Does Immanuel Kant's Categorical Imperative Commit Him to the View that Lying is Always Morally Wrong?

by

Martin Ludwig Perold

Submitted in partial fulfilment of the requirements for the degree of

Master of Arts

at the

University of the Witwatersrand

2010

under the supervision of

Prof L Allais

ABSTRACT

In Immanuel Kant's essay "On a Supposed Right to Lie because of Philanthropic Concerns" (1797) he famously argues that it is never permissible to tell a lie, even when lying could save someone's life. This view has met with a great deal of criticism from philosophers, who argue that his ethical theory must be flawed if it leads to such an undesirable conclusion.

In this report, I explore this claim, arguing that this conclusion does not, after all, follow from Kant's ethical theory. I focus in particular on the three formulations of the categorical imperative – the Formula of the Universal Law, the Formula of Humanity and the Formula of the Kingdom of Ends – and argue that none of these versions of Kant's key ethical principle requires him to make the rigorous claim that we may *never* lie under any circumstances. Although lying turns out to be morally wrong in the majority of cases, based on a proper application of Kant's theory, there are likely to be some situations in which lying is permissible or even obligatory, as I hope to show in this research.

ACKNOWLEDGEMENTS

I would like to thank the following people for their support and assistance throughout this study:

- My supervisor, Prof Lucy Allais of the University of the Witwatersrand, for her valuable advice, assistance and support during this dissertation. Her passion for philosophy is infectious and inspired me to work hard in this study.
- My parents, Roeland and Hildegard Perold, for their undying love and assistance through my whole life; for supporting me unconditionally in all my academic endeavours and for fostering a love of learning in me.
- The members of the Philosophy Department at the University of the Witwatersrand, who provided useful and thought-provoking comments during the reading of my proposal.
- My many close friends, for their unfailing support, patience, pep talks, motivation and assistance throughout the dissertation.

DECLARATION

I, Martin Ludwig Perold, student number 9809721Y, am a student registered for the course of Master of Arts (by Coursework and Research Report) in the year 2010.

I hereby declare that this research report is my own, unaided work except where I have explicitly indicated otherwise. This report is being submitted for the Degree of Master of Arts in the Department of Philosophy, School of Social Sciences in the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination in any other University.

I am aware that plagiarism (the use of someone else's work without their permission and/or without adequately acknowledging the original source) is wrong and is a violation of both the General Rules for Student Conduct and the Plagiarism Policy of the University of the Witwatersrand.

I have followed the required conventions in referencing the words and ideas of others in the attached work.

I understand that the University of the Witwatersrand Plagiarism Policy may be applied if there is a belief that the attached work is not my own unaided work or that I have failed to follow the required conventions in referencing the words and ideas of others, and I understand that application of the Plagiarism Policy may lead to the University taking disciplinary action against me.

Martin Perold 24 November 2010

MARTIN LUDWIG PEROLD DATE

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

AP: Kant's *Anthropology from a Pragmatic Point of View* (1798)

CC: Contradiction in Conception

CW: Contradiction in the Will

EMC: Hanna's Excluded Middle Constraint

FH: The Formula of Humanity (the second formulation of the categorical imperative)

FKE: The Formula of the Kingdom of Ends (the third formulation of the categorical imperative)

FUL: The Formula of the Universal Law (the first formulation of the categorical imperative)

G: Kant's *Grounding for the Metaphysics of Morals* (1785).

LE: Kant's *Lectures on Ethics* (18th Century)

LEP: Hanna's Lesser Evil Principle

NGV: Hanna's No-Global-Violation Constraint

MM: Kant's *The Metaphysics of Morals.* (1797)

RL: Kant's essay "On a Supposed Right to Lie because of Philanthropic Concerns" (1787)

DOES IMMANUEL KANT'S CATEGORICAL IMPERATIVE COMMIT HIM TO THE VIEW THAT LYING IS ALWAYS MORALLY WRONG?

In his infamous essay "On a Supposed Right to Lie because of Philanthropic Concerns" (1797), Immanuel Kant adopts the view that it is never permissible to lie under any circumstances. This is because he claims that telling the truth "in all declarations" is "a sacred and unconditionally commanding law of reason that admits of no expediency whatsoever" (RL, 427, 65)[1]. It seems as though this view follows from Kant's ethical theory, developed in his *Groundwork for the Metaphysics of Morals* (1785) and the *Metaphysics of Morals* (1797), as well as in his lectures on ethics during the eighteenth century. In these works, Kant maintains that telling the truth is a perfect duty to oneself that may not be violated.

This position is certainly controversial, as it is easy to think of cases where lying appears to be morally permissible. For instance, people often tell so-called "white lies" to avoid hurting the feelings of others, such as when someone assures her partner that she loves the expensive gift that he has just bought her, despite secretly despising it. Similarly, an adult might lie to a child about the existence of the Tooth Fairy in order to lessen its distress upon losing a tooth; or a person might lure a friend to her house under false pretences in order to throw him a surprise birthday party. As I will explain in a later section, these cases may all be regarded as lies, and they do not seem to be morally wrong.

Lying is also common in everyday social exchanges, such as the standard response of "Fine, thank you" when someone inquires about one's health even when one is ill; or, to use Kant's own example, the convention of writing "Your obedient servant" at the end of a letter to a person whom one has no intention of serving or obeying (MM, 6: 431, 554). These lies do not seem to be morally reprehensible. As Joseph Margolis

[1] Throughout this dissertation, I will refer to Kant's works through abbreviations, with two page numbers shown. The first page number refers to Kant's original page number, while the second one refers to the page of the version that I am using.

(1963: 414) points out, few people "would subscribe to the principle, 'Lying is wrong'" without exception.

We can also provide examples of lies that seem to be not only morally permissible, but *obligatory*, such as telling a lie to save someone's life. Kant's example (RL) involves lying to a murderer who wants to kill someone in one's house by telling him that his intended victim is not at home, thereby rescuing her. If this lie would prevent the victim's death, it certainly seems as though we are required to tell it. Kant's view, however, is that the lie is morally wrong regardless of the good it might do. Because Kant thinks that "a lie *always* harms another" and "does harm to humanity in general, inasmuch as it vitiates the very source of right", we may not allow even "the slightest exception" to the duty to tell the truth (RL, 426-427, 64-65, my italics).

This view certainly seems objectionably strict and rigorous, since it does not appear reasonable to claim that a life-saving lie may never be told. It is therefore not surprising that many philosophers disagree with Kant on this issue – as James Mahon (2009: 201) points out, "Kant's writings on lies have elicited an unprecedented amount of abuse". Wolfgang Schwarz (1970: 62) emphasises how controversial Kant's position is, noting that Kant "seems to have been carried away by formal considerations and lost all touch with reality" in drawing a conclusion that "has embarrassed even some of his friends and served as a cause for rejoicing among his foes."

