues regarding these various kinds of possibilities. Section 4 is an overview of the issues regarding conceivability and possibility. This section provides a glimpse of the recent discussions on these topics against a Kripkean backdrop. Section 5 also includes a detailed discussion of the concepts of imaginability and conceivability. In this section I argue that there is a distinction to be made between conceivability and imaginability, and that there are three layers to these concepts viz., *imaginable*, *conceivable*, and *fully and coherently describable.* In section 6, I consider the much anticipated question "Is conceivability a guide to possibility?" and argue that it is not. Section 7 is where all this work pays off. I will introduce the concepts of *states* and *cases*, and argue that while conceivability is not a guide to possibility, philosophers have been right to pay close attention to it: it is a guide to *being a case*, and, as we will see in the next chapter, being a case is an important thing indeed.

Section 1. A Few Preliminaries

First, I want to explain what I mean by usefulness of a thought experiment. A thought experiment is useful insofar as we can learn some lesson/s from it, philosophical or otherwise. In

other words, a thought experiment is useful if it is informative about the subject matter in question. For example, Galileo's thought experiment is useful because we learn from this thought experiment that the natural speed with which a body falls is independent of its weight. The Theseus' ship thought experiment informs us that identity, at least for some objects, has to do with something over and above bodily identity.

What I want to show here is that when it comes to useful but impossible thought experiments, we get the "information" by figuring out that a particular counterfactual is true (or some class of counterfactuals are). Consider fission for example: In order to see how fission is informative, we express this in terms of a counterfactual or a series of counterfactuals, one possible counterfactual might look something like the following: "if it were the case that your brain could be transplanted into another body nearly identical to your own, and that either the left or right hemisphere of the brain would be sufficient to support full preservation of all mental properties, and that your right and left hemispheres are successfully transplanted into two different bodies, such that each has all of your memories, beliefs, desires, etc., and that both transplants are successful.…, then you would be identical to neither continuer". This is just one counterfactual and it is not clear whether this counterfactual is true or false. Another counterfactual about fission could be: :"if it were the case that your brain could be transplanted into another body nearly identical to your own, and suppose further that either the left or right hemisphere of the brain would be sufficient to support full preservation of all mental properties, and that your right and left hemispheres are successfully transplanted into two different bodies, such that each has all of your memories, beliefs, desires, etc.…, then you would be identical to both continuers". A third possible counterfactual would be: "if it were the case that your brain could be transplanted into another body nearly identical to your own, and suppose further that either the left or right hemisphere of the brain would be sufficient to support full preservation of all

mental properties, and that your right and left hemispheres are successfully transplanted into two different bodies, such that each has all of your memories, beliefs, desires, etc., and only one transplant is successful…., then you would be identical to one of the continuers". We may need to consider a couple of possible counterfactuals with the same antecedent but different consequents depicting the possible outcomes to figure out which of these counterfactuals is true. After figuring out which counterfactual is true we will be in a position to say what information the particular thought experiment is able to give us. And the philosophical debate is fruitfully considered a debate about which counterfactual is true, perhaps?

Again consider the auditory world: One possible counterfactual is the following; "If there is such a being as Hero, who resides in a completely auditory and non spatio-temporal world with only sound fragments that could serve as objective particulars, and Hero has one and only one mode of outer sense, viz., hearing, and he distinguishes between different sound particulars against the background of "universal sound" or "M-sound", then in such a world objectivity would amount to…". Again we may need to consider a few other counterfactuals with the same antecedent but different possible consequents to see what the thought experiment can teach us.

In both the auditory world and the fission example, and other similar ones which are incomplete yet useful, one can extract the relevant "information" by translating the particular thought experiment into a series of counterfactuals and figuring out which particular counterfactual/s is/are true. In some cases many counterfactuals may be true, especially in cases where many lessons may be learnt. Such counterfactuals are special cases in so far as their antecedents are impossible. Similarly one can also extract information from thought experiments that turn into counterfactuals with false but possible antecedents. However, as stated in Chapter I, classical logic, and in particular in accounts of counterfactuals built on classical logic, the first sort pose a special problem in that anything and

everything follows form an impossible antecedent. Relevance Logic, as we will see in Chapter IV, when coupled with Kripke-Stalnaker style semantic machinery, provides us with the tools to explain why any and every counterfactual with an impossible antecedent is not true. Thus with the help of Relevance Logic we can find out why only certain counterfactuals with impossible antecedents are true and what this can help us analyze a certain class of thought experiments so that we are in a position to accommodate thought experiments that are impossible and incomplete.

I should mention here that initially it seems incompleteness and impossibility go hand in hand, at least for the cases we look at in this thesis. In other words, it seems that precisely the feature that makes a thought experiment impossible also makes it incomplete. That feature is inconceivability. As we will see later in section 5, this is not the case—impossibility does not entail inconceivability. Amongst other things, we need to consider in what ways unimaginability is tied to impossibility. In other words, do things that are inconceivable also turn out to be impossible or vice versa? The question that is more commonly asked is the contrapositive: Is conceivability a guide to possibility? We will see that in asking this question we get our answer regarding whether unimaginability is tied to impossibility.

Section 2. Historical Perspective of Conceivability and Imaginability

In this section we will be dealing with the following two pairs of concepts: imaginability/unimaginability, and conceivability/inconceivability. Considering these two notion from a historical perspective is important because, as we will see, philosophers such as Descartes draw a sharp distinction between conceivability and imaginability, whereas Hume uses the two notions interchangeably. Not making a distinction for Hume has some unpalatable consequences, as we will see. We can learn some lessons from history about why we ought to make a distinction between these

two notions.

Generally speaking, is there a distinction between imagining and conceiving? If there is, then what is the nature of this distinction? Prima facie it seems that imagining is imagistic, and sensory in nature whereas conceiving is non-imagistic and conceptual. So what is the connection between imagining and conceiving? More generally "one might wonder about the relation between imagination/conception, on the one hand, and perception/intellection, on the other: is the former parasitic on the latter…?" (Gendler, 2002, p. 9).

We can trace the distinction between conceiving and imagining back to early modern period, particularly to Descartes and Hume. The distinction between imagining and conceiving is clearly expressed in Descartes'' letter to Messene where he writes: "whatever we conceive without an image is an idea of the pure mind, and whatever we conceive with an image is an idea of the imagination" (Descartes vol. III, p. 186). Descartes draws a sharp distinction between intellection and understanding, on the one hand, and imagination and sensation on the other. While intellection and understanding are cognitive faculties that belong to us *essentially qua* thinking things, imagination and sensation are limited cognitive faculties that belong to us *contingently qua* embodied beings (Descartes, vol. II, p. 51). As Gendler and Hawthorne point out, these faculties, according to Descartes, not only differ in their range of subject-matter in that imagination is "nothing but an application of the cognitive faculty to a body which is intimately present to it" (Descartes, vol. II, p. 50), they also differ in their phenomenology. For, Descartes writes in the Sixth Meditation: "When I imagine a triangle, for example, I do not merely understand that it is a figure bounded by three lines, but at the same time I also see the three lines with my mind's eye as if they were present before me" (Descartes, vol. II, p. 50).

Hume, unlike Descartes, does not maintain a sharp distinction between conceiving and imagining and conflates the two terms. For, he writes:

> 'Tis an establish'd maxim in metaphysics, That whatever the mind clearly conceives, includes the idea of possible existence, or in other words, that nothing that we imagine is absolutely impossible [emphasis my own] (Hume, 1968, p. 32)

It seems to me that Hume is making two claims here: (a) what we imagine or conceive is *presented* as possible and (b) what we imagine or conceive *is* not absolutely impossible. What (a) claims is that the relation between conceivability and possibility is internal, whereas what (b) claims is that there is an external relation (mind to possible world, if you will) between conceivability and possibility. (b) amounts to claiming that what the mind conceives/imagines is *in fact* not absolutely impossible or possible in some states of affairs. (b) is a much stronger claim than (a), what we imagine/conceive is *presented* as possible (to our mind). We can show that (b) is a much stronger claim than (a) by considering conceivable impossibilities. If we were to be able to conceive of impossibilities, would they be *presented* as possible (as (a) requires) or would they possible in some states of affairs? The latter would be an absurd claim to make because according to this claim whatever the mind can conceive *is not absolutely impossible or possible in some states of affairs,* but what we are talking about is conceivable impossibilities. How can impossibilities be also possible in any sense? Although these claims, viz., (a) and (b), might not be directly related to the distinction between conceiving and imagining, for Hume the question whether conceivable impossibilities are only *presented* as possible arises, one might say, because he fails to make a distinction between conceivability and imaginability. This is because ordinarily an act of imagination presents the imagined entity as possible. If imagining = conceiving, then conceivable impossibility = imaginable impossibility. Thus making conceivable impossibilities *presented* as possible.

This is a brief overview of the history for the distinction between conceiving and imagining. The natural progression would be to consider whether there is any distinction to be made between

conceiving and imagining and what my take is on this distinction. However, before going on to discuss the distinction between conceiving and imagining, I shall consider the following questions: What kinds of possibilities/impossibilities are there? What is the relationship between conceivability and possibility? There are two reasons for this diversion: first I want to consider the notion of possibility/impossibility separately from conceiving and imagining. As we will see there is a natural tendency to assume that conceiving and possibility are so closely tied together that being able to conceive x entails x's possibility. And secondly, the notion of possibility/impossibility is closely tied to conceiving in such a way that we have to sort out the different kinds of possibilities/impossibilities before answering the question whether conceiving is a guide to possibility. In other words, the question becomes what kind of possibility is conceivability a guide to, if any?

Section 3. Different Kinds of Possibilities and Some Key Issues

In their recent work, Gendler and Hawthorne divide possibility into two broad classes, viz., epistemic possibility and non-epistemic possibility. Epistemic possibility is "defined relative to some subject (or sets of subjects) in terms of some body of knowledge or evidence available to (or otherwise associated with) the subjects(s) in question" (Gendler and Hawthorne, 2002, p. 3). There are two kinds of epistemic possibility----*strict* and *permissive*. On the *permissive* account,

P is epistemically possible for S just in case S does not know that not-P.

On the *strict* account,

P is epistemically possible for S just in case P is consistent (metaphysically compossible) with all that S knows. (Gendler and Hawthorne, 2002, p. 3).

There are important differences between these two accounts, as Gendler and Hawthorne point out. While on the strict account epistemic possibility entails metaphysical possibility, on the

permissive account it does not.

Non-epistemic possibility is the other kind of possibility. They distinguish three kinds of non-epistemic possibility viz., logical, metaphysical, and nomological possibility, which they define in the following way:

(1) **logical**: P is logically possible just in case no contradiction can be proven from P using the standard rules of deductive inference.

(2) **metaphysical**: This can be defined in terms of a primitive notion along the lines of "how things might have been"/"how God might have made things". In possible world terminology, it is actual that P just in case P obtains in the actual world. It is (metaphysically) possible that P just in case P obtains in some possible world.

(3) **nomological**: P is nomologically possible for a relevant body of laws just in case P is logically consistent with the body of truths implied by those laws (for example, if we consider the true laws of Physics, P is possible if P is consistent with the laws of physics).

In other words, P is nomologically possible just in case S is not ruled out by the (logically contingent) laws of nature (that counts as the relevant body of laws in that context). So it is nomologically impossible for two objects separated by a finite distance not to exert gravitational force on each other, but it is nomologically possible that there are ponds of ice cream.

According to Gendler and Hawthorne, generally speaking logical possibility is the broadest sort of possibility. According to their definition, a proposition is said to be logically possible if there is no logical contradiction involved in its being true. "George Bush is a bachelor" is logically possible, although it is actually false. "George Bush is a married bachelor" is on the other hand logically impossible. Most philosophers have thought that statements like "If I flap my arms very hard, I will fly" are *logically* possible, although they are *nomologically* impossible.

Now let us consider some details about these possibilities. First let us consider metaphysical possibility and logical possibility. In this context consider the following quote from Gideon Rosen's "Modal Fictionalism Fixed".

> In the first place, the feature of the standard semantics for counterfactuals which Hale's objection exploits is plausibly regarded as a *defect* in that analysis. As Hartry Field has observed in another context, we do seem to be able to make discriminating use of counterfactuals whose antecedents we suppose to express necessary falsehoods (Hartry Field: *Realism, Mathematics and Modality*, (Blackwell 1989).[3], pp. 237-8): if arithmetic were inconsistent, set theory would be inconsistent; if the God of the philosophers (i.e., a perfect, necessary being) existed, the righteous would have nothing to fear; if the Queen were your mother, Diana would be your sister-in-law. There may be no good systematic semantics for counterfactuals of this sort. But this does not mean that they don't make sense, or that a philosopher may not avail himself of them in trying to explain his view. The significant feature of these examples is that the alleged impossibilities supposed in the antecedents are not logical impossibilities. They are substantive impossibilities, metaphysical or mathematical; and while there may be insuperable obstacles to making sense of counter-*logical* conditionals, conditionals whose antecedents are impossibilities of these substantive sorts seem much better behaved—as indeed we all tend to acknowledge whenever we explore the consequences of a metaphysical or mathematical view we in fact reject (and so, presumably, regard as impossible) by saying such things as 'If that were true, then this would be true; but this is absurd; so that must be false.' (Rosen, 1995, pp. 70-71).

So, according to Rosen, there are two kinds of impossibilities viz., logical and substantive. Substantive impossibilities include at least two kinds, viz., metaphysical and mathematical. It is apparent from the above that Rosen subscribes to the view that there is a distinction to be made between the two kinds of substantive possibilities, viz., mathematical possibility and metaphysical possibility. I would like to mention here that I will ignore the distinctions between the two sorts of substantive possibility (mathematical and metaphysical). It is worth pointing out in this context that Rosen thinks that there may be no good systematic semantics for mathematically impossible counterfactuals such as “if arithmetic were inconsistent, set theory would be inconsistent”. There is a systematic semantics for counterfactuals of the sort mentioned above, as we will see in Chapter IV.

In the context of possible worlds the modality in question is “metaphysical” possibility. Metaphysical possibility is either equivalent to logical possibility or distinct from it, depending on the philosopher’s view. Following Kripke, some philosophers hold that there is a distinction to be made between these two kinds of possibilities. In statements like "Water is H_2O”, which they think are metaphysically necessary, ‘water’ functions as a demonstrative. So, ‘water’ is a rigid designator. A term *t* is a *rigid designator* if *t* picks out the same thing in all possible worlds, in which the thing exists. Now since, ‘water’ is a rigid designator, and since the chemical composition of water is H_2O, and it is an essential property of water, it is metaphysically necessary that "Water is H_2O". That is, ‘water’ picks out the stuff with the chemical composition of H_2O in all possible worlds. However, although it is metaphysically necessary that “Water is H_2O”, it is not logically necessary that “Water is H_2O” as there is no formal contradiction involved in saying "Water is *not* H_2O". Thus according to them, "Water is *not* H_2O" is logically possible, even though it turns out to be metaphysically impossible.

Pre-Kripke, philosophers in the opposite camp considered logical and metaphysical necessity to be coextensive. They would deny that “Water is H_2O” is necessary at all. Claims like “Brain states are mental states”, “Water is H_2O” were regarded by most philosophers in the pre-Kripke era as *contingent identities*, since they are clearly not logically necessary, and since the only sensible notion we can attach to the phrase “metaphysically necessary” is “logically necessary”.

In this context, consider the following claims:

(1) H_2O contains hydrogen atoms.

(2) Water contains hydrogen atoms.

Alexander Pruss discusses how by using a Kripkean account of natural kind names, one can defend a claim that one cannot use the distinction between logical and metaphysical necessity to

distinguish between (1) and (2)[10]. For those who maintain a distinction between logical and metaphysical necessity, some propositions, such as (1) are *logically* necessary since it is logically necessary that anything that has two atoms of hydrogen and one atom of oxygen in each molecule (and that, after all, is the definition of "H_2O") contains hydrogen atoms. (2) is only metaphysically necessary since it does not follow from the logic of the terms. Similarly, "Water *is* H_2O" is not logically necessary, since it does not follows from the logic of the terms. Philosophers in this camp would claim that (1) and (2) have different modal statuses because they express different sentences.

Those who claim that there is no distinction to be made between logical and metaphysical necessity would reply by saying that they cannot have different modal statuses, because modal status belongs to propositions, not to sentences, and (1) and (2) express the same proposition, and hence have the same modal status. If following Kripke we say that "water" in (2) functions as a demonstrative pointing to the natural kind of that paradigm body of water that was involved in a Kripkean baptism, and if that natural kind just is H_2O, then the difference between (1) and (2) involves no change in modal status. Thus one cannot use the notions of logical necessity and metaphysical necessity to distinguish between (1) and (2). This would imply that on this view the claim "Water is H_2O" is logically/metaphysically necessary.

I will not go into details about modal status of sentences vs. modal status of propositions. It is sufficient for the purpose of our discussion to say that we will take the metaphysical and logical necessity to be distinct

Now let us consider the relationship between metaphysical possibility and nomological possibility. In this context let us consider a recent view developed by Edgington on metaphysical possibility. In "Two Kinds of Possibility" Edgington argues that metaphysical necessity "derives from

[10] I have relied on Pruss, A.R.: *Possible Worlds: What They Are Good For and What They Are* http://www9.georgetown.edu/faculty/ap85/papers/PhilThesis.html#_Toc515941274).

a modal concept we all use, in distinguishing things which can happen and things which can't, in virtue of their nature, which we discover empirically" (Edgington, 2004, p. 1). Thus she claims that metaphysical possibility is constrained by laws of nature. This view, according to her, is Kripkean in spirit and is a consequence of a natural reading of what Kripke says about metaphysical possibility and necessity. As we have already seen, on a Kripkean interpretation, statements like "Water is H_2O", are metaphysically necessary. "Water is H_2O" is metaphysically necessary because among the essential properties of water is that it's chemical composition is H_2O. In case of "Water is H_2O", laws of chemistry (a law of nature) dictate that water cannot be something other than H_2O.

To gain a clear understanding of Edgington's claim, it is important to understand a traditional debate about the modal status of laws of nature. Following Hume, many philosophers hold the view that the laws of nature are *metaphysically contingent*—that there could have been different natural laws than the ones that actually obtain. If this is the case, then for example it would be both *logically* and *metaphysically* possible that the law of gravity does not hold. So, for example, apples might not fall from a tree but would float in mid air. However, it seems that there is an important sense in which this is *not* possible; *given* that the laws of nature are what they are, there is no way that apples would float.