The question that I propose to explore in this dissertation is whether the seemingly unreasonable conclusion that lies are always morally wrong indeed follows from Kant's ethical theory. As Allen Wood (2008: 1) explains, if Kant's ethics lead to the extreme result that lying is *never* morally permissible, then we may be able to show that "there must be something fundamentally wrong with Kantian ethics". The position that I will defend here is that the undesirable conclusion does not, after all, follow from Kant's ethical principles and that his famous categorical imperative does not require us to tell the truth at all times. I hope to show that, although lying is *generally* morally wrong, there are some cases in which it is permissible (such as

telling a lie to save a life), and that these cases do not undermine Kant's ethical thought.

Of course, I must begin by being explicit about how I will understand the term "lying", and this will be discussed in the first section of the research. Next, I need to explain what Kant's ethical theory entails and how he reaches his conclusions about lying. The focus of this dissertation will be on the categorical imperative, which is Kant's rule for determining the morality of an action under consideration. According to most authors, there are at least three different formulations of this rule (Korsgaard, 1986; Wood, 1999; Guyer, 2007). In the second section of the dissertation I will explain the three versions of the rule in detail and show how they are supposed to be used to reach conclusions about whether actions are permissible or not. Here, I will explain why Kant thinks that lying is impermissible and abhorrent. This section will also discuss Kant's distinction between perfect and imperfect duties (LE, MM), and show why he thinks that telling the truth falls under the former category.

In the next part of the dissertation, I will apply the various formulations of the categorical imperative to the question of lying and specifically to the case of lying to the murderer at the door. There is a section devoted to each version of the theory, and I aim to show that none of them leads to the conclusion that lying is *always* wrong. The analysis will show that we may justified in lying to the murderer at the door, for reasons that I will make explicit, and that Kant's moral theory is not as rigorous as it appears at first glance.

The final section of the research focuses on an alternative interpretation of the categorical imperative, as suggested by Philip Stratton-Lake (2000), Jens Timmermann (2007) and Robert Hanna (2009). This view regards the categorical imperative not as a *decision-making rule* that guides our actions, but rather as an *explanation* for why a given action is permissible or not. I will critically discuss this position, showing that it is a plausible way of interpreting Kant's moral thought and that it does not lead to the rigorous conclusion that lying is always forbidden.

Finally, I will conclude that Kant's moral theory is not as strict and extreme as it appears and that he is mistaken to conclude that lying is always wrong. None of the formulations of the categorical imperative entail the conclusion that lying is *never* impermissible, although it is probably impermissible in the majority of cases. Furthermore, the alternative interpretation of the principle provides us with another, plausible way of making sense of Kant's ethics that does not require us to tell the truth at all times, but explains why we find ourselves in a moral dilemma when we are confronted with situations like those involving the murderer at the door. Consequently, Kant's moral theory is not as objectionable as it seems, since it permits lying under certain circumstances. I will now consider the concept of lying and what counts as instances of it.

I. The Concept of Lying

Before we are able to apply Kant's ethical theory to the question of lying, it is necessary to clarify exactly what he takes the term to mean, and what kinds of statements should be classified as lies. Kant explains that lying "comprises every intentional untruth, or every intentionally false statement of my disposition" (LE, 27: 605, 351), so that "in ethics... no intentional untruth in the expression of one's thoughts can refuse [the] harsh name" (MM, 6:429, 552) of a lie. These definitions highlight the two key features of a lie: first, it must be an *untruthful* statement; and second, its untruth must be *intentional*. I will discuss each of these features in turn.

To understand the first feature, we must distinguish between an untruthful statement and a false one. A false statement is simply a proposition that is not true in the world we inhabit. For example, if I make the assertion "Paris is the capital of Germany", I make a false statement, since the proposition I affirm does not hold true in our universe. An *untruthful* statement, on the other hand, is an assertion that is made in the *belief* that it is false, even if it is really true. Suppose that I firmly believe that Paris is the capital of Germany – if I then make the claim "Paris is the capital of France", the statement that I make might be *true* (since it corresponds to a fact about the world), but it is nevertheless *untruthful* (because I do not believe it to be true and am miscommunicating my beliefs). An untruthful statement is indeed "a false statement of one's disposition" since one is asserting something that one does not

believe to be true. On my reading of Kant, lying requires the statement one makes to be untruthful, but not necessarily false (although it may be false).

This means that a true statement such as "Paris is the capital of France" may be considered to be a lie if the speaker asserts it in the misguided belief that it is false. Frederick Siegler (1966) points out that lying requires the liar to believe the proposition he communicates to be false, but not that it actually *be* false. I think that this is a plausible view to hold. It seems reasonable to say that I have lied to you if I tell you something that I firmly believe to be untrue, even though it may later turn out to be true after all – as Thomas Carson (2006: 284) notes, "a necessary condition of lying" is "that the liar cannot believe the statement she makes is true".

It may be argued, however, that lying requires the statement to be false in addition to its being untruthful. This is the position that Carson (2006: 284) holds, claiming that "In order to tell a lie, one must make a false statement". This means that a true proposition can never be a lie, since, as Carson (2006: 284) explains, one may refute the claim that one has lied by "[s]howing that [the] statement [one has made] is true". In this dissertation, however, I will adopt the view that lies need not be false statements. This is because, as I will show later, what Kant considers to be abhorrent about lying is that it involves deception and the frustration of other people's ends; and, as I will argue, it is possible to deceive others without making any false statements.

The second feature of Kant's definition of lying is that the untruth must be intentional. Thus, if I mistakenly believe that Paris is the capital of Germany as the result of a poorly-drawn map, and if I then communicate this statement to you, I am not lying. I say something that is false, but, since I do not believe it to be false, I am not *intentionally* deceiving you. It is not my aim to cause you to acquire a belief that I know to be false – I am merely mistaken in my belief. As Siegler (1966: 128) notes, a "liar must... intend to deceive" someone else for her statement to properly be regarded as a lie, and a lie can occur "only if there is an express declaration of my willingness to inform the other of my thought" (LE, 27: 447, 203).

Of course, it might be argued that it is the truth or falsity of a proposition that determines whether or not it is a lie, and not the intention behind it, as Carson (2006) does. However, it seems as though we would not blame or censure someone who makes a false statement (even under oath) if she sincerely believes that she is being honest. What is at stake in this dissertation is whether or not there are any lies that are morally permissible, and it seems as though we should discount false statements that are told unintentionally, because it may be argued that any moral wrong that is committed in these cases is purely accidental. Here I am concerned only with deliberate untruths, where the liar is fully aware that she is telling what she believes to be an untruth, and where she intends to cause someone else to believe the statement she makes. As Kant says, "if it be that the other is *ever* meant to *believe* it, then... it is a lie" (LE, 27:62, 28). In this paper, therefore, I will consider lies to be intentionally untruthful statements, discounting accidental ones.

Now that I have explained how I will understand the terms "lying" and "lies" in this dissertation, I will consider some examples of "intentional untruths". Clearly, telling the murderer at the door that my friend is out when I know for a fact that she is in meets these criteria – I deliberately tell him a proposition that I do not believe to be true, with the intention of causing him to believe it. This is an obvious case of lying, and would be considered a lie under most authors' definitions of the term (such as Siegler, 1966 and Carson, 2006). However, there are other cases that I consider to be instances of lying that are not obviously lies.

One such example is when a parent tells her child that there is a Tooth Fairy who will give him money for the tooth he has lost if he places it next to his bed at night. The mother's intentions may be very noble, in that the child is upset about the gap in his mouth and needs comfort and solace. However, on Kant's definition of lying, she lies to her child. First, she does not herself believe that the Tooth Fairy exists – she probably intends to continue the charade by placing the money next to the child's bed herself during the night. Thus she is untruthful in her assertion. Moreover, the assertion is blatantly false, as it does not correspond to any real fact about the world (and would therefore be considered a lie on Carson's definition of the term too).

Second, the untruth is certainly intentional. The mother wishes the child to acquire the belief that the Tooth Fairy exists and intends to deceive him, at least for the time being. He will certainly cease to hold this belief when he is older but, for the moment, his mother's aim is for him to act on what she tells him; perhaps he will stop crying and put his tooth in a box, looking forward to his nocturnal visitor. Thus this statement meets the requirements of a lie, as we have defined it here.

It may be argued that the mother's story is not a lie because it is mere fantasy or pretence, just as a science fiction novel is, and that it is not to be taken seriously. However, the key difference between the mother's utterances and a novel is that fiction does not involve any intention of getting the reader to believe the story, while the tale about the Tooth Fairy does. Consequently, only the latter statement is a lie, since fiction does not meet the requirement of intending to deceive someone. A book marked "fiction" does not purport to tell the reader anything that is true – the reader is well aware that the sentences she reads are not facts, but stem from the imagination of the author and are not to be believed. In the case of the Tooth Fairy, on the other hand, there is a clear intention from the mother to cause her child to acquire the belief in the fairy, and this intent is absent from fictional novels.

Even though the mother seems to intend the child to acquire a false belief, we might argue that she is not doing him any wrong. This is because of the unique status that is awarded to children on Kant's views. For Kant, a child is "a passive citizen", that is, someone like "a minor... whose preservation in existence (his being fed and protected) depends not on his management of his own business but on arrangements made by another" (MM, 6:314, 458).

In other words, children are not fully morally developed, especially if they are very young, and they are incapable of looking after themselves. As a result, they lack "the capacity to set ends according to reason" (Wood and O'Neill, 1998: 198) and we are therefore entitled to treat them differently than adults. We must regard them as having "insufficient autonomy" (Korsgaard, 1996: 351) to make rational choices, and we are therefore entitled to "use manipulative tactics" in order to make them do what is in their own best interests, since they lack the ability to determine this themselves.

Thus, when we tell a child an untruth to influence his behaviour and to get him to make a good choice, we may defend this action on the grounds that it is permissible to lie to "persons who should be regarded as incompletely developed" (Korsgaard, 1996: 351).

Tamar Schapiro (1999: 723) regards childhood as "a condition that prevents human beings from achieving autonomy 'all at once'" and emphasises the point that children have a "special status" in the "ethical commonwealth" (Schapiro, 1999: 721). She regards childhood as a temporary "predicament" (Schapiro, 1999: 732) in which the child is prevented from making autonomous choices, although its ability to do so is developing. On Schapiro's view, children are able to become agents through "play" – that is, they inhabit fantasy worlds through playing with each other, and with adults, where they can "more or less 'deliberately' try on selves and worlds to be in" (Schapiro, 1999: 732). Even though the play they engage in is not real, it enables them to "adopt[-] one or other persona" and "act the part of full agents" (Schapiro, 1999: 732), thereby developing their capacity to make autonomous choices.

We may plausibly see the untruth told to a child about the Tooth Fairy as a form of play. When the parent invites the child into a fantasy world, she brings him into a realm where his actions count – he is able to "act the part... of one who can act" (Schapiro, 1999: 733). She creates a make-believe world in which his decisions (such as leaving his tooth in a box next to his bed) have consequences (such as receiving the money the next morning), even though his actions have "an essentially provisional, experimental nature which adult action lacks" (Schapiro, 1999: 733). Thus, even though the mother tells the child a deliberate falsehood, it seems as though her action may properly regarded as play rather than lying. However, it may also be argued that it is a lie nonetheless, since it meets Kant's definition of lying; but, if it is a lie, we may excuse it because of the child's special status as an underdeveloped agent.

Even if untruths told to children in a fantasy world are not lies, it seems as though apparently innocuous falsehoods told to adults with the intention of deceiving them should certainly be considered as lies. Playing a practical joke on someone by telling her something that is false can be regarded as a lie, because the joke can succeed only

if the victim believes the untruth. For instance, I might trick my friend into attending a formal party dressed as a Disney character by telling him that it is a fancy dress occasion. Since I know it is a formal party, I speak untruthfully, and since my aim is for him to *act* on the false information I give him, the untruth is clearly intentional. Thus "[j]oking lies" (LE, 27:62, 28) told for the sake of amusement should also be regarded as lies, as should "a white lie" and the lies that politicians have to tell "to achieve their aims", since all of these involve some form of deception.

Kant also refers to "the tellings of tall stories, or braggings in company" (LE, 27:700, 427) as instances of lying. Embellishing a tale with exaggerations to make the story more interesting, or misreporting the size of the fish that one caught entail deception, even if there is some element of truth in the story. For example, if I tell my friends that I caught an enormous carp weighing forty pounds, when it weighed only thirty pounds in reality, much of what I say is true – I did indeed catch a large fish. Despite this element of truth, though, I still deceive them with an intentional untruth, since I deliberately overstate the weight of the fish. As I shall argue later, Kant seems to regard exaggerations of facts as harmless lies, but they count as lies nonetheless.

One further category of lies merits discussion here; that is, lies told "out of politeness" (LE, 27:701, 427) such as Kant's example of following social conventions by writing "Your obedient servant" (MM, 6: 431, 554) at the end of a letter, despite not being the recipient's servant. It may be argued that this sentence is not a lie, since there is no apparent deception involved in this complimentary close – the recipient of the letter will certainly not take us to mean that we intend to serve and obey her. This might not count as a lie, since nobody is expected to believe it. Perhaps Kant's example is ill-chosen, but we may consider another case in point that shows how social conventions may plausibly be regarded as lies.