While this is certainly the more prevalent view, there have been dissenting voices. Sidney Shoemaker, for instance, argues that natural laws are in fact *necessary*, not contingent. On this view, nomological possibility is equivalent to metaphysical possibility. Edgington agrees with Shoemaker in holding that laws of nature are necessary in some serious sense. She points out, in "Two Kinds of Possibility", that there has been a competing tendency in Philosophy to either assimilate laws of nature with necessary truths or to differentiate laws of nature from necessary truths. And before Kripke, in so far as the *a priori* and the metaphysically necessary were not separated, this dilemma

remained unresolvable. In Edgington's words:

> Very roughly, the prevalent philosophical view before Hume, correctly perceiving that such statements have the mark of necessity, was that we have to think of them as somehow deducible from self-evident truths about how the world must be—if we only knew how. Hume showed definitively that reason was not up to the task. The prevalent view since Hume, recognizing that they can't be known *a priori*, concludes that they are contingent regularities….Only with Kripke's separation of the *a priori* and the metaphysically necessary is there room for the view that the pre-Humeans were right in thinking that laws of nature are necessary but wrong in thinking they are knowable *a priori*, and the post-Humeans made exactly the opposite mistake. *If* Kripke has shown that there is a class of necessary truths which can be known empirically, and *if* there are reasons for treating laws as necessary in some serious sense, why should this class not be the natural home for natural laws? (Edgington, 2004, p. 3)

Edgington offers two reasons for treating natural laws as necessary in some serious sense. First, "the modal idiom is the natural reason for distinguishing laws from accidental generalizations. Nothing *can* travel faster than light. These plants *can't* be grown under freezing temperatures. These other plants, merely, never are, in the history of the universe, grown at freezing temperatures, although they could have been" (Edgington, 2004, p. 3). The second reason is that, according to Edgington, most neo-Humean theories trying to distinguish between natural laws and necessity in some serious sense, fail. The best known neo-Humean theory which she calls "'Mill-Ramsey-Lewis'" theory maintains that laws are those true contingent generalizations that occur as axioms or theorems in the true deductive system that achieves the best combination of simplicity and strength, by our standards of simplicity and strength. Now, according to Edgington, the problem with this theory is that although we may very well like better a world in which laws fit into nice systems and we may have acquired reasons to think we live in such a world, it is far from obvious that our concept of law precludes the possibility of relatively isolated laws governing the workings of their own subject matter, not clashing, but not exactly cohering either. Conversely, as Edgington points out using Bas van Frassen's example, might there not be highly informative and simple generalizations which are not laws?

Consider a world which contains just two kinds of objects: iron cubes and gold spheres—whizzing around according to Newtonian mechanics. Suppose also that it just so happens that there were no collisions, although there could very well have been, that altered the shapes of these objects. It is hard to deny that in such a world "all and only cubes are iron" and "all and only spheres are golden" add a lot of informational content to the description of this world, at little loss of simplicity. But they are not laws.

As we can see, this debate regarding whether natural laws are necessary in some serious sense has bearing on our definition of metaphysical possibility and nomological possibility. If metaphysical possibility is constrained by laws of nature then what is metaphysically possible is also nomologically possible. There is definitely something to be said about laws of nature being necessary "in some serious sense" in that they are not merely accidental but uniform regularities. In the context of this thesis one consequence of collapsing the lines between metaphysical possibility and nomological possibility is the following: As I shall argue, *fully and coherently describable* implies *possible.* If this is so then it yields a surprising result, given that metaphysical possibility and nomological possibility amounts to the same as on the Edgington-Shoemaker view. How can something about language or our linguistic capabilities (i.e., x being fully and coherently describable) be a guide to laws of nature (i.e. , x being nomologically possible)? I agree with Edgington that laws of nature are necessary "in some serious sense". However, I don't think that metaphysical necessity is constrained by laws of nature. So my eventual view is compatible partly with both approaches, because I concur with the Edgington-Shoemaker view in that laws are necessary is some serious sense and I also concur with the Mill-Ramsey-Lewis view that nomological possibility and metaphysical possibility are not one and the same.

Going back to van Frassen's examples of a world which contains just two kinds of objects,

iron cubes and gold spheres, I would like to say the following: I concur with Edgington in saying that, in such a world, statements like "all and only cubes are iron" and "all and only spheres are golden" accurately describe the world and these statements are not laws. Facts about our language do not determine laws of nature; Edgington would agree, since she thinks that generalizations like the ones mentioned above are not laws. Although it is accurate in our world to say that "all apples that fall off the tree, fall downwards", it is not a law itself. The law behind this generalization is the law of gravitation. All generalizations that hold uniformly are not alike. Generalizations like "all and only cubes are iron" in a world where all objects are iron cubes and gold spheres are the same as generalizations like "all ravens are black". However, generalizations like "all objects fall upon being dropped" explain cause and effect. Laws are what explains such causal generalizations. So, I agree with Edgington in that not all generalizations, even those which hold uniformly in a given world, are laws. After all, such uniformity may in fact be accidental.

Now, let us consider the relationship between metaphysical possibility and epistemic possibility. It is impossible for somebody to square the circle. This is metaphysically impossible. However, it was only in 1882 that the task was proven to be impossible, as a consequence of the fact that "pi" is not a algebraic but a transcendental number. It had been known for some decades before then that *if* pi *were* transcendental then the construction would be impossible, but that pi *is* transcendental was not proven until 1882. Although before 1882, it was epistemically possible for somebody to square the circle, it was always metaphysically impossible. Since all mathematical proofs are a priori, that it is impossible to square the circle is a priori. Thus in this context it is important to distinguish between "knowable a priori to be false" and "epistemologically impossible". Notice that the distinction between two kinds of epistemic possibility that Gendler and Hawthorne talk about comes into this. If something is provable but has not been proven yet, then it is knowable a

priori. But it's negation is still epistemically possible in the weaker (permissive, to use Gendler and Hawthorne's terminology) sense. In this context it is important to distinguish these two notions of possibility.

The epistemically possible and the metaphysically possible are orthogonal. That is to say, it can be epistemically possible that P without being metaphysically possible that P and vice versa. To illustrate that it can be epistemically possible that P without being metaphysically possible, consider Edgington's own example: Suppose the security men inside the tradesmen's entrance to 10 Downing Street find a large parcel. Upon examining it closely they hear a ticking noise and call the bomb squad because that bomb might explode. But it turns out that the parcel contains an eighteenth century clock just back from repair. Now, there was no metaphysical possibility that the parcel would explode, although there was an epistemic possibility.

The converse, that there are cases where something is not epistemically possible but is metaphysically possible is illustrated with one of Kripke's examples (see Edgington, 2004, p. 7): Leverrier, noticing some irregularities in the orbits of the planets, concludes that they must be caused by another, as yet unseen planet, and decides to call it "Neptune". Now, it is epistemically possible that his hypothesis was wrong—that there is no such planet. But if his hypothesis is right—if Neptune exists—it is the planet causing these irregularities. This conditional is known a priori, at least by Leverrier. It follows from his stipulation about the use of 'Neptune'. There is no epistemic possibility that Neptune exists and has nothing to do with the irregularities. But there is a metaphysical possibility. Thus it was metaphysically possible that Neptune, which caused the irregularities, was knocked off course a million years ago and did no such thing.

Following Edgington, we can further suppose that an unconnected group of astronomers in China develops a new, more powerful telescope, and discovered this same heavenly body and decided

to call it 'Buddha'. It is not knowable a priori that Neptune is Buddha. But if this is a situation where the same object was named twice in two separate incidents, any metaphysically possible thing that could have happened to Neptune could have happened to Buddha. There is no metaphysically possible situation in which they are different objects. But for anyone wondering whether Neptune is Buddha, it is epistemically possible that it is not.

Now, coming back to Gendler and Hawthorne's definition of logical possibility, notice that they do not describe a very sophisticated notion of logical possibility. First of all, in this definition one cannot say that things like square circles would be logically impossible which it should be. Also some second order "contradictions" are not provable. But they are contradictions nevertheless and thus remain logically impossible. However if we follow Gendler and Hawthorne's definition of logical possibility these contradictions would not be logically impossible since they are not provable. To elaborate: In first order logic the following holds:

if $s|-A$, then $s|= A$ (this is soundness)

if $s|= A$, then $s|- A$(this is completeness).

This means that in first order logic A is provable if and only if A is logically true (i.e., valid). However, in second order logic, there is no complete proof procedure. In other words, for any particular system of proofs for second order logic, there are "logical truths" that are not provable. It follows that there are "contradictions" that are not provably inconsistent—that is, there will be A, B which are inconsistent but cannot be proven to be inconsistent. Let T be an unprovable logical truth. Then $\neg T$ is logically impossible. But if from $\neg T$ we could derive $P \wedge \neg P$, we could thereby prove $\neg\neg T$, so we can prove T after all. Hence there are no contradictions provable from $\neg T$. Thus, we can see that Gendler and Hawthorne's definition of logical possibility is inaccurate.

This is an overview of the different kinds of possibilities and some key philosophical issues

relating to these kinds of possibilities. Now let us look at the notions of conceivable and possible.

Section 4. Overview of Conceivability and Possibility

It is important to investigate these two notions and delineate the differences between the two because there is a natural tendency to presuppose that conceivability of an entity entails its possibility. Therefore, before investigating the question whether conceivability is a guide to possibility or not, we should try to get a clear understanding of the notions independently of one another. The point is nicely put by Timothy Williamson:

> Although there are truth and falsehood about conceivability and inconceivability, they concern our mental capacities, whereas metaphysical modalities are supposed to be mind-independent. They are not contingent on mental capacities (Williamson, 2007, p. 4).

So conceivability is a mind-dependent and possibility a mind-independent notion.

Kripke, in *Naming and Necessity*, identified the two way independence between the knowable *a priori* and the metaphysically necessary. *A priori* and *a posteriori* are epistemological and necessity and contingency are metaphysical categories. In other words, it is an epistemological question to ask whether a proposition is a priori or a posteriori. Whether a proposition is necessary or contingent is a metaphysical question.

With Kripke we learnt not only that these concepts differ in their intension, they also differ in their extension. Besides necessary a priori and contingent a posteriori truths there are also necessary a posteriori and contingent a priori truths[11]. Kripke's own example of necessary a posteriori truth is the following: the standard one meter rod in Paris is one meter long. Suppose a person when he sees the standard meter rod does not know that this is the standard meter rod, and measures it. Upon measuring it he finds out that "the standard one meter rod is one meter long". The proposition that "the standard

[11] Notice that Kripke wasn't the first to consider contingent a priori truths. Kant famously defends the synthetic a priori, where synthetic roughly correspond to some sort of contingency.

one meter rod is one meter long" is necessary but in this example, one learns it a posteriori.

Now, the question is why might we have thought otherwise, prior to Kripke? In Kripke's own words: "I guess it's thought that….if something is known *a priori* it must be necessary, because it was known without looking at the world. If it depended on some contingent feature of the actual world, how could you know it without looking?" (Kripke, 1980, p. 80).

What is the connection between a priori and a posteriori and, necessity and contingent, on one hand and conceivability and possibility on the other? This is a more complicated picture post-Kripke. Gendler and Hawthorne explain succinctly:

> On the traditional picture…, there is a straightforward…way to explain the connection between conceivability and possibility….P is possible iff it is not necessary that not-P. Let us introduce… that P is *Conceivable* iff it is not a priori that not-P. If all and only a priori truths are necessary truths, then all and only Conceivable truths are possible truths. For the Conceivable truths are just those whose negations are not a priori, and the possible truths are just those whose negations are not necessary. And since the latter two classes coincide, so do the former two.
>
> ….[s]ome version of this form of reasoning is implicit both in Descartes and Hume…Put crudely, in Descartes the direction of explanation runs from the metaphysical to the epistemic: something is knowable by reflection because it is necessary; in Hume, the direction of explanation runs from the epistemic to the metaphysical; something is necessary because the mind treats it as such. Yet in each case the metaphysical and epistemic categories coincide.
>
> On the post-Kripkean picture, however, no such explanation is available. For if there are a posteriori necessities and a priori contingencies, no such grounds can be appealed to in establishing a conceivability—possibility link. On the post-Kripkean picture, even if it is not necessary that not-P, it may still be a priori that not-P (contingent a priori); and even if it is not a priori that not-P; it may still be necessary that not-P (necessary a posteriori). But then, by substitution, it may be possible that P. Thus the contingent a priori seems to guarantee that there will be cases of possibility without Conceivability; the necessary a posteriori seems to guarantee that there will be cases of Conceivability without possibility (Gendler and Hawthorne, 2002, pp. 32-33).

Defining conceiving broadly as any sort of mental depiction of a scenario, Gendler and

Hawthorne put forward the following list of mental activities (any ones of which or any natural cluster of which) might qualify as candidates for conceiving:

(1) rationally intuiting that it is possible that P

(2) realizing that not-P is not necessary

(3) imagining that P

(4) conjecturing that P

(5) accepting that P for the sake of argument

(6) describing to oneself a scenario where P obtains

(7) telling oneself a coherent story in which P obtains

(8) pretending that P

(9) make-believing that P

(10) supposing (that) P

(11) understanding the proposition that P

(12) entertaining that P

(13) mentally simulating P's obtaining

(14) engaging in off-line processing concerning P

In their words,

> …the wide variability among their features suggests that the notion in question may be highly elusive. Some for example are propositional attitudes; some are attitudes towards scenarios or states of affairs; and still others are activities. Some seem explicitly sensory; still others are neutral on this question. Some are highly conceptual; others are strongly language based; still others are, perhaps, non-conceptual. Some seem to take place primarily spontaneously; others only under our deliberate control; others in both ways. All seem capable of being directed both towards propositions (or states of affairs) involving particular individuals as well as propositions (or states of affairs) that are general. And both within and among them there seem to be variations in degree of

> privileged access associated with the attitude/activity and its content/object. In light of these differences, one might reasonably wonder which, if any of these features alluded to is required by conceivability in the sense we seek (Gendler and Hawthorne, 2002, pp. 7-8).

Notice that Gendler and Hawthorne do not seem to make a clear distinction between conceiving and imagining[12]. What they say about whether or not there is a distinction between the two is the following:

> There is a traditional distinction made between (sensory) imagining on the one hand, and (non-sensory) non-imagistic conceiving on the other. But it is far from settled whether the distinction has a proper role to play in circumscribing the appropriate subject-matter for an investigation of conceivability as a guide to possibility (Gendler and Hawthorne, 2002, p. 9)

Their list above suggests that there is no clear cut distinction between imagining and conceiving. However, Hawthorne and Gendler themselves do not point out whether they are aware themselves that there are some problems with some of the items on the list, as I will point out shortly. Gendler and Hawthorne merely uses this list to toss out some possible candidates for conceiving. I will use this list as a starter for my argument that there is a distinction to be made between conceiving and imagining. And more importantly, as we will see in the next section, I claim that the distinction between the two does have "a proper role to play in circumscribing the appropriate subject-matter for an investigation of conceivability as a guide to possibility".

Another problem with the list is that some items in the list assume possibility in describing conceivability. Although it is all too common to slip from one to the other without even noticing, this is something I want to avoid. I want to examine these notions independently in order to be able to investigate whether conceivability is a guide to possibility or not.

[12] Gendler, and Hawthorne point out (p. 1 footnotes 1 and 2, 2002), that conceiving can be used in the narrow and broad sense of the term. In the broad sense, conceiving refers to the activity of representing scenarios to ourselves using concepts, imagery (actual or non-actual) etc. The word 'conceive' which is traceable to the Latin verb *concipere* shares its root with the word 'concept', which is traceable to the past participle, *conceptus* (see, Gendler, 2002, footnote 1).

To see what I mean, let us look at the list again: (1) is "rationally intuiting that it is possible that P". This clearly presumes a connection between conceivability and possibility. Similarly, (2) (realizing that not-P is not necessary) presumes a connection between conceivability and possibility because in classical Modal Logic $\neg\Box \neg p = \Diamond p$ and thus (2) presumes P's possibility too.

Not all items on the list suffer from the defect of assuming a connection between conceivability and possibility. Some are problematic for other reasons though. (3) (imagining that P) implies conceiving = imagining. And, as I shall argue, there is a crucial distinction to be made between the two.

It is worth pausing to remark on a couple of other items on the list. With regards to (6), notice that there is a difference between "describing to oneself a scenario where P obtains" and "describing that P". This is not a distinction pointed out by Gendler and Hawthorne. As an example consider the following: I can describe to myself a scenario where Dave squares the circle [Dave squaring the circle = P]. I do not need to describe P itself in describing a scenario where P obtains. As I shall suggest later, describing the scenario does not entail the possibility of P itself, only P's conceivability.

In this context also notice that there is a difference between *squaring the circle* and producing a square circle. The first, is conceivable but not possible. The latter is not even conceivable. A square circle is an impossibility much like a hot and cold thing. That is, no single object can have both properties, at once. To belabour the obvious, suppose S is a circle with centre C, and also is the square ABCD. Since S is a square, the line CA is longer than a line from C to a point P half way between A and B. But since S is a circle, CA = CP. This is a contradiction[13].

[13] "Squaring the circle" is the problem of constructing a square with the same area as a given circle using a finite number of steps with only a compass and straightedge. This entails constructing pi. The impossibility of squaring the circle follows from the fact that pi is a transcendental number—that is, it is non-algebraic and therefore a non-constructible number. If one solves the problem of the quadrature of the circle using only compass and straightedge, then one has also found an algebraic value of pi, which is impossible. Note that the transcendence of pi implies the impossibility of exactly "circling" the square,

Going back to "describing to oneself a scenario where P obtains", one can conceive of a situation where an impossibility holds i.e., one can describe to oneself a scenario where P obtains (i.e., a scenario where Dave squares the circle). Notice that this does not involve knowing what exactly is involved in squaring the circle. However, one cannot even describe a scenario in which there is a square circle. Exactly why there is this difference remains to be seen, but that "a square circle" is an obvious analytical impossibility while "squaring the circle" is a subtle substantive impossibility presumably is part of the answer. So not only is there a difference between squaring a circle and describing a square circle, there is also a difference between describing a square circle and describing to oneself where somebody squares the circle. Describing a situation where somebody has squared the circle amounts to describing what would count as someone having squared the circle, but not what it would be for someone to have produced a single shape that is both a square and a circle at once. In light of our discussion, does "describing to oneself a scenario where P obtains" entail possibility of P? The answer is "No". It entails only conceivability of P.

I will discuss the concept of conceivability in much more detail in the next section where we look at how the conceivable and the imaginable are related.

Section 5. Conceivability and Imaginability Revisited

Here, we need to ask the question, is there a distinction to be made between "imagining" and "conceiving"? As I have said before, I argue that there is a distinction between the two. There are various aspects to these concepts—psychological, epistemological, metaphysical, and semantic. As a starting point, Imagining and conceiving can be characterized in the following way:

• **Imaginable**: This has to do with visual imagery. Imagining x involves both "seeing" x and "seeing as" x. One would know how something would look in one's mind's eye, if you will. One can imagine,

as well as of squaring the circle.

for example what a red couch would look like i.e., one can imagine a red couch. Similarly one can imagine a golden mountain. Notice though, one can of course be mistaken about whether something has been successfully imagined or not.

• **Conceivable**: This has to do with forming a general conception of the entity or entities in question, not necessarily involving any particular figurative detail of the entity in question. One could compare it to Locke's general ideas. As Locke said, when we think about a triangle, we do not think about a particular kind of triangle such as a right angle triangle or an isosceles triangle. We think about a triangle in general. Similarly, consider the example of motherhood. One does not conceive a particular mother, but invokes the general notion associated with mothers. Some may object to this description of "conceivable" and might say that when they think about a triangle or motherhood they think about a particular triangle, a right-angle triangle, for example or a particular mother. If they are right, conceiving for them does involve figurative detail of the entity in question, but what is important is that in conceiving a particular entity, it need not necessarily involve any particular detail. In the case of a triangle, it does not need to be a right-angle triangle, it can very well be another kind of triangle.

Conceiving that P or conceiving of P seems to involve knowing what would count as a case of P. For example, if one can form a concept of a triangle, one must know what would count as a triangle, if one were to come across one. Which is *not* to say that one *could* come across something, merely because it is conceivable. Consider conceiving that someone solves the problem of squaring the circle. One cannot *imagine* this, but one can *conceive* it—one knows what would *count* as a solution to the problem, even though one knows that no such solution is possible. As we will see later, conceivability is only a sufficient but not a necessary condition for something to be a case. Conceiving could involve a broad range of mental activities, for example where conceiving involves being able to tell oneself a coherent story in which P obtains, where P could be something one pretends, where one supposes (that

P), or where it otherwise involves understanding P.

Notice that these categories are not exclusive. Conceiving might involve mental imagery in certain cases, for example. Or imagining might involve forming a concept. Here we can introduce a further distinction, which at first may appear psychological, between sorts of conceiving. This is what I call "fully and coherently describable". It can be defined in the following way:

• **Fully and coherently describable**: This has to do with having a sufficient grasp of a concept or concepts in order to be able to describe it fully and coherently. For something to be fully and coherently describable one needs to be sure that the description is complete and fits together cohesively. For example, consider a chiliagon. Although one may have never come across one, one knows that it is a hundred-sided figure. From the concepts of other many sided figures one can have a sufficient grasp of the concept and one is able to not only extrapolate the notion of a hundred sided figure but also know that the description is complete and cohesive.

At first this appears to be a psychological distinction. But while we sometimes can conceive of impossible things, we can only fully and coherently describe possible things. Indeed, that is part of the *definition* of the notion. Thus fully and coherently grasping for that reason cannot be a merely psychological matter. We have a psychological(ish) distinction between the imaginable and the conceivable; then, a different distinction between the fully and coherently describable and the not-fully-and-coherently describable. Consider a particular case that is conceivable but not possible, e.g., someone squaring the circle, a bit more in detail. We can conceive it because we know what would count as a solution. But when we try to give a complete description of the situation we cannot, for a complete description would involve describing the construction of pi, which cannot be done.

Next let us return to the question of how imaginability and conceivability are related. Is the imaginable a subset of the conceivable? Is everything that is imaginable conceivable too? The answer

to this question is yes. If one can imagine it, i.e., see it in one's minds' eye, one must have a concept of it available to him/her. Of course, what I say here is plausible only if one distinguishes *imagining* from merely *picturing in one's mind*. A distinction parallel to that between *seeing* and *seeing as* is needed here. It is a commonplace that someone unacquainted with tennis presented with a tennis raquet, in some sense, does not *see* a tennis raquet. But nobody supposes that the person's visual field is blank in the area in question—rather we note that the person sees the raquet, but does not *see it as* a raquet, and for many purposes that is what is crucial. Imaginability, properly so called, requires the same sort of conceptual resources as seeing as. One might be able to produce a mental picture of something one saw yesterday but had no idea what it was, but one could not *imagine* it. For example, consider someone presented with a cricket bat who had no idea what it actually was, but perhaps thought that it was used for putting pizzas into ovens at restaurants. If the person in question were to think about the object at a later time, this would not qualify as imagining since it does not involve seeing the cricket bat as a cricket bat. This would count as a case of conceiving since one had the concept of something being used for putting pizza into ovens. If one knows what x would look like i.e., one can imagine x, then if one were to come across a case one would know that it counts as a case. This is because one has a concept of x, i.e., one can conceive x. Everything that is imaginable is conceivable but not vice versa, as have already seen in case of the example of squaring the circle. Imaginability requires more accuracy than conceivability. Imagining requires both seeing something and seeing it as.

To recap, let us consider various combination of categories and see whether those are possible combinations or not and if not, why not?

- imaginable but not conceivable: this is not a possible combination since everything that is imaginable is also conceivable.
- imaginable (and a fortiori conceivable) and fully and coherently describable: For example, take the

concept of a golden mountain. Can one imagine it? Yes, I can see what a golden mountain would look like. Can one conceive it? The answer is "yes"—if one were to come across a case one would know that it counts as a case. Is it a fully and coherently describable case. "yes".

● imaginable (and a fortiori conceivable) and not fully and coherently describable: For example, Poonam being a member of the royal family. I can conceive and even imagine what it would be like to be a daughter of the Queen. However, I cannot fully and coherently describe it because as I fill in the details, the story eventually becomes incoherent. So, in this particular example, what makes it metaphysically impossible for me to be a member of the English Royal Family is that identity conditions are determined by biological origins (following Kripke). Thus as we fill in the details that make me actually a member of the Royal Family, we are eventually going to have to change my ancestry (or theirs), so eventually it is not me who is a member of the Royal Family, but someone a lot like me in some ways but with different origins.

● conceivable and fully and coherently describable and not imaginable: Prima facie, it seems that this is not a possible combination. For it seems, since one can conceive it and fully and coherently describe it then one can very well imagine it. However, consider for example, a chiliagon. Can one imagine it? The answer is not so straightforward for this case is a little bit more complicated than others. It is tempting to think that since one can conceive and fully describe x, one can imagine x. However, as Descartes famously argued (in the Sixth Meditation), it seems that a chiliagon, while conceivable, is not imaginable. Although one might think that it is imaginable, how could one tell that one in fact had imagined a hundred sided figure and not a ninety eight sided figure? One can conceive it though. Is this a fully and coherently describable case? Given what we know about many sided figures, one would be able to have a sufficient grasp of the concept so as to tell whether a figure is a chiliagon or not, if one were to come across one. So it is an example of a conceivable and fully

and coherently describable but not imaginable case.

As we have seen, for some cases it is fairly easy to discern which category or categories they belong to. However, there are some not-so-straightforward cases for which the answer may not be so clear cut. We will see in section 7, that the list above comes in handy in devising what count as *cases* and what does not.

As far as the distinction between conceiving and imagining goes, we have established that there are things that we can conceive but cannot imagine. Strawson's auditory world is another case in point. Why is this not imaginable? Many people think imagining should allow many sensory modalities. I can imagine now what apples taste like. Apparently, smell is a crucial accompaniment for taste. It is believed by some scientists that if we did not have a sense of smell apples and potatoes would taste the same. Thus, if this theory is true, in imagining the taste of an apple, I am (consciously or unconsciously) imagining the associated smells as well. So, I cannot imagine the taste of an apple without having the accompanying smell of an apple. However, in the auditory world case, we have no way to tell what a completely auditory world would be like. For, it would not or should not look like anything but only sound like this or that. The auditory world case is similar to the chiliagon. Namely, it may be that one accidentally "images" a chiliagon when trying to imagine one—but that is insufficient. Similarly for the auditory world—we don't know enough to know the difference between accidental accuracy and inaccuracy. It might be useful to remind the readers about the cricket bat example, where we noted that imagining involves both seeing and seeing as. In the auditory world example, there is no way of making sure that the world imagined is *completely* auditory (as Strawson requires it to be, in his project) without any other accompanying information from other sensory modalities. Thus we are not in a position to see whether the world we are considering which we may think is auditory is in fact so. So we may not be seeing it as the auditory world. So, auditory world is

conceivable but not imaginable. We see once again that there are cases of impossibilities that are conceivable but not imaginable. The above examples further establish my argument that there is a distinction between the notions of conceiving and imagining.

One important point that emerges from the discussion of various categories above is that *impossible* things are not fully and coherently describable. If a description is really both complete and coherent, that would imply possibility. So, we can be *mistaken* about whether we have fully and coherently described (or imagined) something, as would happen when I think I have figured out how to build a perpetual motion machine or some such thing. It seems right to say that it could happen that I have conceived of such a machine, and that I *thought* I had fully and coherently described it, but was mistaken. It is not possible to build such a machine, so my description must have been either incompletely specified or incoherent in some way.

Section 6. Is Conceivability a Guide to Possibility?

There is a natural tendency to think that conceivability implies possibility. Conceivability in this sense is construed in a broad way that includes mental imagery of actual or non-actual things, representation of a concept, or grasping of a concept that implies linguistic expressibility. Often we come across situations where if one is asked whether X is possible or not, one tries earnestly to conceive of X. Under such a broad construal of 'conceive', it seems to imply that what we conceive is *presented as possible*. However, sometimes we make the mistake of jumping to the conclusion that what is *presented* as possible is *in fact* possible.

Upon carefully examining the notions of conceivability and possibility we find that they are distinct notions. As Williamson points out, conceiving is mind-dependent and possibility mind-independent. But even if conceivability does not *imply* possibility, we might still ask whether

conceivability is a guide to possibility. If it is, what kind of possibility is it a guide to?

Recall that Gendler and Hawthorne distinguish between epistemic and non-epistemic possibility the latter of which is, in turn, divided into three kinds; logical, nomological and metaphysical. They argue that conceivability definitely cannot be a guide to epistemic possibility. This can be easily demonstrated taking their own example: if I know that the cat is on the mat, then it is not epistemically possible for me that the cat is not on the mat. Yet, I can conceive a situation in which the cat is not on the mat. So, I can conceive something epistemically impossible (Gendler and Hawthorne, 2002, p. 4). So, in talking about conceivability as a guide to possibility we must mean some other kind of possibility (other than epistemic possibility). According to Gendler and Hawthorne, the question whether conceivability acts as a guide to possibility arises only in the case of metaphysical possibility.

As we have seen, they define logical possibility as "P is logically possible just in case no contradiction can be proven from P using the standard rules of deductive inference". Based on this definition of logical possibility they suggest that conceivability is "superfluous" as a guide to logical possibility. It is better determined by logical proofs and methods than by scenario depiction (Gendler and Hawthorne, 2002, p. 5). In their words, "whether or not a contradiction can be derived from P seems better determined by proof procedures than by scenario depiction" (Gendler and Hawthorne, 2002, p. 5). I have to disagree. First with regard to their definition of logical possibility; as already noted, in some cases, proof procedures do not exist. Thus, even logical impossibility may not be better determined by logical proofs and methods than by conceivable scenario depiction. But notice that *formal* "proofs" of possibility are *not* formal proofs at all, at least typically. Rather, a proof of logical possibility usually involves presentation of a *model* in which a sentence can be seen to be true. This obviously has a lot in common with "scenario depiction" even if often the scenarios involve assigning mathematical properties to predicates and working with domains of numbers rather than

other objects. Surely Gödel's Second Incompleteness Theorem should warn any philosopher not to expect that formal proofs of *consistency* are going to tell the whole story about logical consistency.

Gendler and Hawthorne think that conceivability is not a guide to nomological possibility for the same reason as it is not a guide to epistemic possibility— we can conceive of many things that are nomologically impossible. It seems all too easy to conceive of things that are not possible in the relevant sense. They seem to think that if conceivability is a useful guide for anything, it is a guide to metaphysical possibility. Gendler and Hawthorne recognize that while the above clarifications dispel some of the confusion surrounding these issues, they do little to solve philosophical puzzles regarding conceivability as a guide to possibility. The very notion of metaphysical possibility is highly loaded and elusive.

> While these clarifications dispel a certain amount of confusion, they do little to resolve an obvious puzzle: on the face of it, the idea that conceivability is a guide to metaphysical possibility is extremely problematic. According to current orthodoxy, metaphysical possibility can neither be reduced to, nor eliminated in favour of, linguistic rules and conventions; it constitutes a fundamental, mind-independent subject-matter for thought and talk. Given that picture, it is rather baffling what sort of explanation there could be for conceiving's ability to reveal its character. It seems clear that the causal explanation for the reliability of perception is unsuitable here — and it is profoundly difficult to see what to put in its place (Gendler and Hawthorne, 2002, p. viii).

I agree with Gendler and Hawthorne, that in most cases when we talk about conceivability as a guide to possibility we talk about metaphysical possibility. As Chalmers points out:

> [t]here is at least some plausibility in the idea that conceivability can act as a guide to metaphysical possibility. By contrast, it is very implausible that conceivability entails physical or natural possibility. For example, it seems conceivable that an object could travel faster than a billion meters per second. This hypothesis is physically and naturally impossible, because it contradicts the laws of physics and the laws of nature. This case may be metaphysically possible, however, since there might well be metaphysically possible worlds with different laws. If we invoke an intuitive conception of a metaphysically possible world as a world that God might have created: it seems that God could have created a world in which an object traveled faster than a billion meters per

> second. So in this case, although conceivability does not mirror natural possibility, it may well mirror metaphysical possibility (Chalmers, 2002, p. 146).

But as I will argue, conceivability is *not* a guide to possibility—not even metaphysical possibility. The question whether conceivability is a guide to possibility or not arises because, when we engage in conceiving that involve depiction of scenarios to ourselves "the things we depict to ourselves frequently present themselves *as possible*, and we have an associated tendency to judge that they *are possible*. Indeed, when invited to consider whether something is possible, we often engage in deliberate effort to conceive of it; upon finding ourselves able to do so, we conclude that it is. We may even decide that something is impossible on the basis of our apparent inability to conceive of it" (Gendler and Hawthorne, 2002, p. 2).

Williamson's words illustrate the case why conceivability is not a guide to possibility, both when we talk about conceivability in non-philosophical terms as well as when we talk about conceivability in philosophical terms:

> The impression is that, outside philosophy, the primary cognitive role of conceivability is propaedeutic. Conceiving a hypothesis is getting it onto the table, putting it up for serious consideration as a candidate for truth. The inconceivable never even gets that far. Conceivability is certainly no good evidence for restricted kinds of possibility that we care about in natural science or ordinary life. We easily conceive particles violating what are in fact physical laws…On this view, conceiving, outside philosophy, is not a faculty for distinguishing truth and falsity in some domain, but rather a preliminary to any such faculty (Williamson, 2007, p. 4).

Although sometimes conceiving, at least prima facie, may seem to be a guide to metaphysical possibility, there are well known misleading cases. That is to say, there are conceivable impossibilities. For example, an Ancient Greek might have found it conceivable that Hesperus is not Phosphorous although this is not actually possible. There are examples like the auditory world where Hero resides in a completely auditory world, which is metaphysically impossible but conceivable.

Examples like trisecting an angle with only a ruler and compass demonstrate the same point. That is why conceivability cannot be a guide to possibility. Although conceivability does not imply possibility, it divides impossibilities into two classes—the conceivable ones and the inconceivable ones. That is to say that not all impossibilities are conceivable. A hot cold thing, for example, is not even conceivable.

Now conversely, does inconceivability imply impossibility? Prima facie it might look like inconceivability implies impossibility. For example, when we are asked to think about things like a hot cold thing, or a tall short person, one might jump to the conclusion that these things are impossible because they are inconceivable. As we noted in the beginning in examples like the auditory world, it may seem like impossibility is tied to inconceivability. But surely there are things that are inconceivable because we do not even have the concepts available to us, but that are possible. Advancement in the Sciences is a good proof for things that were inconceivable once but are certainly possible. The first test tube baby was born in 1978, and presumably the idea of a test tube baby was conceivable in 1978. But it is plausible that such a thing was *inconceivable* one or two thousand years earlier. However, it was not impossible. This is an example of something that was inconceivable once but its inconceivability did not imply its impossibility. Thus we find that the relationship between conceivability/inconceivability and possibility/impossibility is orthogonal—neither implies the other.

Now coming back to our three layers of the imaginable, conceivable, and the fully and coherently describable, we will see that these three layers are closely related to the question whether conceivability is a guide to possibility or not. It may look like that what I call “fully describable cases” and conceivable cases are one and the same. Let’s see how the imaginable, the "fully and coherently describable" and the merely conceivable are related, especially the last two. A person would typically be able to tell whether something was conceivable or not, but *not* whether it was fully and coherently

describable. So, someone might *think* they had conceived of something in complete detail, but be mistaken about that. This is important because something that is *fully described* in this sense is *possible*. So "fully described conceivability" *is a guarantee of possibility* because—to put the matter bluntly—it guarantees being a fragment of a possible world. Unfortunately, as a matter of epistemological fact, human beings often mistakenly think they have fully described something, e.g., a being with all perfections, when in fact they have not. Similarly, all fully and coherently describable cases may not in fact be imaginable. Again as a matter of epistemological fact one might mistakenly think what one has imagined, i.e., formed a mental picture of, is a chiliagon. But it may very well be a figure with 97 sides. In this context I want to point out that number 6 (describing to oneself a scenario where P obtains) on Gendler and Hawthorne's list of candidates for conceivability, does not entail possibility. Describing to oneself a scenario in which P obtains, does not involve fully and coherently describing P.

In this context, it is worth pointing out that making a distinction between conceiving and imagining and spelling out the above three layers, help us show that this distinction has a "proper role to play in circumscribing the appropriate subject-matter for an investigation of conceivability as a guide to possibility". As we saw in section 4, Gendler and Hawthorne express doubt about this. Interestingly, we have arrived at a negative result for our first question—conceivability *does not* imply possibility—but conceivability turn out to be suitable to play an important role in evaluation of counterfactuals.

Section 7. States and Cases

The reason conceivability is such an important notion is because for something to be a case it has to be at least potentially conceivable. The point of all this discussion has been to make it possible

to isolate a notion that will play a role in our account of counterfactuals similar to that played by *possible worlds* in Lewis-Stalnaker accounts. That is the notion of a *case*. As we will also see in Chapter IV, we will use *cases* in our semantics for a new theory of counterfactuals with impossible antecedents. Here we will use the notion of conceivability to distinguish between those inconsistent classes of statements that count as cases and those that do not.

We borrow from the notion of a *state* used in Relevance Logic, to define what a *case* is. States can be both inconsistent or incomplete. In Relevance Logic *states* represent ways that parts of the world could be, ways that parts of the world could not be, ways the whole world could be, and ways the whole world could not be. In our semantics for a new theory of counterfactuals, *states* can be extended to *cases*. Any random set of sentences that can be *extended* to something conceivable is a case. By adding enough details to make something conceivable out of a set of sentences that contains some inconceivability or incompleteness, we can turn it into a case.

The notion of a case is important because, as we will see in Chapter IV, we use cases as a (modal) tool to analyze counterfactuals with impossible antecedents. The reason all this is important is because in Chapter IV, I need to distinguish those inconsistent classes of statements that count as *cases* in my semantics for counterfactuals with impossible antecedents, from those that do not. Let us remind ourselves what counts as cases: Squaring the circle and flying dogs are some examples of cases. Examples of non-cases are a flat-spherical object, a tall short person. The former ones count as cases because, one would know what would count as squaring the circle by using only a finite number of steps with a compass and a straightedge, although one cannot actually do it. These are conceivable impossibilities. Whereas for some non-cases, one would not know what such an object be like. When it comes to the above sort of non-cases not only can one not imagine it but one cannot make sense of such things conceptually. As we will see in Chapter IV, there are some non-cases that may be

potentially conceivable, thus may turn out to be cases. These are inconceivable impossibilities. They are inconceivable due to human conceptual or cognitive limitations. Some of these entities may in time (in the future due to changes/advancement in scientific theories) come to be regarded as conceivable. It turns out that any random group of sentences that can be *extended* into something conceivable is a case.