Following Carson (2006), we may suppose that someone decides to invite an unpleasant old uncle to her wedding, even though she desperately hopes that he will decline the invitation, since she despises him. Perhaps she extends the invitation to him only because her grandmother wishes her to, or because she feels pangs of guilt. If she then sends him an invitation worded "We sincerely hope that you will come to

our wedding", it may be argued that she is lying to him, since she most certainly does *not* hope that he will attend. Once again, the statement is an untruth, and it is also designed to deceive him into believing that he is welcome at the gathering, thereby meeting Kant's two criteria for being a lie. This example is somewhat different to Kant's own example, but I think that it might plausibly be regarded as a lie out of courtesy, even if Kant's "obedient servant" example may not be.

This section has shown that there are several cases that may reasonably be said to be lies, since they are deliberate untruths, but which, intuitively, do not seem wrong. I will argue that Kant's ethical theory does not necessarily require us to say that all these instances of lying are morally impermissible. Of course, one might object that there is something wrong with this definition of lying, and that it is too broad, but it seems as though the action of not telling the murderer at the door the truth is a clear case of lying, under any definition. Thus, even though I will consider some of the examples that are not obviously lies in the following analysis, I will focus specifically on the case of the murderer at the door and how Kant's moral principles can be applied to it. First, however, the categorical imperative has to be made explicit, and this will be done in the next section.

II. The Three Formulations of the Categorical Imperative

In the second section of the *Grounding for the Metaphysics of Morals*, Kant develops his ethical principle known as the categorical imperative. The categorical imperative is described by Allen Wood (2009: 1) as "the fundamental principle of ethical duties" and can be seen as a moral decision-making rule – it "says what action possible by me [or, indeed, anyone] would be good" (G, 414, 25). Kant and his commentators, such as Wood (1999, 2009), Christine Korsgaard (1986) and Paul Guyer (2007) distinguish at least three different formulations of this principle, namely, the Formula of the Universal Law (hereafter abbreviated as the FUL), the Formula of Humanity (FH) and the Formula of the Kingdom of Ends (FKE). In this section I will adopt the same division and explain each of the three versions in turn.

The Formula of the Universal Law (FUL)

To understand the first formulation of the categorical imperative correctly, we must first consider Kant's conception of a *maxim*. Kant defines a maxim as "a subjective principle for acting" (G, 421, 30, footnote 9) – in other words, it is an underlying, general rule to which we appeal when we perform actions. Furthermore, Kant's theory requires that it be subjective; that is, "it is the principle of a particular subject or agent" (Guyer, 2007: 83). This means that a maxim does not hold *objectively*, as a law of nature would, but is determined by reason "in accordance with the conditions of the subject" that is acting on it (G, 421, 30, footnote 9).

The concept of a maxim may be explained more clearly with the use of an example, as Wood (1999) does. He asks us to consider a situation where we have borrowed a book from him and have promised to return it by a specified time. When we perform the actions necessary to bring the book back into his possession, we are following the maxim of keeping our promises and this, according to Wood (1999: 41) ensures that these actions are "to be esteemed".

Importantly, there seems to be no inherent moral value in the action of getting into my car and driving to Wood's house to return the book. Instead, what makes the action meritorious is the fact that I perform it in accordance with the noble maxim of always keeping my promises. If I were driving to Wood's house with the aim of assassinating him, I would be acting on a different maxim (perhaps of killing any philosophers I meet). This maxim would render my action morally wrong, even though driving to his house appears morally neutral. On this example, we can see why the maxim is subjective: the morality of the action can depend on my intentions in performing it. In this way, the maxim tells us where "the moral worth of the action is located" (Wood, 1999: 40).

Now that the concept of a maxim has been made explicit, we may examine the first formulation of the categorical imperative, or the FUL. Kant states this principle as follows: "Act only on that maxim whereby you can at the same time will that it should become a universal law" (G, 421, 30). He goes on to add that this "universal

law" should be seen as "a universal law of nature" (G, 421, 30)[2]. Simply put, this means that we may determine whether or not a given action is morally permissible by considering a world in which *everyone* habitually acts on the maxim on which that action depends – that is, a world in which the maxim is a universal law of nature. If we can consistently will the maxim to become a universal law of nature, then the action is permissible; if not, then it is morally wrong. Kant intends this criterion of universalisability to be a rule from which "all imperatives of duty can be derived" (G, 421, 30).

The universalisability criterion is best explained by using an example. Kant considers a "man in need" who borrows money and makes a false promise to repay it when "he knows well that he won't be able to" do so (G, 422, 31). This is an obvious instance of lying, as the promise is intended to deceive the lender. Now, to determine whether the false promise is morally permissible, we must identify the maxim on which it is based. Kant expresses it as follows: "when I believe myself to be in need of money, I will borrow money and promise to pay it back, although I know that I can never do so" (G, 422, 31).

We may now apply the FUL's universalisability test to this maxim, by considering a world in which everyone habitually acts on it; that is, a world where everyone borrows money and makes false promises to repay it whenever they are in need. Is the maxim universalisable? Kant explains that it is not, because "the end to be attained" from making a false promise in a world where everyone habitually does so would be "quite impossible, inasmuch as no one would believe what was promised him but would merely laugh at such utterances as being vain pretenses" (G, 422, 31).

This maxim cannot, therefore, be willed to become a universal law of nature, since one would not be able to borrow money successfully by acting on that maxim. As Guyer (2007, 85) explains, "no one in his right mind would accept such a promise, and thus one's own plan of getting out of trouble by making a false promise would be

[2] Some authors like Guyer (2007) regard this alternative statement of the first formulation of the categorical imperative to be a separate version of it, distinguishing between the FUL and the "Formula of the Law of Nature" (FLN). For the purposes of this paper, I will regard the FUL and the FLN as equivalent and refer to the principle as the FUL only, where "the universal law" is to be understood as a universal law of *nature*.

impossible" in a world where making false promises is a universal law of nature. Since this maxim is not universalisable, we must conclude, from the FUL, that making a false promise is morally wrong.

According to the FUL, we may perform similar thought-experiments such as this one in order to derive the duties that we have to others and to ourselves. To determine whether any given action is morally permissible, all we have to do is consider what would happen if the maxim on which the action depends were to be a universal law of nature. Any maxim that fails the universalisability test violates a duty, since Kant thinks that the FUL expresses "the universal imperative of duty" (G, 421, 30).

Wood (1999) and Guyer (2007) point out a further feature of the FUL; that is, that it allows us to distinguish between perfect and imperfect duties. A perfect duty is one that "permits no exception in the interest of inclination" (G, 421, 30, footnote 12). This means that it is a duty in the strictest possible sense – Kant's perfect duty to refrain from committing suicide applies to all people in all circumstances, and we are never justified in violating that duty, regardless of our inclinations. As Kant says, one may *never* take one's own life, even when one is "reduced to despair by a series of misfortunes" and "feels sick of life" (G, 422, 30).