In short, the ones that are cases, even though they are impossible, are the ones that can be extended to something *conceivable*. If something is not even potentially conceivable, then it does not count as a case. So cases are of two kinds: possible and impossible.

So where are we at after all these? Let me clarify by using a now familiar example: The auditory world is metaphysically impossible. Is it imaginable? No, at least not completely, which is why in the end Strawson gives up that project. But a more important question is this: is the auditory world conceivable? The answer is yes. This is the reason why it is not completely useless. This is where we are at now. We are going to take advantage of the conceptual spadework we have just completed to further develop a theory counterfactuals. This will allow us to explain the amount of mileage Strawson is able to get out of the sound world, in spite of its impossibility.

Chapter IV: Counterfactuals with Impossible Antecedents

In this chapter I will sketch a theory about how to make sense of counterfactuals with impossible antecedents. The aim is not just to address the key problem of counterfactuals with impossible antecedents, but to then shed some light on a contested class of thought experiments, namely incomplete/impossible yet useful ones. By no means do I dare pretend that my account is complete. However, I contend that it does make progress towards making sense of these notoriously ill-behaved conditionals.

This chapter is divided into five sections. In section 1, I explain why we need a better story about these conditionals. Section 2 includes a preliminary discussion of Relevance Logics and the rudiments of the semantics of Relevance Logic that is used in analyzing counterfactuals with impossible antecedents. In section 3, I discuss the idea of 'closeness' with regards to possible worlds. In section 4, I consider some details that need to be ironed out regarding this new theory of counterfactuals. In this section I discuss the notion of 'closeness' with regards to impossible worlds. In Section 5, I present a new theory of counterfactuals that can handle counterfactuals with impossible antecedents.

Section 1. The Need for a Better Story about Counterfactuals

We need a better story about counterfactuals with impossible antecedents—a better story than classical two valued logic (classical logic for short, henceforth) has given us. There are two reasons for this: First, the more direct and general reason, as we noted in Chapter I, is that classical accounts of counterfactuals cannot handle counterfactuals with impossible antecedents—not very satisfactorily, anyway. I will assume the deficiencies of classical logic and focus on how to tell a better story about these conditionals.

Secondly, it is an important contention of this thesis that the payoffs of an improved account of counterfactuals are to be found in various areas of philosophy. To make the case, I have chosen to focus on one particular philosophical payoff: if we have a better explanation of these counterfactuals we are going to able to better analyze thought experiments in terms of counterfactuals. In earlier chapters, I have argued both that there are important advantages to analyzing thought experiments in terms of counterfactuals and that many thought experiments that involve impossible scenarios are both informative and useful. Since, in the counterfactual analysis of thought experiments, the antecedent amounts to a description of the scenario, this account does not go very far if all counterfactuals with impossible antecedents come out true. In order to defend these claims we need an account of counterfactuals that is more comprehensive. In particular, we need one that can give us a better story about counterfactuals with impossible antecedents.

As we will see, to construct a more comprehensive account of counterfactuals we will make use of a logical concept of "impossible worlds". At this point, it is natural to wonder what these impossible worlds are or why we need them at all. As Daniel Nolan puts it:

> [T]here are a variety of areas in which it is useful to be able to reason about impossible situations and to do so in a nontrivial way (so that it is not good enough to just throw up one's hands and say that everything follows). The mere fact that we can think about what is impossible does not commit us to impossible worlds, any more than the mere fact that we claim that some claims are necessary or possible commit us to possible worlds. But just as it is a natural way to cash out our talk of necessity and possibility in terms of possible worlds, it is tempting to talk about impossible worlds, or situations, or ways things couldn't be (Nolan, 1997, p. 536).

In order to avail ourselves of concepts such as impossible worlds we turn to recent works in Relevance Logic.

Section 2. Semantics of Relevance Logic

Relevance Logics are non-classical logics that developed out of the main idea that a necessary condition for the conditional A → B to be true is that there must be some kind of a "connection between A and B". In other words, for A → B to be true, A must be relevant to B in some sort of way. In turn Relevance Logics also attempt to avoid paradoxes of material and strict implication.

Before going into the semantics for Relevance Logic I want to explain these paradoxes. This is important because Relevance Logics are an attempt to preserve some sort of relevancy between the antecedent and the consequent when it comes to conditionals. As such, the philosophers who have developed Relevance Logics have needed to develop tools to avoid those features of classical logic that result in classically valid reasoning where no such connection exists—for instance, everything following from a contradiction—that are very similar to features of standard accounts of counterfactuals that we shall also try to avoid, as we shall be able to see by considering the paradoxes. In these paradoxes there is evident loss of relevancy between the antecedent and consequent, which is what Relevance Logics strives to avoid. So naturally these paradoxes are distasteful for advocates of these logics. These are not paradoxes in the strict sense of the term, i.e., the ones that entail a contradiction, but they are paradoxes in the sense that they seem to go against ordinary intuition.

First let's see what the paradoxes of material implications are.

The truth table for material implication is familiar:

p	q	p ⊃ q
T	T	T
T	F	F
F	T	T
F	F	T

This says:

(1) whenever the antecedent is false, the whole conditional is true and

(2) whenever the consequent is true, the whole conditional is true, otherwise

(3) the conditional is false.

If we intend the 'p → q' operator to represent some notion of "q follows from p" or "p implies q" this truth table is counterintuitive to say the least. In particular, it gives us the following three paradoxes of material implication:

(a) p → (q → p)

(b) ¬p → (p → q)

(c) (p → q) v (q → r)

From (2), we get the first paradox i.e., (a) a true proposition is implied by anything whatsoever! Thus, " the moon is made of green cheese" implies "2 + 2 = 4". Let p = "2 + 2 = 4" and q be "the moon is made of green cheese". Since, "2 + 2 = 4" is true, the consequence above is true. From (1) we get the second paradox, i.e., (b) if p is false it implies anything whatsoever! According to (b) I am a monkey's uncle implies that the earth is round. Let p be "I am a monkey's uncle". Since it is false, we can demonstrate the consequent of the above i.e. "if I am a monkey's uncle then the earth is round". Consider, now, (c): it has this as a special case: (p → q) v (q → p). This is known as Dummett's scheme, and it amounts to saying "for any two propositions, one implies the other or vice versa" ... which seems strange when you can have entirely unconnected p and q (either (Dave is a philosopher implies cats have tails) or (cats have tails implies Dave is a philosopher)). But (c): (p → q) v (q → r) is one step worse, because it means that either p implies q or else q implies anything whatsoever: Either (Dave is a philosopher implies cats have tails) or (cats have tails implies the sun is about to explode). Well, cats do have tails, so since the sun is not about to explode it must be that Dave

is a philosopher implies that cats have tails.

If we take "implication", as in the case of material implication, as a truth functional connective then we have to pay a big price. If the truth-value of a whole conditional depends on the truth-values of its antecedent and consequent alone, then what matters is the *truth-value,* and not the *content* of the antecedent and consequent. But if the content of the antecedent and consequent is irrelevant, then they may be utterly unrelated to one another. We have abandoned the requirement of ordinary implication that antecedent and consequent be mutually relevant or somehow connected. So, if we adopt "⊃", for implication we embrace truth-functionality in exchange for relevancy. The paradoxes are counterintuitive because the truth-value of one component can determine the truth-value of the whole compound, regardless of the truth-value or content of the other. That is, they disturb us precisely because of this loss of mutual relevancy.

The paradoxes of material implication are not the only paradoxes that relevance logicians try to avoid. There are the so-called paradoxes of strict implication, in which we again see a loss of relevancy between the antecedent and the consequent. Long before Relevance Logics came along, C. I. Lewis invented the strict conditional to avoid the paradoxes of material implication. He introduced a new symbol called a "fishhook".

p ⥽ q (this is called a strict conditional)

is interpreted as "it is not possible that p be true and q be false". So p ⥽ q is equivalent to ¬◊ (p & ¬ q).

In this system, □ p is equivalent to ¬ ◊¬ p. With this equivalence, substitution, and double negation elimination, we get:

□ (p →q) is equivalent to ¬◊ (p &¬ q).

Therefore, $\Box (p \rightarrow q)$ is equivalent to $p \dashrightarrow q$.

There are remnants of the "paradoxes of material implications" in this. For example we get what is called the "paradox of strict implication" which is a modal version of (a) in the list of paradoxes of material implication, which looks like the following:

$\Box p \dashrightarrow (q \dashrightarrow p)$

This says that a necessary proposition is strictly implied by any proposition. Suppose p is necessarily true. Then it cannot be false, and thus it cannot be the case that q is true and p is false. Thus a necessary proposition is implied by any other, however irrelevant that other may be.

Since strict implication is defined so that p strictly implies q if and only if it is logically impossible for p to be true and q false, it follows that a contradiction strictly implies any proposition, and any proposition strictly implies a logical truth. Thus we get the following paradoxes:

$(p \ \& \neg p) \dashrightarrow q$.

$p \dashrightarrow (q \dashrightarrow q)$.

$p \dashrightarrow (q \ \mathrm{v} \neg q)$.

These so-called paradoxes of strict implication seem counterintuitive because again, just as in the paradoxes of material implication, the antecedent seems to be irrelevant to the consequent.

This is enough about the paradoxes of implication. Suffice it to say that relevant logics are created in an attempt to construct logics that avoid these paradoxes and require "$\rightarrow$" to mean something that makes the antecedent relevant to the consequent[14] when conditionals are true.

Before going into showing how, with the help of Relevance Logic and Lewis-Stalnaker style

[14] Re: paradoxes of material and strict implication, see: Mares, E 2006, and Mattey, G.J, 1998

account of counterfactuals, we can account for counterfactuals with impossible antecedents, I will introduce a few semantic primitives taken from Restall's recent work, that this chapter alludes to.

Ways / States

One of the important semantic notions used in Relevance Logic is 'ways' (=states). In possible worlds semantics for classical modal logic, we can only talk about the ways the (actual) world could be. The different ways the actual world could be are in turn different possible worlds. So, in classical possible worlds semantics, 'ways' is just a synonym for 'worlds'. We will see shortly why this is not the case in Relevance Logic. In Relevance Logic, we can talk about the ways the world could be as well as the ways that this world *could not* be. Consider some examples of the ways the world could not be: the world could not be such that there are hot cold things. It also could not be such that there are square circles…. These are some examples of the ways the world could not be since these are examples of inconsistencies.

Classical accounts of possible worlds do not allow us to talk about inconsistencies since these accounts are only and purely a systematization of the possible. As Restall points out, it is obvious that allowance for inconsistencies (a la Relevance Logic) is incompatible with Lewis-style extreme realism. As we know, Lewis himself rejects impossible worlds (Lewis, 1986). This incompatibility between allowance for inconsistencies and Lewis-style extreme realism is a direct consequence of Lewis's construal of the notion of possibility. Nolan captures this point rather nicely in the following:

> On his [Lewis's] conception, possiblia do in fact have the features we associate with them: the merely possible blue swans are literally blue and literally swans, for example. Possible worlds for Lewis, notoriously, are just large objects much like our own cosmos—so the worlds where there are blue swans are just cosmoi with blue swans (among other things) in them. Extending this approach to impossible objects produces literal impossibilities…: if the impossibilium corresponding to the blue swan-and-not-a-swan is literally a swan and is literally not a swan, then a contradiction is literally true… (Nolan, 1997, p. 541).

The problem with literal impossibilities is specific to Lewis' extreme realism only. However, the problem of theorizing about inconsistencies is not specific to Lewis' account alone. It is a problem with all classical accounts of possible worlds.

Also, it is worth noting that classical modal logic deals only with the *total* or *complete* ways the world could be, and it does not accommodate merely partial ways. This is a handicap of classical modal logic. This idea of completeness being essential (for classical logic), clashes with what is often useful in thought experiments, where incompleteness is in some cases harmless, and may even be beneficial. Intuitively, my working on my thesis at this moment, for example, has implications for some things like the department of philosophy (hopefully it will have one more Ph.D. student graduating soon) or my family (viz., that I could not go out skating with them) but what implication does it have on things like the colour of the shirt President Bush is going to wear or whether the U.S. is going to wage a war against Iran next? So, there is at least a prima facie advantage in being able to consider the ways parts of our world could be, without regard to the rest. Putting these ideas together, it seems that prima facie there is an advantage in being able to consider ways parts of world could not be. For example, Dave's squaring the circle has implication for whether or not he would be famous, but it does not seem to have any implication for whether or not the sun would explode. So we will borrow this idea of partial ways and incorporate it in our new theory for counterfactuals. Note that when we talk about 'partial states/ways' we are talking about it in the sense of states which are incomplete—not in the sense of parts of a state.

The entities that represent ways that parts of the world could be, ways that parts of the world could not be, ways our whole world could be, and ways whole world could not be, are called "states". States can be *inconsistent* (some states might answer both 'yes' and 'no' to some issues, so, P & ¬P can come out true in some states, unlike in traditional possible worlds semantics), and *incomplete* (states

need not answer every issue with a 'yes' or a 'no'). In other words, in some states there is just no fact of the matter about whether P or ¬P is true. Or in other words, the law of excluded middle is not valid.

Now we can see why "states" come in handy in analyzing the kind of counterfactuals (and in turn the kinds of thought experiments) we want to: they admit of both "incompleteness" and "inconsistency". If we can use states to analyze counterfactuals and thought experiments, then we no longer have to discount counterfactuals and thought experiments either on account of inconsistency or incompleteness.

Cases

Instead of "states" in our new account of counterfactuals, we will be using "cases". Our discussion of "states" sets the stage rather nicely for "cases". Cases are much like states in that they can be incomplete and inconsistent. In Chapter IV we saw that cases include both possible (including the actual case) and impossible ones. We will be focusing our discussion on impossible cases—naturally, because we are interested in analyzing counterfactuals with impossible antecedents.

Recall the following important details about cases:

- there are possible ones and impossible ones
- impossible cases must be conceivable or at least potentially conceivable.
- not just any group sentence of a world *w* forms a meaningful 'case' which we would need if it were going to be used for evaluation of a thought experiment or a counterfactual.

Let us consider some details about this last point. Consider the following two sets of sentences:

(1) {Poonam is related to Mary, Mary is imprisoned in a tower in Scotland, Mary is the sister of Elizabeth II}

(2) {it is snowing outside, this cup of tea is both hot and cold, Gandhi was killed by a gunman named

Nathuram Godse}

Although (1) is impossible because it implies genetic links that do not exist, nevertheless, it is conceivable. Let me explain why: As I mentioned in Chapter IV, all sets of sentences that can be *extended* to something conceivable are cases. If we can add details to make something conceivable out of a set, that shows that the set is not so fundamentally incompatible that it can be disregarded. (1) is conceivable because we can reckon that I (Poonam) was the genetic progeny of one of the lesser royals, who tossed me over the side of the Queen Elizabeth II yacht when I was born, where I was discovered by a passing cruise ship, then placed in an orphanage… Thus we can add details to (1) and tell a story about the link between me and Mary. So we can add enough details to (1) and thus extend it to something conceivable as a case. (2) is not a case. We cannot do the same to (2). In other words, we have no idea what would be involved in making all of these claims true, because no single thing can be both hot and cold at the same time, thus making (2) a non-case.

Similarly, there are cases where Dave squares the circle and the sun explodes, because supposedly Dave is born on a distant galaxy far in the future, just before the sun goes supernova. You just have to add those details to {Dave squares the circle; the sun explodes} to "link them up". But {This is hot at time t, this is cold at time t} cannot be extended in that way.

Thus we can see conceivability is at least a *sufficient condition* for being a case. Hence impossible but conceivable situations count as cases. However, as I argued in Chapter IV, since there can be inconceivable possibilities (e.g., due to human conceptual or cognitive limitations), conceivability cannot be a necessary condition for being a case, because all possible situations must count as cases (because our new semantics is an extension of Lewis-Stalnaker approaches).

Now the question still remains: What about inconceivable impossibilities? As we just saw, some inconceivable impossibilities, like a hot and cold thing or a tall short person, cannot be cases.

However, could there be some inconceivable impossibilities that are cases, but we can not conceive of them because of human conceptual or cognitive limitations? It seems that there will be sets of sentences which at one time are regarded as inconceivable and so as non-cases, but which (e.g., after changes in scientific theories) come to be regarded as conceivable, though still impossible. Thus such sets of sentences may come to be regarded as cases at a later time. Take the notion of "faster than light travel", for instance. At one time, when the idea that light travelled at finite speed had not occurred to anyone, the notion itself made no sense. Nowadays it is generally regarded as physically impossible, but conceivable. Thus some inconceivable impossibilities may come to be regarded as cases.

Thus "conceivability" is not going to give us a decision procedure for determining whether a given set of sentences can be extended to something conceivable, and so to be a case. I like to think of cases as compact “micro worlds” that could have added features of conceivability and that may have the properties of incompleteness and inconsistency.

Since my goal is to give a semantics that is structurally similar to a Lewis-Stalnaker semantics for possible worlds, but with cases in place of worlds, I will need some account of the relationship of “closeness” among cases, since it plays a fundamental role in the Lewis-Stalnaker account.

As we have noted, there are two kinds of cases: possible (for which I will use the variables x, y etc. including the actual case a); and impossible ones (for which I will use the letters w, z). There are some details that needs to be worked out about these cases. What is the relationship between these two kinds of cases in terms of their closeness? The following are some questions that one could raise about these cases, in terms of closeness: Could it be that a is closer to w than z? In other words, could it be that the actual case is closer to one impossible case than another? Or more broadly, could it be that a possible case is closer to one impossible case than another?

More generally we can ask the following question: what is “closeness” in relation to

impossible cases? How do we spell out this notion in this new context? Before answering, we must take stock of "closeness" more generally. That is to say what does closeness amount to in possible world semantics?

Section 3. "Closeness" vis-à-vis Possible Worlds

The general agreement is that closeness involve some sort of similarity. But similarity of what kind? As Bennett points out, after Lewis's *Counterfactuals,* initially it was thought that closeness is all-in similarity. However that this was a misunderstanding was clarified in the paper entitled "Counterfactual Dependence and Time's Arrow" (Lewis, 1979). In Bennett's words:

> Lewis's theory evidently needs to be based not on untutored offhand judgments about all-in similarity, but rather on similarity relation that is constrained somehow—it must say that A → C (Bennett writes A > C) is true just in case C is true at the A-worlds that are most like the actual world *in such and such respects*. The philosophical task is to work out *what* respects of similarity will enable the theory to square with our intuitions and usage (Bennett, 2003, p. 196).

For example, it was pointed out that we want constraints on similarity relations so that it will allow some counterfactuals of the form A → Big-difference to be true. To make this clear consider Bennett's own example:

(1) *If on July 20 Stauffenberg had placed the bomb on the other side of the trestle, Hitler would have been killed*. Or

(2) *If at time T, the trajectory of asteroid X had been one second of arc different from what it actually was, the dinosaurs would have survived to the present*.