An imperfect duty, on the other hand, allows "leeway" in its fulfilment (Russell in Audi, 1999:249). Kant explains that we have a duty to develop our talents, but this duty is imperfect because we are able to choose *which* talents to develop. For instance, a person might have a natural aptitude for operatic singing as well as for pole vaulting, but financial and time constraints might make it impossible for her to develop *both* of these talents. She indeed has a duty not to allow her talents to be neglected, but, since this duty is an imperfect one, it is permissible for her to choose either talent to develop while neglecting the other one. As long as she develops one talent and does not "indulge in pleasure rather than... improving [her] fortunate natural aptitudes" (G, 423, 31), she does not violate her duty by training for the pole vault rather than singing. Russell (in Audi, 1999: 249) makes the distinction clear: "the duty to help those in need is an imperfect duty since it can be fulfilled" in many

ways, while "the duty to keep one's promises" is perfect, since it does "not allow one to choose which promises to keep" – we must keep them all.

Kant claims that we may determine whether a duty is perfect or imperfect by making use of the universalisability criterion. A duty is perfect if "the opposite cannot become a universal law" (LE, 29:609, 232), while it is imperfect if the opposite *can* logically be a universal law, but if we cannot consistently *will* that it become one. This distinction leads to two separate forms of the universalisability test, as Wood (1999: 84) explains – these are known as "the 'contradiction in conception' (CC) and the 'contradiction in the will'... (CW)" tests[3]. Failing the former means that a maxim violates a perfect duty, while failing the latter means that it violates an imperfect one (Wood, 1999).

If a maxim fails the CC test, it means that we cannot even conceive of its being a universal law without contradiction. Making false promises falls into this category. For our false promise to succeed, we require a world in which people habitually do *not* make false promises; otherwise, as explained previously, our promise will never be believed by anyone and our aim of borrowing money on the basis of a lying promise will be frustrated. This is contradictory because we prescribe a rule of action for the rest of the world (making only true promises) while doing the opposite ourselves. Thus the maxim of making false promises "cannot without contradiction even be thought as a universal law of nature" (G, 424, 32), which means that it violates a perfect duty.

It seems as though most cases of lying would fail the CC test, since lying depends on others believing the lie to be successful. Therefore, with lies, we generally prescribe a different rule of action to the rest of the world than the one we intend to follow ourselves. These considerations allow Kant to reach his rigorous conclusion about telling the truth, since the CC test identifies it as a perfect duty.

[3] As Wood (1999: 84) points out, the terminology used to name these two forms of the universalisability test was first devised by Onora O'Neill.

Even if a maxim passes the CC test, it might still fail the CW test. This test considers whether we could consistently *will* the maxim to be a universal law of nature – if we cannot, it violates an imperfect duty. Helping others is an example of an imperfect duty. There is no logical impossibility or contradiction in a world where everyone refuses to help those in need – we can easily envisage a society in which people never do anything for others. However, Kant thinks that we could not rationally choose to live in such a society: "I cannot will that lovelessness should become a universal law, for in that case I also suffer myself", because nobody will help *me* when I am in need (LE, 29: 609, 233). Since we cannot will the maxim to become a universal law of nature, helping others is an imperfect duty. The following passage summarises the application of the two tests clearly: "With perfect duties, I ask whether their maxims can hold good as a universal law. But with imperfect ones, I ask whether I could also will that such a maxim should become a universal law" (LE, 29: 609, 232). In a later section of this dissertation, I will apply the FUL to the question of lying to the murderer at the door in detail.

The Formula of Humanity (FH)

Kant's second formulation of the categorical imperative (the FH) reads as follows: "Act in such a way that you treat humanity, whether in your own person or in the person of another, always at the same time as an end and never simply as a means" (G, 429, 36). This means that we must always allow others the freedom to pursue their own goals without interference and treat them as rational beings capable of making their own choices, since "every rational being... exists as an end in [it]self and not merely as a means to be arbitrarily used by this or that will (G, 428, 35). For Kant, all our ethical duties "arise from the obligation to make each human being's capacity for autonomous choice the condition of the value of every other end" (Korsgaard, 1986 :331).

As Guyer (2007: 90) points out, Kant is not very helpful regarding what it means to treat someone as an end: "Kant's comments... tell us a little about what it is *not* to treat a being as an end in itself, but do not tell us very much positive about what it *is* to treat someone as an end in itself." However, the basic idea seems to be that we should respect the wills of others and refrain from manipulating them. Kant

emphasises that "[m]an is not a thing" (G, 429, 30) that may be used by others to reach their own aims; instead, "rational beings... should always be esteemed... as beings who must themselves be able to hold" their own ends (G, 430, 37).

This understanding of what it means to treat someone as an end in herself is supported by Kant's examples. He points out that someone who makes a false promise to repay money "intends to make use of another man merely as a means to an end which the latter does not likewise hold" (G, 429, 37). The lender cannot share the same end as the borrower in this situation, since he is being deceived about what the end of the transaction is. Korsgaard (1986: 333) states the point clearly: "[i]f you make a make a lying promise to get some money, the other person is invited to think that the end she is contributing to is your temporary possession of the money" when "in fact, it is your permanent possession of it."

In this example, the borrower is manipulating the lender in order to achieve his own goals, and the lender is unable to make a free, autonomous choice: she "cannot possibly concur with" this "way of acting toward" her (G, 429, 37) since she is deceived. This is what makes the action morally impermissible, on Kant's view. Guyer (2007: 93) notes that this gives us an idea of what it means to treat others as ends in themselves: it is "to ensure that they retain what is essential to their own humanity, namely the right to set their own ends freely or to make the ends of others their own by freely consenting to them." On the FH, therefore, lying is impermissible because it deceives others about what the ends to which they are assenting are, rendering their "free consent" to them impossible.

The Formula of the Kingdom of Ends (FKE)

The final formulation of the categorical imperative requires us to regard ourselves "as legislator[s] in a kingdom of ends rendered possible by freedom of the will" (G, 434, 40). This means that we look at morality from the point of view of a lawmaker who is able to create the laws that govern a kingdom of which the lawmaker is himself a member. The legislator must also obey the laws, since Kant describes him as someone who "legislates... universal laws while also being... subject to" them (G, 433, 40). Kant thinks that any actions that violate the laws we would make as universal

legislators in a kingdom aimed at protecting the autonomy of all its subjects are morally impermissible. The formal statement of the FKE is as follows: "Act in accordance with the maxims of a member legislating universal laws for a merely possible kingdom of ends" (G, 439, 43).