Counterfactuals like the above seem true. But if "closeness" amounts to overall similarity then we will have to consider counterfactuals of the form A → Big-difference like the above false. (1) is almost certainly true. But if we take closeness to mean overall similarity, any A-world at which Hitler dies on

July 20 1944, is less close to actual world a than some of those at which Hitler miraculously survives. Thus we are forced to make this counterfactual false. Again in case of (2), that dinosaurs still exist and roam the Earth seems less like the actual world than are some at which X miraculously swerves just after time T into a trajectory identical with its actual ones so that it hits our planet and extinguishes the dinosaurs. So, we want constraints on similarity relations that will allow counterfactuals like (1) to be true, "implying that some of the worlds at which Hitler dies on July 20, 1944 are more like the actual world in the relevant respects than are any at which Stauffenberg puts the bomb within reach of Hitler and the fuse fails. This must be achieved without also declaring true some conditionals that informed people are sure are false" (Bennett, 2003, p. 197)

There are two constraints that are placed on closeness by Lewis, making closeness of worlds amount to similarity in certain specified respects. The theory makes the truth value of A → C depend on whether C is true at the A-worlds that are

(1) like the actual world in matters of particular fact up to the antecedent time and

(2) perfectly like the actual world in respect of causal laws.

(1) says that we must compare closeness of worlds in respect of their states up to the time that the antecedent is about—call it TA. This could not be the whole theory because what a world is like up to a particular time implies nothing about what it is like later unless we take causal laws into consideration. Consider two worlds w1 and w2, in each of which a bomb with a fuse is placed at TA, with the fuse failing at one of them, and third world war ensuing in the other. The difference between war and no-war makes no difference to how close they are to the actual world because it pertains to the post-antecedent time. Alongside the bomb-world w1 which is like the actual world up to TA and happens to be the world with the third world war, there is a bomb-world w2 that is equally like w1 up to TA and happens to be the world at which the third world war does not obtain. If we consider

constraint (1) alone, nothing said so far about the constraints lets us choose between these two worlds. This is why we need (2) to bring causation into the story. The closest A-worlds must not only be like the actual world *a* up to TA but also must conform to the causal laws that govern *a*. In case of the Stauffenberg example,

> The bomb, the room, and the people were so structured and interrelated that the bombs being placed on Hitler's side of the trestle supporting the table would causally suffice for Hitler's death; so that the only way for Hitler to survive is through a miracle, a breach of the causal laws of a [Bennett writes α]. The proposed confining of ourselves to causally possible worlds is, precisely, the exclusion of all worlds in which miracles occur. So we have what we want: a theory that makes it true that if a bomb had been placed a foot to the right Hitler would have been killed (Bennett, 2003, p. 198).

This is a brief version of one of the important views on what closeness amounts to. It will suffice for my purposes in this thesis to assume that this account is more or less correct. Now that we have seen what "closeness" vis-à-vis possible worlds amounts to, let us look at what closeness vis-à-vis impossible cases amounts to.

Section 4. Closeness vis-à-vis Impossible Cases

We can take Lewis's idea of closeness as a jumping off point to construct a new idea of closeness in relation to impossible cases. However, we cannot just borrow the idea of closeness vis-à-vis possible world wholesale, keeping all the causal laws fixed, and all the matters of particular fact up until the time in question, and apply it to impossible cases. As we will see shortly, this is not compatible with the partiality of ways.

We want the new account to handle counterfactuals with various sorts of impossible antecedents, including causally impossible antecedents. So it is not easy to see how those are going to work if all the causal laws must be held fixed. For Lewis, such counterfactuals with impossible antecedents come out vacuously true, by the mere fact that there is no such world where the antecedent

is true. However, to determine counterfactuals like "If humans had three eyes, it would be harder to make glasses", where the antecedent is not metaphysically impossible one needs to look at worlds where the antecedent is true, if the consequent is true also. If we borrow Lewis's idea, then we would have to end up saying that counterfactuals with metaphysically or causally impossible antecedents are vacuously true. However, this is not what we want. We want some of these counterfactuals to be true and some of them to be false.

The real problem, I think, is that impossible cases, even logically or metaphysically impossible cases, will require a very different structure to the universe if things are going to be arranged "overall" so that they would be true. "If there were finitely many primes..." may be an antecedent for interesting counterfactuals, but if this were so then the laws of most areas of mathematics would have to be very different, and so too, presumably, would be the laws of physics. Since we can, as things stand now, encode all well-behaved formal languages via Gödel numbering, which involves the use of primes, much of linguistics and computer science will be false ... In other words, the partiality of the impossible ways is important for handling antecedents such as the above. For, we can not really take seriously the idea of "making things overall so that things can turn out this way". So the idea of partiality plays an important role in figuring out closeness.

The impossible cases are necessarily partial, and we can handle them only by insulating much of the rest of the "world" from the influence of the impossible bits. But what distinguishes counterfactuals like "If Dave squared the circle, he would be more famous than Gödel" from "If Dave squared the circle, the sun would explode" is that the revisions involved to actuality to make Dave famous in a "squares the circle" way need not require wholesale revisions to the actual laws and the actual facts—there is a practice of publication of mathematical theorems, and solution to long-standing and famous problems brings fame (of a sort) to mathematicians—and the changes involved in making

the impossibility true put Dave in a position to publish a theorem that is a solution to a long-standing and famous mathematical problem. On the other hand, to make the sun explode in a world where Dave squares the circle requires significant changes to the physical facts or the physical laws (and probably both), changes which have no direct connection to the changes needed to make it so that Dave has squared the circle.

So closeness of one case to another depends on whether the changes we need to allow for the consequent to be true in a case where we have already allowed changes for the antecedent to be true, are related to the actual case and whether these two sets of changes are connected or not. Thus the Dave-famous case is closer to the actual world *a* than the sun explodes case, because, in the actual case, publishing of papers relating to proving a theorem in mathematics brings you fame. So the changes we need for the impossibility of Dave squaring the circle and thereby becoming famous to be true is connected to the facts about the world where fame follows directly from publishing difficult theorems in mathematics.

Thus, we try to explain closeness in the following way: Let

A = Dave squares the circle

B= Dave will be famous

C = the Sun explodes

a = actual world.

Is this AB (Dave squares the circle and he is famous) case closer to *a* than AC (Dave squares the circle and the sun explodes) case? If it is, why? What does it mean for one case to be closer to the actual world than another? One can say the AB case is closer to *a* than the AC case because for AB to be true, the changes that we allow for A to be true are connected to the changes we have to allow for B to be true. It follows directly from the changes we need to envisage for the impossibility of Dave

squaring the circle to be true, that for such a major accomplishment, Dave would publish the results, because in the actual world people publish the results of major mathematical accomplishments like proving such a theorem, and in the actual world publishing such results bring fame. Therefore, in that AB case, it will bring fame to Dave and thus make the conditional true. However, for AC to be true, we need rather far fetched changes such as changes in the laws of physics for C to be true in a world where A is true. And the changes we need to envisage for A to be true are mathematically related whereas the changes we need for C to be true are causal, and as there is no link between the two sets of changes.

Notice that such an account makes room for a role for the actual case *a* in explaining which counterfactual is true. It seems likely that how things actually are should have some influence on which counterfactuals with impossible antecedents turn out to be true. This way of viewing closeness also makes room for capturing relevancy between the antecedent and the consequent, which is what Relevance Logics tries to capture. From "Dave squares the circle" one can infer that "Dave would be more famous than Gödel" , because he could publish the results, which would be more impressive than Gödel's incompleteness results (because up to this point squaring the circle was considered a mathematical impossibility, whereas the incompleteness theorem was not)….But from "Dave squares the circle" one cannot infer "the sun would explode" because proving a mathematical impossibility has nothing to do with the sun exploding.

Thus, some impossible cases are closer to the actual case than others. In other words the answer to our initial questions is "yes". Because it seems a case where Dave squares the circle and he is very famous (AB case) is closer to the actual case *a* than a case where Dave squares the circle and the sun explodes (AC case). In other words, impossible cases that are closer in some relevant crucial respects (i.e., the same laws of physics hold etc.) are far less dissimilar. This way of viewing

"closeness" in terms of relevant similarity, with regard to impossible cases, facilitates the distinction between counterfactuals like "if Dave squares the circle then he will be more famous than Gödel" and "if Dave squares the circle then the sun will explode". We can say that the former is true and the latter is false, if the closest and most relevantly similar Dave-squaring-the-circle impossible cases are also cases where he is more famous than most famous mathematicians. But Dave-squaring-the-circle impossible cases are not the cases where the sun explodes. And that is why they are dissimilar.

Notice that the difficulty with the Lewis-Stalnaker style account is going to be explaining what it means for one case to be closer to another. Let me explain this with the help of our original example: Consider why the Dave-famous counterfactual is true: any nearby case in which Dave squares the circle is also a case in which he is famous, because part of "closeness" in this case is that theorems are the kinds of things that get published and can make you famous. The content of the consequent links to that sort of feature of the world, so it is natural to include them in the specification of the case, even if cases can be incomplete. On the other hand, there is no link of content between the mathematical practice of constructing and publishing proofs and the sun exploding. So, one is not completely unconstrained in including what sorts of things should count as being held constant between two cases for them to count as nearby. Indeed, this is less problematic for analyses using cases than those using worlds—a notorious problem for Lewis and Stalnaker because they have to consider whole worlds, including the vast, uncountable realms of facts not relevant to what is actually stated in the antecedent.

Now that we have sorted out these crucial details about impossible cases, we are in a position to discuss the new theory for counterfactuals.

Section 5. A New Theory of Counterfactuals

What is this new theory of counterfactuals, and how does it manage counterfactuals with

impossible antecedents in a systematic way so that some of these counterfactuals come out true and some come out false?

In our new theory for counterfactuals we use the basic concepts of incompleteness and inconsistency made available by Relevance Logic, incorporate that with our concept of "cases" that have the added feature of conceivability (at least potential) and combine that with the basic ideas of Lewis-Stalnaker style analysis of counterfactuals to produce a new theory—one that manages counterfactuals with impossible antecedent in a fairly systematic way.

As we have noted before, the basic idea in Lewis-Stalnaker style analysis is the following:

In analyzing a conditional of the appropriate sort (i.e., P $\Box\!\!\rightarrow$ Q) one checks the *closest* possible worlds where the antecedent (P) is true, and if in all those possible worlds (where P is true), the consequent (Q) is true as well, the conditional in question is true.

On this new theory:

P $\Box\!\!\rightarrow$ Q is true at *w* iff there is a (P and Q) case *y* that is closer to *w* than is any (P and ¬Q) case *z*. If there are no P cases at which either Q or not Q is true, then P $\Box\!\!\rightarrow$ Q is indeterminate. Also iff for every P & Q case there is a closer P & ¬ Q case, then P $\Box\!\!\rightarrow$ Q is false at w.

So w $\vdash$ P $\Box\!\!\rightarrow$ Q iff there is some y such that:

*1. If $y \vdash P \wedge Q$, $y \nvdash \neg P$, $y \nvdash \neg Q$, and $z \vdash P \wedge \neg Q$ then *y* is closer to *w* than z.

There is another alternative:

*2. $y \vdash P \wedge Q$, $y \nvdash \neg P$, $y \nvdash \neg Q$, and if $z \vdash P$ and $z \nvdash Q$ then *y* is closer to *w* than *z*.

To see the difference this change would make consider the following counterfactual: "If Dave squared the circle then the earth would continue to circle the sun". Adopting *2 makes it easier to say that the counterfactual being considered is indeterminate since truth value of Q is unknown in *z*. But by adopting this analysis the risk we run is that on this formulation counterfactuals like "if Dave squares the circle then the sun will explode" also turns out indeterminate. This may not seem problematic. However, it is counterintuitive to the premise this thesis is originally based on. This thesis started with the intuition that counterfactuals like "If Dave squares the circle then the sun will explode" seems false and counterfactuals like "If Dave squares the circle then he will be famous" seems true. At this point I am tempted to stay with *1.

The price we pay for adopting this is that counterfactuals like "If Dave squares the circle then Monica (his daughter) will have toast for breakfast", and "If Dave squares the circle then the earth will continue to go around the sun" will come out true, though we can say that "If Dave squares the circle then the sun will explode" is false. But this is not a high price to pay. Recall that we define closeness in relation to actual cases. It is quite plausible that if Dave squares the circle then he will be so excited that he will offer to make breakfast that day and that he will decide to make French-toast for Monica or if Dave squares the circle then although it is a considerable mathematical achievement, the world will still largely be the same as it was before he squared the circle, and hence the earth will continue to circle the sun. How about the sun-explodes counterfactual then? It is also a case, as we said earlier, because of the example in which Dave is born on a distant galaxy long in the future, just before the sun goes nova. But it is false and the sun-explodes case is not as close as the Monica-having-toast case because the changes we need to allow for the consequent to be true, when the antecedent is true, have nothing to do with the actual case. There is no connection between the sun exploding and Dave squaring the circle. In the actual world, there is no connection between such things as publishing

mathematical theorems and stars exploding. Thus the way the actual world is determines which impossible cases are closer to a. Thus I am quite comfortable allowing for counterfactuals like "If Dave squares the circle, Monica will have toast for breakfast" coming out true, much like counterfactuals such as "If Dave squares the circle, then he will be famous".

Chapter V: Application of the New Theory to Thought Experiments

One of the key objectives of this thesis is to provide a tool that can analyze thought experiments involving impossible scenarios, such as the auditory world and fission, in a way that accounts for their usefulness/informativeness. As we will see this kind of thought experiments is deemed uninformative in two highly regarded book length treatments of thought experiments, due to Roy Sorensen and Tamar Gendler. In this chapter, I will show how by applying the new semantics for counterfactuals to these thought experiments we can explain their usefulness. With the help of these two and two other crucial examples of thought experiments (Ship of Theseus, and Chinese Room), I will show that the new semantics allows for a nice representation of what is at issue in the philosophical debates that surround them, and that it provides guidance to philosophers about what needs to be shown to make a compelling case about what they show.

This chapter is divided into five sections. In section 1 and 2 I will discuss Sorensen's and Gendler's accounts of thought experiments. Section 3 and 4, are critiques of the aforementioned accounts, in particular in relation to what they imply about impossibility. I show that these two accounts fail to handle thought experiments with impossible scenarios, because they imply that thought experiments with impossible scenarios are incoherent, and so fail to explain why some such thought experiments are much-discussed by philosophers, and generally regarded as potentially informative. In section 5 I will discuss how the new semantics for counterfactuals apply to the four example thought experiments (Theseus's Ship, Chinese Room, Fission, and Auditory World). I will also discuss the main pay offs of such an application.

The availability of the new account of counterfactuals brings two related advantages when it comes to thought experiments. First, for impossible scenario thought experiments, the new account can

explain how they can still be informative. Secondly, for thought experiments such as the Chinese Room, where it is not clear whether there is a subtle impossibility in the scenario or not, this new account with its continuous treatment of possible and impossible cases makes clear why the debate about them looks the way it does. The crucial question about the Chinese Room is not whether there is such an impossibility, but what is the "nearest" situation in which there is a Chinese Room (whether it is impossible or not) and what we would say in such a case (about the intentionality of the room). On traditional accounts, including Gendler's and Sorensen's, it ought to be paramount to deal with the possibility question, because if it is an impossible scenario the lessons we learn are very different from the ones we learn if it is possible.

Thus, there are two related deficiencies in the standard book length accounts. First, they fail to account for incomplete/impossible yet informative thought experiments. Secondly, even for the possible ones, they are unable to explain, at least in some interesting cases, what are the important philosophical issues at stake.

Section 1. Sorensen

Sorensen, in his book *Thought Experiments*, aims to present a general theory of thought experiments that would account for: what they are, how they work, their virtues and vices. Based on the definition of an experiment as "a procedure for answering or raising a question about the relationship between variables by varying one (or more) of them and tracking any response by the other or others" (Sorensen, 1992, p. 186), Sorensen defines a thought experiment to be "an experiment that purports to achieve its aim without the benefit of execution" (Sorensen, 1992, p. 205). By “purports to achieve its aim without execution”, Sorensen means that “the experimental design is presented a certain way to the audience. The audience is being invited to believe that contemplation of the design justifies

an answer to the question or (more rarely) justifiably raises its question" (Sorensen, 1992, pp. 205-206).

Thus, Sorensen's view is that the thought experiment constitutes a "limiting case" of experiments and thought experiments can achieve their aim without being executed. There are five models of thought experiment that are suggested by philosophers and psychologists, claims Sorensen. The five models are the following: *the recollection model, the transformation model, the homuncular model, the re-arrangement model,* and *the cleansing model.* Thought experiments are categorized thus according to the purpose they serve in "improving the epistemic state of the thinker without the addition of new information".

Many thought experiments that fall into the category of *the recollection model* serve their purpose as reminders. These function by bringing about the recollection of previously acquired empirical knowledge. Sorensen's example of the following standard thought experiment in physics is a case in point: This thought experiment asks us to picture a pilot who forgets to fasten his seat belt when performing a front-back loop. At the top of the loop, when the pilot is upside down, does he fall down? One is naturally inclined to answer "yes" since gravity should pull the pilot down. However, one figures that the question would not have been asked if the answer was so simple. So one examines the forces that might counteract gravity. Many are led to think of centrifugal force because the pilots hypothetical stunt *reminds* them of carnival rides. Now the similarity between the pilot's stunt and a carnival ride is increased by imagining that the pilot is doing a left-right loop. The pilots feet would press against the floor, he would not be leaning just on his side. This similarity leads the thinker to a negative answer for the vertical loop case: centrifugal force would keep the pilot on the plane.

Thought experiments in the *transformation model* achieve their results by codifying knowledge of linguistic rules. Explaining what *transformation model* thought experiments are, Sorensen says: "Linguistic philosophers portray the a priori refinement of implicit knowledge as

proceeding from knowing *how* to knowing *that*. The idea is that we all have knowledge about how to speak the language in which we are philosophizing. Since philosophical theses concern meaning and meaning is governed by rules of the language, we can settle philosophical questions by codifying our mastery of linguistic rules" (Sorensen, 1992, p. 92). As an example, Sorensen cites Harry Frankfurt's refutation of 'A person is responsible only if he could have done otherwise'. Frankfurt asks us to imagine that a scientist has wired up a man with a fail-safe device. The device will cause the man to do a bad thing if he does not do it on his own. As it turns out, the device is not activated because the man does the wicked deed on his own. Now, is the man responsible? Since we are inclined to describe him as responsible for the deed, our mastery of English give us evidence against the entailment rule. Knowledge of how to use 'responsible' thus transforms into knowledge that responsibility is compatible with the inability to do otherwise.

The *homuncular model* thought experiments are mental information processing as if by homunculi. Sorensen here endorses Dennett's idea of viewing the agent as an imperfectly coordinated complex of cognitive systems---a crew of homunculi (little men). According to Sorensen, this model suggests that the imagination will have a distributional role. Some thought experiments in this model reveal hidden disagreements within the "internal committee". Sorensen provides an illustration of this model through the following: consider the belief that <u>all harms must make a discriminable difference to their victims</u>. Jonathan Glover, with the help of the following thought experiment, shows that this belief conflicts with our belief that <u>little differences can add up to a big difference</u>:

> Suppose a village contains 100 unarmed tribesman eating their lunch. 100 hungry armed bandits descend on the village and each bandit at gun-point takes one tribesman's lunch and eats it. The bandits then go off, each one having done a discriminable amount of harm to a single tribesman. Next week, the bandits are tempted to do the same thing again, but are troubled by new-found doubts about the morality of such a raid. Their doubts are put to rest by one of their number who does not believe in the principle of divisibility. They then raid the village, tie up the tribesman, and look at their lunch. As

> expected, each bowl of food contains 100 baked beans. The pleasure derived from one baked beans is below the discrimination threshold. Instead of each bandit eating a single plateful like last week, each takes one bean from each plate. They leave after eating all the beans, pleased to have done no harm, as each has done no more than sub-threshold harm to each person. (Glover, 1975, pp. 174-175)

This thought experiment helps to bring out the contradiction in two beliefs—disagreement among the internal committee of homunculi.