It is important to note that these laws are intended to govern the kingdom of ends, which is a realm with very specific features. This kingdom of ends is "merely possible", so Kant does not intend it to exist in reality, but rather in thought. As Wood (1999: 166) explains, "the realm of ends... is a *moral* realm, the idea of which determines what *ought* to exist". In other words, the kingdom of ends is an ideal realm in which every member respects the autonomy and freedom of choice of every other member, so that the ends are "harmonious and reciprocally supportive".

The legislator in the kingdom of ends would, consequently, make laws to respect the autonomy of his subjects, aimed at the "mutual furthering of the ends of all rational beings in a single unified teleological system" (Wood, 1999: 166). This means that the laws cannot be arbitrary – instead, the laws in the kingdom of ends must proceed "from reason" (Wood, 1999: 157). It seems that the FKE can therefore be used as a decision-making rule, similar to the FUL. To determine whether or not an action is morally permissible, we have to consider whether its underlying maxim is consistent with a law that we would make in our position as universal legislator of an ideal realm where autonomy is revered and protected.

If we apply the FKE to the case of false promises, it is clear that making a lying promise is inconsistent with the laws that a universal legislator would make in this ideal realm. Autonomy cannot be preserved if false promises are permitted, because lying "treats someone's reason as a tool" (Korsgaard, 1986: 334) – it does not permit her to make her own, autonomous decisions. As we have seen earlier, the lying promise deprives the lender of the freedom to choose her own ends and "the deceiver tries to determine what levers to pull to get the desired results" (Korsgaard, 1986: 334). As Korsgaard (1986: 334) explains, deception cannot be permitted in the kingdom of ends because "it is a direct violation of autonomy", that is, of the principal, sacred value of the kingdom of ends. However, as I will explain in a later

section, there are some cases where violations of autonomy may be permitted, even in the kingdom of ends.

From this exposition of Kant's three formulations of the categorical imperative, it seems that he indeed has reason to believe that lying is always morally wrong. The FUL would outlaw lying because the maxim of never telling the truth cannot be universalised; the FH claims that it does not treat others as ends and, therefore, goes against their humanity; and it is incosistent with the laws of a kingdom of ends because it violates the principle of autonomy. In the subsequent sections, however, I will argue that lying is not necessarily always incompatible with these three principles and that there is a strong case for believing that Kant's moral theory does not require us to tell the truth at all times. I will consider the three formulations of the categorical imperative in turn, with specific reference to the case of lying to the murderer at the door.

III. The FUL and Lying

As explained in the previous section, it is necessary to determine what the maxim is on which we act when we perform an action in order to apply the FUL to it. Then we consider whether we can conceive of this maxim to be a universal law of nature, as well as whether we would *will* it to be one. The first part of this section will be devoted to determining what the maxim is on which we act when we tell a lie to the murderer at the door. Next, I will try to show that the relevant maxim is, indeed, universalisable so that the lie is permissible according to the FUL, concluding that the FUL does not require us to tell the truth at all times and in all situations.

In the *Metaphysics of Morals,* Kant seems to endorse the position that there are different possible instances of lying and that these should be considered individually (MM, 6:431, 554). For example, he asks whether "an untruth from mere politeness" can "be considered a lie"(MM, 6:431, 554) and claims that other lies, such as those that are told to cover up a crime, appear to be more serious. It is therefore plausible to argue that there are different maxims underlying different instances of lying – a principle such as telling the truth "can be embodied... in many different ways" (O'Neill, 2002: 331). We therefore need to consider a maxim that is quite specific to

the case when we apply the FUL. As Guyer (2007) notes, it is essential to specify the maxim with exactly the right degree of generality if we are to make use of the FUL meaningfully.

What type of maxim meets this requirement in the case of lying to the murderer at the door? It seems as though a very general maxim such as: "I shall sometimes lie" (hereafter abbreviated as M1) is not appropriate, since it fails to capture some morally relevant features of the situation. Thomas Pogge (1998: 189) stresses that a maxim contains three parts: it is "an ordered triplet consisting of a type of circumstances S, a type of conduct C and a type of material end E". In this way, it can function as a guide for action – whenever we are under similar circumstances and have the same end, we propose to conduct ourselves in the way described.

The maxim M1, above, omits both the end and the circumstances. It tells us nothing about the conditions that hold, since the word "sometimes" is too vague. Furthermore, it does not express what we hope to achieve by lying – is our end to injure someone, to save a life or to start a nuclear war? Intuitively, it seems as though we would permit the lie if it saves a life, and condemn it if it injures someone or starts a war. M1 simply does not tell us enough about the circumstances surrounding our proposed lie for us to make a meaningful moral judgement. As Marcus Singer (1954: 590) points out, the categorical imperative "must always be applied to an action considered as taking place in certain circumstances, or for a certain purpose".

Barbara Herman (1993: 142) explains that, if our maxims are too general, then we end up with "rigoristic moral requirements that vitiate any hope that the [categorical imperative] procedure can be morally supple". To prevent the FUL from yielding moral duties that are too strict, we ought to formulate the maxims at "the correct level of description" (Herman, 1993: 142); that is, the level at which all the morally relevant features of the situation are taken into account.

However, we must also guard against making our maxims too specific. Guyer (2007) points out that *any* maxim that is highly restricted is likely to pass the universalisability test. He considers a red-headed bank robber, Ignatz

MacGillycuddy, who proposes to act on the maxim "[a] red-haired person named Ignatz MacGillycuddy should rob any bank that is open north-east of his house at 5 p.m. on any Thursday" (Guyer, 2007: 139) and shows that this maxim is universalisable. The maxim passes the CC test, since there is no logical impossibility or contradiction that would arise if everyone were to act on it.

Furthermore, the maxim even seems to pass the CW test. There are likely to be so few red-headed people named Ignatz MacGillycuddy with banks north-east of their houses that, even if *each of them* were to act on the maxim, there would not be enough bank robberies "to bring down the whole banking system" (Guyer, 2007: 139) or to put even one bank out of business. It therefore seems as though this very specific maxim can be universalised successfully, but this result is problematic because we would ordinarily say that most instances of bank robbery are morally impermissible (assuming there are no mitigating circumstances that might need to be considered).

From the preceding discussion it is clear that the maxim that applies to the case of lying to the murderer at the door can be neither too general nor too specific. I propose that the maxim "Whenever I sincerely believe that telling a lie is necessary to save someone's life, I will lie to prevent that person's death" (M2) is a suitable one. It meets Pogge's (1998) requirements of having a set of circumstances ("whenever I sincerely believe that telling a lie is necessary to save someone's life"), a type of conduct ("I will lie") and a material end ("to prevent that person's death"). Moreover, this maxim is not a general action description that would yield undesirably strict duties (such as M1 is), but it is also not so specific that it will automatically pass the CC and CW tests (as the bank robber's maxim will). Having expressed the maxim in this way, we are now in a position to apply the CC and CW tests to it to determine whether or not it is universalisable.