The *re-arrangement model* thought experiments re-arrange information into more convenient formats. "This model is inspired by situations in which the information at hand is made more digestible by changing its form" (Sorensen 1992, p. 99). As an example, Sorensen provides James Rachels's thought experiment that serves a critique against American Medical Association's position on euthanasia. Association policy forbids active euthanasia ("mercy killing") but permits passive euthanasia. According to Rachels there is no morally relevant difference between killing someone and letting someone die. Rachels provides the following scenario in which he places a pair of hypothetical bad moral deeds side by side. The first involves a man, Smith, who will inherit a large estate if his six-year-old cousin dies. One evening while the boy is taking a bath, Smith sneaks in and drowns the boy. The second involves a man, Jones, who, like Smith, has a six-year old cousin standing in the way of Jones inheriting a large estate. Jones, like Smith, decides to kill his cousin. However, when Jones sneaks in to kill the boy, he sees the boy slip, hit his head, and land with his face in the water. The delighted Jones stands over the boy, ready to push him under if he recovers. But the boy drowns on his own. Smith killed his cousin. Jones merely let his cousin die. But having controlled all the extraneous variables carefully, we see that the distinction does not make a moral difference. Thus if we rearrange the information in new ways, into a more convenient format, the information is made more digestible.

The *cleansing model* thought experiments expose and eliminate acts of irrationality in belief-

formation. This includes a familiar situation where an inconsistency is noticed, and is then weeded out (Sorensen, 1992, pp. 88-109). "The *cleansing model* is inspired by incidents in which you recognize your own irrationality and then change your beliefs to remove the flaw" (Sorensen, 1992, p. 104). According to Sorensen, Plato's allegory of the cave serves to demonstrate this. The allegory of the cave demonstrates that Plato's theory of forms could be right although it contradicts common sense.

The difference between the first four models and the cleansing model is that "whereas the previous models cast the epistemic improvement as a matter of adding positive features, the cleansing model concentrates on subtracting negative features" (Sorensen 1992, p. 104). Sorensen further adds: "These negative features are intellectual vices that diminish your efficiency at tasks such as argument, explanation, inquiry, prediction, planning, problem solving, and teaching" (Sorensen, 1992, p. 104).

None of these five models preclude the others, according to Sorensen. So, he says, "we are free to pick and mix. There are thought experiments conforming to each—and some that fit all the models simultaneously" (Sorensen, 1992, p. 109). Sorensen acknowledges that although all these five models may have some application, he believes that only one can be elaborated on further. For developing the first four would require psychological theories more sophisticated than we at present possess. Given this situation, he deems only the fifth, i.e., the cleansing model, suited for development. In his words:

> Although all of the models have applicability, only one has prospect of immediate elaboration. We have a very limited understanding of how the mind works. This psychological obscurity fogs in the positive models. We can say that thought experiments function as reminders, transformers, autosimulators, and rearrangers, but we cannot go much beyond that. Future progress may enable us to go further. But for now, the positive models only provide vague sketches of how thinkers improve without new information (Sorensen, 1992, p. 109).

Thus, he goes on to develop an account of *cleansing model* thought experiments. Henceforth

when we talk about Sorensen's account of "thought experiments" we mean Sorensen's account of "*cleansing model* thought experiments".

Sorensen contends that, in general, thought experiments are reactions to inconsistencies in sets of statements. More specifically, thought experiments aim at refuting a statement (which he calls the thought experiment's "source statement") by disproving its alethic modal consequences, i.e., those consequences of the statement having the form "It is necessary/possible that p". In those cases where the alethic modal consequence of the statement is of the form "it is necessary that p"—a source statement that entails that p holds in all possible worlds—a successful thought experiment involves finding a possible world in which p is false. Conversely, in cases where the alethic modal consequence is a statement of the form "it is possible that p"—a source statement that implies that p holds in some possible world—a successful thought experiment involves establishing that there is no such possible world.

Sorensen claims that any argument purporting to refute a source statement by disproving one of its modal consequences can be laid out in a standard form, consisting of five propositions that are jointly inconsistent, of which one is the source statement. "Necessity refuters" and "possibility refuters" are what Sorensen calls the arguments that purport to overthrow the source statement that implies that p holds in all possible worlds, and p holds in some possible worlds respectively. This can take the following form:

(1) Modal source statement:

(2) Modal extractor/possibility extractor: this proposition draws the relevant modal implication/a possibility consequence from the source statement.

(3) Counterfactual: this proposition claims that the antecedent which is the conjunction of the implication and the imagined situation, has a weird consequence.

(4) Absurdity: this proposition explains the weirdness as an impossibility.

(5) Content possibility/content copossibility: this asserts that the content of the thought experiment is a possibility/that the statement extracted at 2 is true only if it is compatible with the content of the thought experiment.

The joint inconsistency of the above statements imposes an obligation to deny at least one of the statements. While the successful thought experiment establishes that it is the source statement that should be denied, unsuccessful thought experiments lead to the denial of other statements in the set. On the basis of these observations Sorensen proposes a taxonomy of thought experiments according to the particular member-statement of the set which they undermine (Sorensen, 1992, pp. 135-60).

To see how a thought experiment can be translated into the above form, consider Gettier's thought experiment. Gettier attempted to overthrow the "JTB" definition of knowledge (the definition that knowledge is justified true belief). According to JTB, A knows that p if and only if (1) A believes that p, (2) A is justified in believing that p, and (3) p is true. Gettier's objection was that the definition is too broad. He proved it through the following imaginary situation: Imagine that Smith and Jones are candidates for the same job. Now imagine that Smith acquires a justified belief that

(a) Jones is the man who will get the job, and Jones had ten coins in his pocket.

For example, Smith's evidence might be that the president of the company told Smith that Jones will get the job and that Smith previously counted the coins in Jones's pocket. Smith notices that (a) entails:

(b) The man who will get the job has ten coins in his pocket.

And thus comes to the justified belief that (b). However, despite what the president said, Smith is the man who will get the job and, as it happens, Smith has ten coins in his pocket. Hence (b) is true, and Smith justifiably believes it. Yet Smith does not know (b). Too much luck was involved.

Thus JTB does not hold.

This thought experiment can be put in the following form:

(1) The definition of knowledge is justified true belief.

(2) If knowledge is justified true belief, then necessarily, if a person has a justified true belief that p, then he knows that p.

(3) If all justified true believers that p have knowledge that p and Smith is justifiably right but for the wrong reason, then Smith knows that (b) because of luck.

(4) It is impossible for anyone's knowledge to be due to luck

(5) It is possible for Smith to be justifiably right for the wrong reason.

Most epistemologists agree that the Gettier cases are counterexamples to the JTB definition of knowledge. So they reject the first member of the set. Some deny that the possibility of a justified false belief and thus deny the fifth member of the set on the ground that Smith was not justified in believing in (b).

This in a nutshell is Sorensen's theory of thought experiments. Now we come to Gendler's account.

Section 2. Gendler

Tamar Gendler, in her recent work, *Thought Experiment: On the Powers and Limits of Imaginary Cases*, discusses how imaginary cases provide (wherein lies their power) or fail to provide new knowledge (wherein lies their limits). My interest is not to provide a synopsis of Gendler's book. Thus I will not be discussing all the valid points she makes about thought experiments and criticisms she offers of her predecessors. Instead, I want to discuss certain crucial and novel points that she makes about thought experiments.

The following, according to Gendler, characterizes the fundamental structure of a thought experiment:

(1) An imaginary scenario is described.

(2) An argument is offered that attempts to establish the correct evaluation of the scenario.

(3) This evaluation of the imagined scenario is then taken to reveal something about cases beyond the scenario (Gendler, 2000, p. 21).

Gendler finds the negative counterpart of this characterization particularly useful in classifying criticisms directed at thought experiments and their real-world import. The grounds of attack are the following:

(1a) *Unimaginability*: the scenario described is not (fully) imaginable.

(2a) *Unsound argument*: although the scenario described is imaginable, the argument establishing the correct evaluation of the scenario is unsound.

(3a) *Inapplicability*: although the scenario described is imaginable, and the argument establishing the correct evaluation of the scenario is sound, the conclusion does not reveal what the author takes it to reveal about the actual world (Gendler, 2000, p. 22).

Gendler also provides a taxonomy for thought experiments. Depending on three basic sorts of questions that different sorts of thought experiments are trying to answer, we can divide thought experiments into three kinds. The questions are the following:

(1) What would happen,

(2) How, given (1) (i.e., what would happen), should we describe what would happen,

(3) How, given (2) (i.e., how should we describe what would happen), should we evaluate what would happen.

The first type is called *factive* (e.g., Galileo), the second *conceptual* (e.g. The Ship of

Theseus), and the third *valuational* (e.g. Fission). In the first type we are concerned about what the facts of a situation would be, in the second, what would we take the proper application of the concepts to be, and the third, what would be the proper moral or aesthetic response to a situation. Gendler also cautions us that the line between these types of thought experiments may sometimes collapse.

According to Gendler, the powers and limits of a thought experiment depend on whether or not the case provided by a thought experiment is correctly treated by the theory, i.e., the thought experimenter maintains a distinction between theories with "norm-driven-exceptions**",** and theories with "exception-driven-norms". By theories with norm-driven exceptions and theories with exception-driven norms, Gendler means the following: There are two main ways in which a theory can explain exceptional cases. In other words, there are two main strategies users of a theory can employ in explaining exceptional cases. According to Gendler, "[T]he first strategy is to use exceptional cases as a way of progressively narrowing the range of privileged characteristics… According to this strategy, one uses exceptional cases to ascertain the theory's exception-driven norms. The exceptions drive interpretation of the norms; what is taken to matter about normal cases is whatever it is that they have in common with exceptional cases…" (Gendler, 2000, p. 8).

So, let's suppose, following Gendler, that entities under inspection by the theory in question generally have the characteristics *a*, *b*, *c*, *d*, and *e*. Suppose further that an exceptional entity is found that falls within the purview of the present theory, but which has only characteristics *b*, and *d*. Under this strategy, it follows that no characteristics other than *b* and *d* can be *privileged* characteristics in the sense that they necessarily belong to any entity that falls within the purview of the theory. "Such an attitude towards exceptional cases involves using them as test cases to ascertain necessary and sufficient conditions", says Gendler (Gendler, 2000, p. 8). The conclusion one draws is that even in non-exceptional cases, the characteristics that really matter are those that are present in the exceptional

cases as well, viz., *b* and *d*, in this case.

The second strategy is to view exceptional cases as evidence for the strength of the theory's core. "On the basis of this strategy, one concludes that what it is that allows the exceptional cases to be cases at all is that they have enough in common with the normal cases". So, for example, suppose again that entities under the theory in question generally have the characteristics *a*, *b*, *c*, *d*, and *e* and suppose further that some entity is found that has only *b*, and *d*, but which nonetheless seems to fall within the purview of the theory. According to the second strategy, what we ought to say about the entity in question is that it falls within the purview of the theory, but only because it is similar in certain crucial ways to more typical instances of entities under the theory. Under this strategy, one uses exceptional cases to ascertain the theory's norm-driven exceptions. It is the norms that drive the interpretation of the exceptions (Gendler, 2000, p. 9).

Gendler contends that both the powers and limits of imaginary cases can be traced back to the fact that when such contemplation brings us to new knowledge, it does so by forcing us to make sense of an exceptional case (Gendler, 2000, p. 12). Why is this so? She writes:

> My answer is three-fold: First, thinking about exceptional cases can lead us to a reconfiguration of our conceptual commitments, allowing us to organize information in a way that renders it newly meaningful. Moreover, exceptional cases are good test cases; they help prevent us from mistaking accidental regularities for regularities that reflect a deeper truth about the world. But, third, exceptional cases are dangerous; if we fail to keep straight the distinction between theories with norm-driven exceptions and theories with exception-driven norms, we are likely to draw radically misguided conclusions (Gendler, 2000, p. 12).

Galileo's thought experiment (see Section 3. Chapter I, for a detailed description), according to Gendler, brings new knowledge to the Aristotelian by allowing him to see all cases involving falling bodies through the lens of an exceptional case of strapped bodies wherein the Aristotelian realizes, through the paradox, the inconsistencies in his own theory. In this case the exception dictates the

interpretation of the norms. This is what explains the thought experiment's power in this case.

In the fission case, precisely the same feature [i.e., the feature that exceptions are taken to drive the interpretation of the norms] explains the thought experiment's limitations (Gendler, 2000, p. 159). According to Gendler, the fission case is an obviously exceptional case. It is meant to be unusual/extraordinary. She argues that to make sense of such cases they have to be treated as extraordinary. Otherwise, we are bound to draw "radically misguided conclusions". Our ability to draw any meaningful lessons from such cases relies on our ability to draw lessons from ordinary cases. So, such cases are powerful insofar as we treat them as norm-driven-exceptions rather than as exception-driven-norms.

As Gendler points out in the fission case, a process (the process of transplanting Brainy's brain into Lefty's and Righty's bodies) that is normally identity-preserving would turn out to be identity creating. That is, in the "single-transfer case" the process would result in continued existence of some entity over time. But the in the "double-transfer case" it would end up in creating two new entities. But if these entities are self-conscious, as human beings are, then we are faced with the following puzzle: to the extent that we are talking about the same process (intrinsically), in both cases, how could the rationality of one's attitude towards one's continuer differ from the single-transfer case to the double-transfer case? In other words, how could one's attitude depend on whether the process ended up preserving one's identity (in the single transfer case) or whether the process ended up creating two new individuals? Presumably, one's attitude towards one's continuer would be the same in both the single-transfer and the double-transfer case. Up until now, Gendler agrees with Parfit. However, what Parfit concludes from fission is that what makes my prudential concern for myself tomorrow rational is not the fact that myself-tomorrow will presumably be identical to myself-today, but only that she will be connected to me by the right sort of causal process i.e., ones that will result in the right sort of

relation of psychological continuity and connectedness. So personal identity is not what matters or, to state the point in Gendler's terms, the relation which matters for rational prudential concern is not identity. This is because in the double-transfer case what matters for rational prudential concern is present but identity is absent. So, if the former can obtain without the latter, then identity cannot be what matters for rational prudential concern.

Gendler argues against this conclusion and claims that fission shows much less than what Parfit takes it to show. "It shows only that there are conceivable circumstances where it might be rational to bear a relation of prudential concern towards a continuer with whom one was not identical" (Gendler, 2000, p. 147). The larger lesson to be drawn from fission, according to Gendler, is the following: in the case of fission, we are asked to imagine a scenario where "a pair of features that coincide in all actual situations are imaginatively separated" and we are asked to make a judgment about which of the two features has conceptual primacy. The proper interpretation of the case is exactly the opposite of what it is taken to be. We are not to make a judgment about the normal cases on the basis of the exceptional cases. The reason why fission's implications have been misunderstood is the following:

> [c]ertain patterns of features which coincide only fortuitously may nonetheless play a central role in the organization of our concepts. To the extent that imaginary scenarios involve disruptions of these patterns, our first-order judgment about them are often distorted or even inverted (Gendler, 2000, p. 147).

The Ship of Theseus is another example of an obvious exceptional case in the sense of being extraordinary. It presents us with a case in which a seemingly entity-preserving process apparently becomes entity-creating after sufficient iteration. According to Gendler, it would be wrong to draw the conclusion in this case that such processes are *ordinarily* entity-creating. "For, if cases like Theseus were the norm rather than the exception, it would not make sense even to speak of identity-candidacy

for ships. Ships would be like amoebae or cloud formations or World Wide Web sites—messy sorts of entities with obscure criteria for individuation and persistence" (Gendler, 2000, p. 155). So, according to Gendler, to properly interpret this thought experiment would be to say that this is a case of a norm-driven-exception. And we can make sense of such an unusual case only against the background of normal cases. Hence, thought experiments are powerful or limited depending on what conclusions can legitimately be drawn from them, according to Gendler.

Section 3. Sorensen's Diagnosis vis-á-vis Impossible Thought Experiments

Since Sorensen's account is in essence an account of the *cleansing model* thought experiment, it is not clear how this account applies to some of the impossible-scenario thought experiments under consideration in this thesis. According to Sorensen, the *cleansing model* thought experiments work by exposing and eliminating acts of irrationality in belief-formation. This includes a familiar situation where an inconsistency is noticed, and is then weeded out. However, thought experiments such as the auditory world are not aimed at exposing and eliminating irrationality in our belief formation about objectivity. The auditory world is construed with the aim to expose previously unnoticed ties between our concepts in our conceptual scheme of the real world.

The "source statement" under attack in this thought experiment is that 'space is a prerequisite for objectivity". However, Sorensen will deem this thought experiment not successful. For, according to Sorensen, in those cases where the alethic modal consequence of the statement is of the form "it is necessary that p"—a source statement that entails that p holds in all possible worlds—a **successful** thought experiment involves finding a possible world in which p is false. Conversely, in cases where the alethic modal consequence is a statement of the form "it is possible that p"—a source statement that implies that p holds in some possible world—a successful thought experiment involves establishing

that there is no such possible world.

Applying Sorensen's account to the auditory world thought experiment, the source statement is "necessarily, if there is objectivity, then there is space". So Strawson needs to show that there is a possible world in which there is objectivity but no space. Sorensen would have to say that the auditory world thought experiment is unsuccessful because Strawson is unable to establish whether there is a possible world (namely the auditory world) in which P (space is a prerequisite for objectivity) is false. After all, Strawson does abandon the thought experiment. So this might be seen by Sorensen as having failed to establish a necessity refuter. However, it is clear that Strawson's goal in this thought experiment is not to establish that there is a possible world in which P is false. All he wants to do is to find out more about objectivity in the real world, by construing this auditory world. His aim is not to refute that “space is a prerequisite for objectivity”; rather, his aim is to test whether it is the case. The auditory world thought experiment teaches us valuable lessons even though it is ultimately not fully and coherently describable.

The Chinese Room thought experiment is also a challenge for Sorensen's account. As far as the original thought experiment goes the source statement would be: "It is possible that there is a computer with intentionality", and Searle wants to use the Chinese Room to show that this is false (i.e., he wants to prove "necessarily, there are no computers with intentionality"). And Sorensen's way of looking at the matter requires that the Chinese Room be possible, so that it can be an appropriate analogue of possible computers. However, we will see in our discussion of the Chinese Room in Section 5 that Dennett and others have (in effect) pointed out that it doesn't really matter whether the Chinese Room is possible; what matters is what we should say in a case in which there is a Chinese Room (whether it is possible or not). It is not clear what one would say, if we sit down to consider the details. What comes out of the debate regarding the Chinese Room, as we will see, is that there might

be some subtle impossibilities built into the Chinese Room. For example, speed is an important factor for intelligent system. If there is a person manipulating symbols using a rule book, the speed in which the Chinese Room would operate would be too slow to regard it an intelligent system. So perhaps in order for it to appear intelligent the person inside the Chinese Room would need to operate at a rate that would be humanly impossible. Applying Sorensen's account to the Chinese Room, the source statement would be: "it is possible to have computers with intentionality". Searle with the thought experiment wants to establish: there are no possible cases of computers with intentionality. Now, if Searle's argument works, Sorensen would have to say that his argument involves the claim that if any computer has intentionality, then Chinese Room has intentionality, in order for Searle to justify the claim that since the Chinese Room doesn't have intentionality, no computer does." This argument depends on the Chinese Room describing a *possible world* (i.e., the Chinese Room is possible), if the point is to show "no possible computer". But this is not what the debate turns on, judging from what various philosophers say. Instead the debate is about what to say about Chinese Room, whether or not it is possible.

Section 4. Gendler's Diagnosis vis-á-vis Impossible Thought Experiments

How well does Gendler's diagnosis for failure of thought experiment fare when it comes to thought experiments with impossible scenarios, especially in difficult (i.e., non-obvious/non-standard) cases? Here I will show that Gendler's account fails to account for certain difficult (non-obvious) cases. First, let us consider Strawson's auditory world thought experiment.

This thought experiment is special in that it has an impossible scenario and is non-standard. The two main questions Strawson tries to answer with the help of this thought experiment are: what are the conditions for making a distinction between oneself and states of oneself on the one hand, and what

is not oneself and/or a state of oneself on the other [this is what Strawson calls "non-solipsistic consciousness"]?; and how are these conditions fulfilled? What motivates Strawson to construct the auditory world thought experiment is that he wants to explore the possibility of a non-solipsistic consciousness that does not rely on material bodies as basic particulars.

Before going on to criticize Gendler, I need to say a few words about why the auditory world thought experiment is impossible in a non-obvious/non-standard way. It is non-standard compared to the other thought experiments discussed in this thesis for the following reasons:

(a) It is non-standard in terms of its formulation. Strawson introduces it to test the thesis that space is a pre-requisite for objectivity. He does not clearly state the results this thought experiment produces. He uses it as a test case to put pressure on our ordinary conceptual scheme regarding objectivity and keeps it on the sideline to remind us about how much we take for granted in our ordinary conceptual scheme.

(b) It is also non-standard in that the author himself expresses doubt as to whether the thought experiment is completable.

Gendler would have to dismiss Strawson's auditory world and deem it not useful, on the count of unimaginability. Since the auditory world is a conceptually impossible scenario, it is unimaginable. However, although Strawson himself admits that the auditory world is not fully imaginable, he still finds the thought experiment useful in testing the thesis that space is a prerequisite for objectivity. Strawson admits that it might be better to abandon the auditory world because it is not clear *how* the conditions for the distinction between what is oneself and what is not oneself, in the auditory world, are to be fulfilled. In his words: "…the fantasy, besides being tedious, would be difficult to elaborate. For it is too little clear exactly what general features we should reproduce, and why. It might be better at this point to abandon the auditory world…" (Strawson, 1959, p. 85). However this thought experiment is still useful, not for drawing conclusions about what would happen in such-and-such a case, but as a

test case. Strawson's auditory world is a non-obvious case because it does not fall into the categories of a standard thought experiment in that Strawson's intention is not to draw any direct lessons from the auditory world. Rather he wants to use it as a test case for our conceptual scheme about the actual world. In his own words:

> …there is a certain advantage in keeping before our minds the picture of the purely auditory world, the picture of an experience very much more restricted than that which we in fact have. For it may help to sharpen for us the question we are concerned with; it may help to give us a continuing sense of the strangeness of what we in fact do; and this sense of strangeness we want to keep alive in order to see that we really want to meet it and remove it, and do not just lose or smother it (Strawson, 1959, p. 88).

The auditory world is constructed as a case where bodies are absent. By constructing an auditory world Strawson raises the question, whether we could make sense of the idea of a conceptual scheme which provided for objective particulars, but in which material bodies were not basic particulars. Strawson claims that although some of the conditions of such a scheme could be fulfilled in the terms of the thought experiment, in order to satisfy them all, "we should have to reproduce, in the restricted sensory terms available, more and more general features of the actual world". Thought experiments such as the auditory world, according to Strawson, are "not constructed for the purpose of speculation about what would really happen in certain remote contingencies. Their object is different. They are models against which to test and strengthen our own reflective understanding of our own conceptual structure" (Strawson, 1959, p. 86).

Strawson urges the reader to keep Hero close by, on the sidelines of one's thoughts while thinking about the concept of an ordinary person. For Hero is an oddity in a world of auditory flux. Confronting oddities in the conceptual scheme of the auditory world help us explore connections in our own framework. According to Strawson the auditory world helps in the following way: Hero is a member of the auditory world and he is just a sound. What we need to figure out is what conditions

need to be satisfied in order for Hero to be more—to be a subject of his own experience. Thus, the question is: what are the conditions requiring fulfillment for a non-solipsistic consciousness. Or how are the conditions of subjective/objective experience fulfilled? From this question we can ask the more general question: where do I get the idea of myself as a subject that has experiences of things that are other than myself? The auditory world proves useful in so far as we can extrapolate details about our own conceptual scheme. Thus the use Strawson makes of this thought experiment demonstrates the usefulness of a thought experiment about a scenario which is not fully imaginable.

Gendler's theory fails on more accounts when it comes to applicability to the auditory world thought experiment. Applying Gendler's tools, one might think that Strawson's thought experiment could be open to more criticism. According to Gendler, the powers and limits of a thought experiment depend on whether or not the scenario inquisition is correctly treated by the theory i.e., whether the experimenter maintains a distinction between theories with "norm-driven-exceptions**",** and theories with "exception-driven-norms". Gendler is occupied with determining how new knowledge is produced by a thought experiment in the context of an established theory. So her account is formulated in a way that only applies to those thought experiments with an underlying theory. This is another limitation on the applicability of her theory since for many thought experiments, in particular many involving impossible scenarios, there is no pre-existing underlying theory.

The first impossible thought experiment in hand is Strawson's auditory world thought experiment. There does not seem to be the required theory with respect to which one can decide how the auditory world is to be treated. While a "theory of objectivity" may be a candidate "theory" lurking in Strawson's enterprise, as Gendler might point out, I am not sure that it serves as a theory in the manner that Gendler promulgates. In other words, there is no clear cut underlying theory that would treat the auditory world as either a norm-driven-exception or an exception-driven-norm.

Theseus's Ship is a thought experiment discussed by Gendler where a theory of identity is the underlying theory for which the thought experiment is a test case. We can decide in this thought experiment, according to Gendler, whether the thought experiment is powerful or not on the basis of how the underlying theory, viz., the theory of identity in this case, is able to handle the thought experiment. The case acts as a test case for an already established theory, viz., the background theory of identity.

It is at least not clear what the underlying theory is in the auditory world thought experiment. It won't do to suggest, on Gendler's behalf, that the underlying theory is some implicit theory of objectivity. Unlike with Theseus's Ship, Strawson's main motivation for introducing this thought experiment is not to test an already established theory but to find out what this theory is in the first place—what the various concepts are that serve as the building blocks for the theory of objectivity and how they are related. So, the theory of objectivity doesn't serve the same purpose in Strawson's enterprise as the theory of identity does in the Ship of Theseus case.

Another paradigmatic example of such a thought experiment is Newton's bucket thought experiment. This thought experiment serves as another case that shows that it is no trivial limitation that Gendler's account only applies to those thought experiments with an underlying theory. To recap briefly, in this thought experiment we are to think away all the rest of the material universe except a bucket of water which goes through the following three distinctive stages:

(1) there is no relative motion between the bucket and water, the water surface is flat.

(2) there is relative motion between water and bucket.

(3) there is no relative motion between water and bucket; the water surface is concave.

This is the phenomenon produced by this thought experiment that needs to be explained. The explanation Newton gives is the following: in case 3, but not 1, the bucket and water are rotating with

respect to absolute space. According to Brown, absolute space is not *derived* from the thought experimental phenomenon, rather it is *postulated* to explain it (Brown, 1991, p. 40). For our discussion the particular explanation itself is not important. What is important is the fact that the explanation is postulated to explain rather than derived from the thought experiment.

I have already argued that the auditory world is an impossible scenario. Arguably, Newton's bucket thought experiments is physically impossible. For we are to imagine the rest of the material universe away. One could then ask, where is the bucket hanging from? Some might retort by saying that there is a possible world with only a bucket hanging from a rope. We could still argue that there would still be gravity etc. acting on the bucket if the water is to stay in it. This matters because it seems that in many thought experiments with impossible scenarios there is no underlying theory for which these thought experiments serve as test cases. Thus we cannot assess the limit or powers of these thought experiments against an underlying theory, in the way that Gendler recommends. Hence, her theory is particularly ill-suited for accounting for impossible-scenario thought experiments.

The question that one could ask is : Why is having no background theory more likely the case with impossible-scenario thought experiments than with possible-scenario ones? In other words, what is the connection between impossibility and lack of a background theory? Impossible scenarios often serve the purpose of bringing some previously unthought/unanalyzed concepts or theories to the forefront. Hence they do not necessarily try to overthrow or question well-established theories. This is not the case with possible-scenario thought experiments. One of the reasons people are likely to employ such thought experiments is in the development of new theories, or as justification for the structure and presuppositions made by a new theory, rather than as a "test" of an existing theory. In Strawson's developing an explicit account of objectivity, and Newton's justifying the presupposition of absolute space in his formulation of his mechanics, they are both trying to isolate factors relevant to

some fundamental question from potentially misleading other factors, and it is the consideration of the factors in isolation that makes the scenarios described impossible.

Another reason why Gendler's account is not applicable to cases such as Strawson's is the following: One might think, using Gendler's tools, that Strawson's thought experiment is open to criticism because he does not recognize that if it makes sense at all it must be treated as an exceptional case—especially one in which the norms drive interpretation of the exceptions. We should not derive any lessons about objectivity in the real world, as Strawson does, from this thought experiment, Gendler would have to say. On the contrary, we can make sense of objectivity in the auditory world only against the backdrop of objectivity in the real world. So, it seems that the criticism of inapplicability could be raised against Strawson.

However, such a diagnosis would defeat the very purpose of this particular thought experiment. The main purpose behind Strawson devising this thought experiment is to learn more about objectivity in the real world. At the end of the chapter entitled "Sounds" Strawson says:

> My real concern is with our own scheme [of the actual world], and the models of this chapter are not constructed for the purpose of speculation about what would really happen in certain remote contingencies. Their object is different. They are models against which to test and strengthen our own reflective understanding of our own conceptual structure (Strawson, 1959, p. 86).

We ought to remember that *Individuals* is an essay in *descriptive metaphysics*, as the subtitle indicates. According to Strawson, as opposed to *revisionary* metaphysics, *descriptive* metaphysics is content to describe the actual structure of our thought about the world rather than producing a better structure for our thought about the world. Descriptive metaphysics aims "to lay bare the most general features of our conceptual structure" and thus, "it can take far less for granted than a more limited and partial conceptual inquiry" (Strawson, 1959, p. 9). In this context the purpose of introducing the

auditory world is not to consider it as an exceptional case—especially in the sense that it is a case in which the norms drive interpretation of the exceptions.

For these reasons, I think it is fair to say that Gendler's theory cannot account for thought experiments that are impossible or uncompletable, yet useful/informative. It seems to me that unlike scientific thought experiments (Galileo's falling body thought experiment, for example), many thought experiments (especially the ones popular in Philosophy) are fraught with impossibility and unimaginability. A prime example is fission, not in the sense we have discussed frequently in this thesis but in another sense one finds in the philosophical literature on personal identity. The scenario in question seems to be imaginable only to a very limited extent. Fission—that is, if Brainy were to suddenly undergo real fission, amoeba style (as opposed to brain transplant) is physically impossible. We can only imagine what it would be like if Brainy suddenly divided into Lefty and Righty in front of our eyes to a very limited extent. Would each be half the size, for example? There is something to be learned about a concept like human identity from this sort of case, but it's not *really* imaginable. Unimaginability is a pervasive feature of thought experiments. If it is, how useful is it then to include it as a characteristic or a criticism for thought experiments?

Since impossibility of the scenario considered makes the thought experiment unimaginable, I think in this light, Gendler's unimaginability criterion might need revising. It would seem rash to dismiss a broad class of thought experiments altogether on such grounds. I am therefore, reluctant to include imaginability as a formal requirement of a thought experiment, even if only to a predefined degree.

Under traditional accounts of thought experiments such as Sorensen's and Gendler's thought experiments that consider impossible scenarios are deemed either unsuccessful or unimaginable. Thus one cannot learn very much from such thought experiments. As we will see in the next section the

main advantages of our new semantics for counterfactual, when we apply it to thought experiments are that it can account for usefulness of impossible thought experiments, and that this with its continuous treatment of possible and impossible cases we are able to explain why the debates about thought experiments such as the Chinese Room look the way they do.

Section 5. Application of the New Theory:

The crucial tool my new theory of counterfactuals provides that is relevant to understanding thought experiments is an account of the truth conditions for counterfactuals. Below, with the help of four crucial thought experiments (viz., the Ship of Theseus, Searle's Chinese Room, Fission, and Strawson's Auditory world) I will show how this new theory helps advance the debate on such thought experiments, by showing us the truth conditions for such thought experiments. The problems that the above thought experiments tackle are very hard philosophical problems. So simply knowing what those truth conditions are does not mean that we will be able to determine easily whether the claims are true or not. But it can help clarify what is at issue in these debates. It can help clarify why some matters are relevant to the question and some are not.

Application to Ship of Theseus

In our new semantics, we evaluate a thought experiment, by first translating the thought experiment into a series of counterfactuals. The different counterfactuals would have, in the present case, the following general form:

The antecedent would have the stipulations and requirements that are built into the relevant scenario. The consequents of each counterfactual would express the different anticipated outcomes.

Some possible counterfactuals in the Ship of Theseus (ST) could look like the following:

- If ST, then the newly constructed ship with all of its planks replaced is the Ship of Theseus.

- If ST, then the ship constructed with all the old planks is the Ship of Theseus.
- If ST,... then an identity preserving process results in identity creation.

This thought experiment is clearly a possible thought experiment. As we know. the new semantics systematically deals with both possible and impossible counterfactuals. So, how would this thought experiment be treated under this account? The question that is relevant here is: does the new analysis of counterfactuals change anything with regards to the Ship of Theseus or are all the same questions and answers in place as before the development of the new analysis? The answer is that since the new theory about counterfactuals with possible antecedents is just the same as traditional possible worlds accounts, nothing really changes with regards to the Ship of Theseus. By thinking of thought experiments in terms of counterfactuals we can understand why the debate ought to be about these issues. Because these are the issues that determine "closeness".

Clearly, at most one of these three counterfactuals can be correct. Moreover, since there are clearly possible worlds in which Theseus-style activities take place, the truth conditions for my theory and for standard Lewis-Stalnaker accounts will be the same for the three counterfactuals: it is a matter of whether at the nearest possible two-ship world Theseus's ship is the old plank ship, the new plank ship, or both. What comes out of the discussion above that might help, though, is that some of the same kinds of considerations that apply in determining which of two impossible cases is closer to actuality could be relevant to answering what to say is the case at the nearest possible two-ship world. The question at hand is "what should we say?". In determining what the nearest case in which Dave squares the circle is like, we must raise issues such as "what features of reality are salient and can be held constant?" Similar questions arise in the present case: is this a world where people regularly do what Theseus has done? Then perhaps ships will have identity conditions like (to steal Gendler's examples) world wide web sites and amoeba, and the question of a second christening won't arise.

Perhaps, then, the we want to say that the nearest world is one where only Theseus plays pranks of this sort. But in any case, it is on decisions of this sort, it seems to me, that the question of what it will be appropriate to say will turn.

Application to Chinese Room

This is, possibly, an example of a computationally impossible thought experiment, as we will see. The Chinese Room can be formalized as a counterfactual as follows:

- if you consider a Chinese Room, then there is nothing in the room that understands Chinese, despite how it appears judging only by outside behaviour.

By analogy, computers are similar to the Chinese Room. Thus the conclusion Searle derives is that computers are not intelligent kinds of things, and specifically they do not have intentionality, and that their purely syntactic manipulations are bereft of the semantics that are requisite for intentionality.

Daniel Dennett expresses concerns about the Chinese Room and calls it an “intuition pump”. Intuition pumps, are “not arguments, they're stories. Instead of having a conclusion, they pump an intuition. They get you to say “Aha! Oh, I get it!”” (Dennett, Chapter 10: Intuition Pumps). In "Fast Thinking" he expresses another concern regarding the speed at which the Chinese Room would operate. Although, the operator of the Chinese Room may eventually produce appropriate answers to Chinese questions, the speed at which it would be done makes it a slow system. But however complex they may be, slow thinkers are stupid, not intelligent. In Dennett's words: "speed … is ‘of the essence’ for intelligence. If you can't figure out the relevant portions of the changing environment fast enough to fend for yourself, you are not practically intelligent, however complex you are" (Dennett, 1987, p. 326). Thus some may hesitate to attribute intelligence and understanding to a slow system, such as the Chinese Room, due to concerns regarding speed. It may simply be that our intuitions regarding the

Chinese Room are unreliable, and thus the man in the room, in implementing the program, may understand Chinese despite intuitions to the contrary. Or it may be that the slowness marks a crucial difference between the simulation in the room and what a fast computer does. Thus, we may say that the man is not intelligent while the computer system as a whole is (Dennett)[15].

For thought experiments such as the Chinese Room, it is not clear whether there is a subtle impossibility in the scenario or not. For the person in the Chinese Room to manipulate Chinese symbols, as per the rule book, so as to appear intelligent, perhaps some laws of physics have to be violated. Because intelligent systems are incredibly fast and the Chinese Room would have to be equally fast in order to be a candidate for an intelligent system, or the person inside the room at least would have superhuman powers. Or maybe there are thousands of people and not just one person, inside the Chinese Room , but then the question of how they coordinate their efforts so effectively remains to be answered. So it is really not clear what we should/would say when we sit down to consider the details.

What Searle says is roughly that "in this Chinese Room scenario, we should certainly say X (i.e. computers don't have intentionality)". However, what comes out of the debate about the Chinese Room is that it is not clear at all what would really be required in such a scenario. If one tries to conceive such a scenario one will see that it's not at all clear that you would say about whether it is thinking. Dennett raises the crucial consideration: "well, what would need to be in place ... the guy inside could not just be flipping through a book at the rate normal people do, so what do we change about him? ... when we make all those changes, it's not clear at all what one would say."

Dennett's critique of the Chinese room fits what we suggest is required to turn a set of sentences

[15]For the discussion on Chinese Room above, I have relied on Cole, D.: The Chinese Room Argument, *The Stanford Encyclopedia of Philosophy (Fall 2008 Edition)*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2008/entries/chinese-room/>.

into a case. It needs to be shown to be conceivable by filling in a few more details. However, we need not fill it in entirely (because it need not even be possible). And what Dennett argues, in effect, is that it is not clear that the nearest case in which we confront a Chinese Room is one where it is right to say "not thinking". The crucial question is not whether there is such an impossibility, but what is the "nearest" situation in which there is a Chinese Room (whether it is impossible or not) and what we would say in such a case (about the intentionality of the room). What the semantics makes clear in this case is why Dennett's concerns are exactly the relevant ones, and why the debate has not centred on the question of possibility.