First, it seems plausible to argue that M2 passes the CC test. Although the success of the lie depends on its being believed, the world in which M2 is universalised is not one in which people routinely lie. In this world, statements would *generally* be believed, because people would tell the truth unless they *sincerely* believed that a lie

would save another person's life. Lies would, therefore, only occur under very specific circumstances, and it seems reasonable to think that situations in which telling lies could save lives are likely to be rare. It does not seem as though adopting M2 would undermine the institutions of making promises, drawing up contracts and everyday discourse (which require habitual truth-telling for their success).

This result seems satisfactory, because it squares with how human beings generally view the requirement to tell the truth. Usually, we accept that people are truthful *except* under extraordinary circumstances, and we generally believe their utterances even though we are aware that they sometimes tell lies. The fact that these circumstances exist does not cause us to disbelieve everything we are told and there is, therefore, no inherent contradiction in adopting M2.

Maxims that fail the CC test end up being self-defeating, as Wood (1999: 89) explains, because the universalisation of these maxims "would simply render it impossible to achieve one's ends" by acting on them. For instance, if the maxim of making a lying promise to repay money one has borrowed whenever one is in financial difficulties were to be universalised, it would never be possible to attain money by acting on that maxim; "nobody would trust to a promise, or therefore do anything because of it" and "promising would abolish itself, and thus automatically cease" (LE, 29:608, 232).

Thus, the success of the lying promise *depends on* its being the case that the maxim of making such promises is not a universal law of nature. In order for someone to believe me when I say that I will repay the money, it must be that people generally *keep* their promises to return money that they borrow. Certainly, nobody would lend me money if they knew that everyone habitually made idle promises to gain money with no intention of keeping them – crucially, the promise can only succeed in a world where its underlying maxim is *not* a universal law, which leads to a contradiction. For the action to succeed, I must be the only person who acts on that maxim, so it cannot be universalised. As Kant says, someone "who fails to keep his promise does not will that this should become a universal law; he merely wishes to

exempt himself alone from this law" (LE, 29: 608, 232) while, paradoxically, expecting everyone else to obey it.

The success of an action based on M2, on the other hand, does not require others not to adopt the maxim. I may achieve my aim of saving my friend's life by lying to the murderer if it M2 is a universal law of nature, since other people would generally believe my utterances except under the very special conditions where they knew for certain that I sincerely believed that someone's life was in danger. The murderer has no cause to disbelieve what I say when I tell him that my friend is not at home, since people would generally be truthful in a world where M2 is universalised. Thus universalising M2 would not undermine the success of actions based on it (whereas universalising other maxims that fail the CC test would).

One might argue, however, that lying might not be successful in saving the victim's life in a world where M2 was a universal law of nature. The murderer at the door would also be governed by M2, so he would know that people usually tell lies when they believe that doing so could save someone else's life. If I then lie to the murderer regarding my friend's whereabouts, he will probably not believe me, since everyone would lie under these conditions! In a world where M2 is a universal law of nature, it therefore seems as though telling the truth to the murderer might, paradoxically, be more effective in saving her life than lying to him. Perhaps, then, M2 cannot be universalised without contradiction after all.

However, this argument depends on the assumption that the murderer has perfect knowledge regarding my beliefs. M2 states that I would lie only when I sincerely believe that the lie would save my friend's life – however, the murderer is unlikely to know that I believe this, assuming he does not knock on my door and say, "I have come to kill your friend – is she here?". If he simply came to the door and asked if my friend was in, he would believe my lie unless he knew for certain that I was aware that he was intending to murder her. And if he *did* know this for certain, asking me about my friend's whereabouts would be a meaningless exercise for him, as he would know that I would lie to him anyway. Thus this argument does not show that

adopting M2 leads to a contradiction, and it seems plausible to claim that this maxim passes the CC test.

We now consider the CW test – can we consistently will a world in which M2 is a universal law of nature? Would we assent to living in a world where people routinely lie when they sincerely believe that doing so could save an innocent life? I certainly think that it is reasonable to say that we would. It seems rational to say that we would agree to live in a world where there is a chance of being lied to when we know that the lies are intended to save the lives of others.

Herman (1993: 50-51) suggests that we should supplement "the Kantian procedure with [John Rawls'] veil of ignorance" when applying the categorical imperative to moral judgements. This means that, when we decide whether or not we could will M2 to be a universal law, we do not know whether we will end up being the liar, the would-be murderer or the intended victim in the world in which M2 holds; so we must examine the consequences of acting on M2 from the viewpoints of all three of them.

Telling the truth to the murderer would have devastating consequences for the victim (her death), and perhaps also for the person telling the truth (guilt and regret at contributing to her murder), while the murderer would be able to fulfil his goal of killing her and would not be harmed. Lying, on the other hand, would save the life of the victim while frustrating the aims of the murderer, but this might even be beneficial to the latter as he could be spared a lengthy and undesirable prison sentence. It therefore seems reasonable to say that the comparative inconvenience caused to the murderer by the lie is a much better outcome than the loss of life resulting from telling the truth. If we were in the position of the victim, we would certainly will the lie to be told – therefore, it seems as though M2 passes the CW test.

Once again, this result is consistent with the way in which we normally view lies. We usually do not blame people who lie under extraordinary circumstances. For example, when lies are told in order to avoid insulting someone, to arrange a surprise party or to play a practical joke, the liar is generally not regarded as morally reprehensible –

instead, the special conditions surrounding the lie are taken into account. A lie to save someone's life certainly appears to fall into this category, and it follows that we *are* able to consistently will a maxim such as M2 to become a universal law. We might argue that a world in which lies are sometimes told, with the aim of saving lives or arranging surprise parties, is preferable to one where everyone always tells the truth. Thus I feel that our maxim M2 is indeed universalisable, which means that telling a lie to save a life is permissible on the FUL. This version of the categorical imperative does not, therefore, require Kant to maintain that lying is *always* wrong.

One might object to this conclusion by pointing out that M2 is specified in such a way that it cannot fail the CC and CW tests – in other words, we have tailored it to get the answer that we want. This is a familiar objection to the categorical imperative: as Guyer (2007: 139) notes, the FUL is problematic because it "yields so many false positives, that is, maxims that should be impermissible on any reasonable view of morality but that turn out to be permissible". Because "any particular action could be performed under an indefinite number of different maxims" (Guyer, 2007: 141), it seems as though we merely have to fashion the maxim in the correct way in order to obtain the desired result. If we want a given action to turn out to be permissible, all we have to do is devise a maxim that plausibly describes the action, but clearly passes the CC and CW tests. Wood (1999: 103) points out that a maxim may conceivably be described "in such detail... that its becoming a universal law of nature would foreseeably have *no* consequences" besides one's person acting on it "on this particular occasion", so that an application of the FUL might justify an obviously immoral action.