Application to Fission:

Fission is arguably physically impossible. Fission is '*extraordinary'* and '*counterfactual'* to use Gendler's words. We do not have to look at the world to see if such a thing happens. Without looking at the world we can tell that such a case is purely fictional. The criticism that Gendler levels at this thought experiment is *unsound argument.* According to Gendler, the scenario described here is "imaginable", but the argument establishing the correct evaluation of the scenario is unsound. So the evaluation of the scenario provided by Parfit is "fundamentally misguided".

According to Gendler, fission does not show what Parfit takes it to show, i.e., that identity does not matter. I agree with Gendler that fission does not show what Parfit takes it show. My reasons are very different from Gendler's, though. The fact that I will be the same person tomorrow is because me-today is connected by the right sort of psychological and causal connection to me-tomorrow, as Parfit would say drawing from fission. This conclusion is unwarranted according to Gendler. According to Gendler, fission is a process in which the same process that is identity preserving (in the single-transfer case) ends up being entity creating (in the double-transfer case), which raises the

following puzzle—how can the rationality of our attitude towards one's continuer be different depending on the outcome? In other words, how could we react differently towards the outcome , depending on whether it is a single-transfer case or double-transfer case? "Presumably one's attitude towards one's continuer would—rationally—be the same in both the single-transfer and the double-transfer case. And with this much, I said I agree" (Gendler, 2000, p. 146). Gendler allows for the first step of fission i.e., the single-transfer case to go through. It is the double-transfer case that Gendler calls into question..

What role does the new semantics play in the fission case? The new account helps make clearer what is at issue in the debate; and, in particular, this clearer picture prevents one from being led astray as was Gendler, who is led by her own theory to mis-diagnose what is wrong with Parfit's account. According to Gendler, fission is exceptional and it ought to be treated as an instance where norms drive the exceptions, rather than an instance where exceptions drive the norms (as Parfit does according to Gendler). However, what needs to be sorted out even before we decide whether the first step of fission (i.e., the single-transfer case) goes through is: what are the relevant issues? Gendler's account leads her to focus on the wrong sorts of considerations. That is whether fission is to be treated as a case of norm-driven exception or exception-driven norms. The new account leads us to focus on the right sorts of matters. According to Gendler, Parfit is misdirected because he considered fission as a case where exceptions drive the interpretation of the norms. Parfit tells us 'identity' is not what matters and this is what we learn from the double-transfer case. According to Gendler, what we ought to say about fission is that it is obviously an exceptional case and it is o be treated as a case in which norms drive the interpretation of exceptions. The important question, as we will see, is not whether fission is an exceptional case or not and whether we treat it as a case where norms drive the interpretation of exceptions. As we will see, why issues such as whether the lack of ability to survive a

stroke post-fission means Brainy and Lefty are two distinct people, are the relevant issues in this context. This will help us figure out why fission does not show what Parfit takes it to show and why Gendler's verdict is misguided. Although fission is impossible (at least physically), but, it is conceivable. The counterfactuals corresponding to the fission thought experiment might look like the following:

- If a single-transfer scenario occurred, in which the brain of one of the triplets…(P), then Brainy would be identical with Lefty (Q).
- If a double-transfer scenario occurred…(P2), then Brainy would not be identical to Righty and Lefty (Q2).

How does the semantics help us in this context? What determines whether P $\Box\!\!\rightarrow$ Q is true is whether there is a PQ world closer than any P ¬ Q world? *Closeness* involves scrutiny of what, at *a*, is involved with the sorts of changes required to make a P-world. But that means that factors like what determines identity in the case of brain trauma are relevant. And arguably, some properties that Brainy has that Lefty does not (after the operation) are relevant. For instance, Brainy could have a peculiar stroke that shuts down important functions on the left half of his brain without serious disability, Lefty not. Grounds such as these seem to be the relevant ones in considering whether or not there is a PQ world closer to actuality than is any P and ¬Q world. And the semantics explain why they are the relevant ones.

What considerations of that sort suggest is that the first counterfactual is indeterminate because there are no P cases in which Q or ¬Q is true. Notice, that Brainy is very special because he has a very unusual brain, in that it is replicated in each lobe etc. And after the accident, after the transplant takes place, for Brainy to be *identical* to Lefty, if identity includes such things as effects of

particular sorts of stroke, the whole of Brainy's brain needs to be transplanted into Lefty's. But because only half of Brainy's brain is transplanted, we cannot call him identical to Lefty. So even the first step of the fission argument (that Brainy is identical to Lefty) cannot go through, if this argument is cogent, and the semantics makes clear why. This is what Gendler's account misses. She does let the first step of the argument go through.

Application to Auditory World

Arguably, the auditory world is physically and conceptually impossible. We established in Chapter IV that it is unimaginable. It is, however, conceivable. If we level the criticism of *unimaginability*, (in Gendler's spirit) the auditory world enterprise collapses before getting off the ground. In the present setting, though, we merely translate the thought experiment into a suitable set of counterfactuals, and the unimaginability and impossibility of the scenario merely makes the antecedent impossible. So some of the counterfactuals may look like the following:

- If there is a completely auditory and non-spatial world inhabited by Hero, who is such…(P), then for non-solipsistic consciousness Hero would need an analogue of space (Q).
- If there is a completely auditory and non-spatial world with Hero, who is such…(P2), then the least we need suffice for Hero to make sense of objective particulars is M-sound (Q2).

And so on and so forth.

What Strawson argues is that there is a P and Q case that is closer to actuality than is any P and not-Q case. We see this in his argument when he says, for example, that there must be an analogue of space, for Hero to be able to distinguish between himself and things other than himself. Since Hero lives in an auditory world populated only by sound particulars, without an analogue of space i.e., M-sound, which form the background against which Hero compares different sound experiences, Hero

could not tell one sound apart from another.

What Strawson needs to show, to establish his conclusions, is that these counterfactuals are true. And, as we now know, this amounts to showing that in the nearest case in which the antecedent is true (there is a non-spatial sound world that includes an agent Hero ... and Hero can make sense of objective particulars), the conclusion (there is an analogue of space, e.g., M-sound) is true also. In order to make a persuasive case that he is right, all Strawson needs to show is that M-sound world is close enough to the actual world. However, what Strawson does, as we have seen, is that he tries to sketch a complete picture of such a world. When he realizes that it is not possible to completely and coherently describe such a world, he abandons the auditory world. What we learn from the new analysis is that Strawson argues in a way that he didn't need to. He thought, for the auditory world enterprise to be carried out, it had to be carried out in its complete and coherent form. All that needed to be done was that to show that such a world is the "closest" to the actual world.

Thus we can evaluate our original thought experiment, despite its being unimaginable. Philosophically this thought experiment is useful and informative, at least in so far as it helps us figure out objectivity. So, we cannot dismiss it, as we would have to if we adopt Gendler's account. It is an 'exceptional case' to adopt Gendler's terminology.

Using the new semantics one can say Strawson, in this thought experiment, tries to establish the nearest case where space is not a prerequisite for objectivity. In a non-spatial case that is closest to the the actual world, what would we need to retain in order to preserve objectivity? This is what Strawson tries to find out with the help of this thought experiment. Thus our new semantics nicely captures the main motivation of this thought experiment, that Gendler's and Sorensen's accounts fail to capture.

Conclusion

I hope that the discussion in this chapter makes clear that the theory of counterfactuals that has been the main topic in this thesis has potential for real philosophical payoff. I don't pretend to have used the new tools I have developed to solve the problems I discuss, nor even to have made significant progress on them. Instead, what I hope to have done is made plausible the claim that use of this new tool could lead to such progress. One example of the kinds of uses a clear understanding of counterfactuals with impossible antecedents will have in getting a clear understanding in these other areas is that in case of impossible case scenario thought experiments it allows for a nice representation of what is at issue in the philosophical debates. The new semantics provide some guidance to philosophers about what needs to be shown to make a compelling case. And in the fission case, it helps us figure out that though it may initially seem fission is fully and coherently describable, it is not the case. However, this is not why fission fails to show what it is taken to show. For there to be a case where either the consequent or the negation of the consequent is true, we need to talk about identity in the ordinary sense of the term. For identity-continuation conditions to be fulfilled, in fission-like cases, we need an ordinary case of brain transplantation, where all of Brainy's brain is transplanted into Lefty's body. To make a compelling case for identity, ordinary conditions for identity ought to be fulfilled.

In the auditory world and the Chinese Room examples, looking at the thought experiments through the lens of our new semantics helps us figure out why consistency is not what is important for a thought experiment to be informative. In the auditory world example, with the help of the new semantics we are able to see that what Strawson in fact tries to do is to figure out what is the nearest case in which we are able to preserve objectivity without space. This captures the main motivation of the thought experiment, which the other accounts miss. In the Chinese room example, our new semantics helps us see why the debate surrounding this thought experiment looks the way it does. It is

not the possibility or impossibility of the scenario that is important. What is important what the nearest Chinese Room is like, whether possible or not—it is from that case that we learn from the thought experiment. In the traditional accounts the lessons to be learnt in case of a impossible thought experiments are non-existent.

Bibliography:

Adams, E. W.: 1970, Subjunctive and Indicative Conditionals, *Foundations of Language* **6**, 89-94.

Bennett, J.: 1988, Farewell to the Phlogiston Theory of Conditionals, *Mind* **97**, 509-527.

Bennett, J.: 2001, Conditionals and Explanations, *in* A. Byrne, R. Stalnaker, and R. Wedgwood (eds.), *Fact and Value: Essays on Ethics and Metaphysics for Judith Jarvis Thomson.* MIT Press, Cambridge, Massachusetts. pp. 1-28.

Bennett, J.: 2003. *A Philosophical Guide to Conditionals.* Clarendon Press, Oxford.

Boden, M.: 1988, *Computer Models of the Mind*, Cambridge University Press, Cambridge, pp. 238-251 were excerpted and published as 'Escaping from the Chinese Room', in *The Philosophy of Artificial Intelligence*, M. A. Boden (ed.), Oxford University Press, New York, 1990.

Brown, J.R.: 1991, *Laboratory of the Mind: Thought Experiments in the Natural Sciences.* Routledge, London.

Chalmers, D. J.: 2002, Does Conceivability Entail Possibility?, *in* T. S. Gendler and J. Hawthorne (eds.), *Conceivability and Possibility*. Clarendon Press, Oxford. pp. 146-200.

Chisholm, R.M.: 1946/1949, The Contrary-to-Fact Conditional, *in* H. Fiegl and W. S. (eds.), *Readings in Philosophical Analysis*. Appleton-Century-Crofts, Inc., New York. pp. 482-497.

Cole, D.: 2008, The Chinese Room Argument, *The Stanford Encyclopedia of Philosophy (Fall 2008 Edition)*, E. N. Zalta (ed.)

URL: *http://plato.stanford.edu/archives/fall2008/entries/chinese-room/*

Descartes, R.: [1619-64] 1990, *The Philosophical Writings of Descartes*, vol. I, trans. J. Cottingham, R. Stoothof, and D. Murdoch. Cambridge University Press, Cambridge.

Descartes, R.: [1619-64] 1990, *The Philosophical Writings of Descartes*, vol. I, trans. J. Cottingham, R.

Stoothof, D. Murdoch, and A. Kenny. Cambridge University Press, Cambridge.

Descartes, R.: [1641-2, 1701] 1989, *The Philosophical Writings of Descartes*, vol. II, trans. J. Cottingham, R. Stoothof, and D. Murdoch. Cambridge University Press, Cambridge.

Descartes, R.: [1619-50] 1991, *The Philosophical Writings of Descartes*, vol. III, trans. J. Cottingham, R. Stoothof, and D. Murdoch. Cambridge University Press, Cambridge.

Dennett, D.: 1987, Fast Thinking, *in The Intentional Stance*, MIT Press, Cambridge.

Dennett, D.: 1995, Intuition Pumps, *in* J. Brockman's *The Third Culture: Beyond the Scientific Revolution.*

URL: *http://www.edge.org/documents/ThirdCulture/c-Copyright.html*

Dummett, M.: 1978, *Truth and Other Enigmas,* Harvard University Press, Cambridge.

Edgington, D.: 2004, Two Kinds of Possibility, *Aristotelian Society* Suppl. **78 (1)**, 1–22.

Edgington, D. 1995.: On conditionals. *Mind*, New Series,**104 (414)**, 235-329.

Edgington, D.: 2003, Counterfactuals and the Benefit of Hindsight, *in* P. Dowe and P. Noordhof (eds.) *Causation and Counterfactuals*. Routledge, London. pp. 12-27.

Edgington, D.: 2006, Conditionals, *The Stanford Encyclopedia of Philosophy* (2006 Spring Edition), E. N. Zalta ed.

URL: *http://plato.stanford.edu/archives/spr2006/entries/conditionals/*

Evans, G.:1980, *Things Without the Mind—A Commentary upon Chapter Two of Strawson's Individuals, in* Z. V. Straaten (ed.) *Philosophical Subjects*. Clarendon Press, Oxford. pp. 76-116.

Fogelin, R. J.: 1998, David Lewis on Indicative and Counterfactual Conditionals, *Analysis* ,**58**, (**4)**, 286-289.

Gendler, T.: 2000, *Thought Experiment: On The Powers and Limits of Imaginary Cases,* Garland Publishing, Inc., New York and London.

Gendler, T. and Hawthorne, J. (eds.) (2002). *Conceivability and Possibility,* Clarendon Press, Oxford; Oxford University Press, New York.

Glover, J.: 1975, It Makes No Difference Whether or Not I Do It, *Proceedings of the Aristotelian Society*, Suppl. **49,** 174-175.

Goodman, N.: 1965, *Fact, Fiction, and Forecast* (2nd edition), Bobbs-Merill, Indianopolis.

Hume, D.: 1968, *Treatise of Human Nature,* Clarendon Press, Oxford.

Kripke, S.: 1963, Semantical Considerations on Modal Logic, *Acta Philosophica Fennica*, **16**, 83-94.

Kripke, S.: 1980, *Naming and Necessity,* Harvard University Press, Cambridge, Massachusetts.

Kvart, I.: 1986, *A Theory of Counterfactuals,* Hackett Publishing Company, Indianapolis.

Lewis, D.: 1973, *Counterfactuals,* Harvard University Press, Cambridge, Massachusetts.

Lewis, D.: 1979, Counterfactual Dependence and Time's Arrow, *Noûs*, **13 (4)**, 455-476.

Lewis, D.: 1986, *On the plurality of worlds*, Basil Blackwell, Oxford.

Mares, E.: 2006, Relevance Logic, *The Stanford Encyclopedia of Philosophy (Spring 2006 Edition)*, E. N. Zalta (ed.).

URL: *http://plato.stanford.edu/archives/spr2006/entries/logic-relevance/*

Mattey, G.J.: 1998, Lecture Notes: Strict Conditional (April 6, 1998).

URL: *http://philosophy.ucdavis.edu/mattey/phi134/**strict**.htm/*

Mendelsohn, R.: 2004, Review of Jonathan Bennett's *A Philosophical Guide to Conditionals , in* Notre Dame Philosophical Reviews 2004.02.11

URL: *http://ndpr.nd.edu/review.cfm?id=1388#1b*

Nolan, D.: 1997, Impossible Worlds: A Modest Approach, *Notre Dame Journal of Formal Logic*, **38, (4)**, 535-572.

Norton, J.:1996, Are Thought Experiments Just What You Always Thought?, *Canadian Journal of*

Philosophy, **26, (3)**, 333-366.

Norton, J.: 1991, Thought Experiments in Einstein's Work, *in* T. Horowitz and G. Massey (eds.), *Thought Experiments in Science and Philosophy,* Littlefield Publishers, Inc., Rowman pp. 129-144.

Phillips, I.: Morgenbesser Cases and Closet Determinism

URL: *http://users.ox.ac.uk/~magd1129/Morgenbesser.pdf*

Pruss, A. R.: 2001, *Possible Worlds: What They Are Good For and What They Are*.

URL: *http://www9.georgetown.edu/faculty/ap85/papers/PhilThesis.html*

Restall, G.: 1999, Negation in Relevant Logics, *in* D.M.Gabbay and H. Wansing (eds.) *What is Negation?*. Kluwer Academic Publishers, Dordrecht/ Boston, Massachusetts. pp. 129-144.

Rosen, G.: 1995, Modal Fictionalism Fixed, *Analysis*, **55, (2)**, 67-73.

Sanford, D. H.: 1989, *If P then Q: Conditionals and the Foundations of Reasoning*. Routledge, London and New York.

Searle, J.: 1980, Minds, Brains and Programs, *Behavioral and Brain Sciences,* **3,** 417-457 Sorensen, R. A.: 1992, *Thought Experiments*. Oxford University Press, New York.

Slote, M.: 1978, Time in Counterfactuals, *Philosophical Review,* **87, (1)**, 3-27.

Stalnaker, R. C.: 1968, A Theory of Conditionals, *in* W.L Harper, R. Stalnaker, and G. Pearce (eds.), *Ifs* (1981). Basil Blackwell Publisher. pp. 41-55.

Stalnaker, R. C.: 1978, A Defense of Conditional Excluded Middle, *in* W.L Harper, R. Stalnaker, and G. Pearce (eds.), *Ifs* (1981). Basil Blackwell Publisher. pp. 87-104.

Stalnaker, R. C.: 1975, Indicative Conditionals, *Philosophia* **5**, 269-286. Reprinted in F. Jackson, (ed.), (1991) *Conditionals.* Oxford University Press, Oxford. pp. 136-154.

Stalnaker, R. C.: 1984, *Inquiry,* A Bradford Book. MIT Press, Cambridge, Massachusetts/ London, England.

Stalnaker, R. C.: 2003, *Ways A World Might Be: Metaphysical and Anti-Metaphysical Essays,* Clarendon Press, Oxford.

Strawson, P.F.: 1959, *Individuals: An Essay in Descriptive Metaphysics,* Routledge, London and New York.

Vaidya, A.: 2007, The Epistemology of Modality, *The Stanford Encyclopedia of Philosophy (Winter 2007 Edition)*, E. N. Zalta (ed.).

URL: *http://plato.stanford.edu/archives/win2007/entries/modality-epistemology/*

Wilkes, K. V.: 1988, *Real People*: *Personal Identity* W*ithout Thought Experiments*. Clarendon Press, Oxford.

Williamson,T.: 2007, Philosophical Knowledge and Knowledge of Counterfactuals.To appear in C. Beyer and A. Burri (eds.), *Philosophical Knowledge*: *Its Possibility and Scope*, Rodopi, Amsterdam. pp. 89-123.

URL: *http://www.philosophy.ox.ac.uk/__data/assets/pdf_file/0008/1304/counterfactuals.pdf*

Yablo, S.: 1993, Is Conceivability a Guide to Possibility?, *Philosophy and Phenomenological Research*, **53, (1)**, 1-42.