This point may be clarified in considering a further maxim (M3) that may be used to describe the action of lying to the murderer at the door. I specify M3 as follows: "Whenever I am in a position to lie to someone, I will lie in order to deceive that person". Now suppose that I have an intense desire to deceive others whenever possible, and my only motivation in lying to the murderer is from the satisfaction I will gain from misleading him. Here, I do not care whether or not he kills my friend, and I do not lie out of any concern for her safety – perhaps I would lie to him even if the lie directly lead to the victim's death.

If we now apply the FUL to M3, it seems obvious that it will fail the CC test in the same way that making a lying promise to get money would. Nobody would believe my lie if it was a universal law of nature for everyone to deceive others and my action would be self-defeating. Consequently, actions based on M3 are morally impermissible, according to the FUL. This means that lying to the murderer is wrong if we act on M3, while it is permissible (or even obligatory) if we act on M2. As a result, the morality of an action seems to depend on nothing more than the way in which its underlying maxim is specified.

This means that our ability to use the FUL requires us to know whether an action is right or wrong beforehand, so that we can adapt the maxim to fit our intuitions, and this questions the efficacy of the FUL as a decision-making rule. If we already know whether or not an action is permissible, and if we are able to tailor the maxim in order to get the desired result, it seems as though any attempt to use the FUL is futile. The difference between acting on M2 and acting on M3 is that our intention is to save a life in the former case and to deceive in the latter. Concluding that acts based on M2 are permissible while those based on M3 are not presupposes that we know already that saving a life is a noble aim while deceiving others is reprehensible – as such, using the FUL to determine that one action is right while the other is wrong is nothing more than circular reasoning.

This objection to the FUL is indeed a powerful one. As Wood (1999: 105) explains, the FUL is "not enough" – we "need a further specification of the moral laws themselves that" it "is commanding us not to violate" and the categorical imperative itself is not adequate for this purpose. However, we may overcome this difficulty if we see the FUL not as a decision-making rule, but rather as an *explaining feature* underlying all right action. In other words, the FUL does not determine whether or not an action is right, but rather provides us with a justification for its rightness. It is an objective standard to which we appeal when explaining why a given action (already known, perhaps intuitively, to be wrong) is impermissible under a specific set of circumstances. This view of the categorical imperative, which is the one proposed by Stratton-Lake (2000), Timmermann (2007) and Hanna (2009), will be discussed in more detail in section VI of this dissertation.

A further objection to the preceding analysis of the FUL is the claim that the categorical imperative should not be applied in a consequentialist way. In considering whether M2 is universalisable, we are examining the likely consequences of telling the lie for the murderer, the liar and the intended victim – we therefore draw our conclusions based on consequences, as a different moral theory like utilitarianism would. This is objectionable because Kantian ethics requires us to consider the action itself, independently of its consequences, so one might argue that this way of drawing the conclusion is mistaken.

This objection fails because the application of the categorical imperative does not depend on the consequences of a *particular* action; rather, the conclusion is drawn from a thought-experiment that considers the result of an action-*type*. This is why we consider a general maxim on which the action rests instead of the act itself. We do not know what the consequences of the action will be – at best, we have a belief about what the *likely* result of lying to the murderer will be. It is with this end (the preserving of the victim's life) in mind that we perform the action, and the categorical imperative is applied to actions of the same type under the same circumstances, intended to achieve the same end. We imagine *everyone* habitually performing actions based on the maxim, not the consequences of one person acting on it in a given set of circumstances.

Kant emphasises the point that we do not know the consequences of our actions in advance: "whoever tells a lie... must answer for the consequences resulting therefrom... regardless of how unforeseen... [they] may be" (RL, 427, 65). The universalisation of M2 does not require that we know the result of telling a lie to the murderer in a specific case. Instead, the universalisability test considers what would, in general, happen if lies were usually told with the aim of saving lives.

Kant points out that my act of lying to the murderer might actually *cause* the death of my friend. If I tell the murderer that my friend is not at home and she "has... (unbeknownst to [me]) gone out" (RL, 427, 65), the murderer might leave, find her outside and kill her. My lie has therefore contributed to the death of my friend and I "may be justly accused of having caused" the murder (RL, 427, 65). This shows that

morality should not depend on consequences, since these are so uncertain – the best we can do is to act on acceptable maxims. If we have adhered to the demands of duty, it is reasonable to think that we should not be blamed for the unforeseen consequences that result from our actions. Thus the Kantian application of the categorical imperative is *not* a consequentialist one and the objection to it is unfounded. Since it considers the desirability of universalising an action type instead of the likely consequences of a specific action, the FUL will still allow us to act morally even if our actions end up having undesirable consequences that we could not have foreseen.

Importantly, Kant makes use of the fact that my lie could cause the victim's death to draw the conclusion that lying is impermissible. He says that "a well-intentioned lie can become punishable in accordance with civil law because of an accident" while "if you have adhered strictly to the truth, public justice cannot lay a hand on you" (RL, 427, 65). Thus Kant thinks that I may be punished if lying to the murderer causes the victim's death (as described previously), but not if telling the truth results in her killing – in the latter case, I do nothing wrong as I fulfil my duty of truthfulness, whereas I may be regarded as an accessory to the murder in the former. This means that I should tell the truth to the murderer, because nobody would punish me for causing my friend's death by being truthful.

I think that Kant is mistaken in drawing this conclusion. First, the number of cases where telling a "well-intentioned lie" (RL, 427, 65) will have unforeseen, negative consequences is likely to be comparatively small. It seems reasonable to claim that the lie to the murderer will probably succeed in most cases, and that the exceptional cases do not mean that the lie is impermissible. In the same way, my action of driving to buy groceries could result in a car accident whereby a pedestrian is harmed; however, the fact that such an accident is possible does not render it morally wrong for me to drive to the store. The intention behind telling the lie is a noble one, since we sincerely believe that our lie can save an innocent life if we act on M2, whereas telling the truth appears to betray our friend.

Thus it is difficult to agree with Kant's view that we should tell the truth merely because we cannot be punished for doing so in a court of law. We might counter,

with Singer (1954: 584), that telling the truth in this situation would be wrong "because it would help destroy the bonds of human trust, in terms of which one person may be relied on to shield the other against an oppressor."

Furthermore, Singer (1954: 584) claims that Kant's conclusion that it would be wrong to lie is "question-begging". Kant says that lying to the murderer would be morally wrong because "it does harm to humanity in general, inasmuch as it vitiates the very source of right" (RL, 426, 64-65). In other words, telling a lie is impermissible because it is morally wrong – but "whether it would be wrong is precisely the point in question" (Singer, 1954: 584). This certainly seems to be a valid criticism of Kant's conclusion.

We have another reason for disagreeing with Kant's conclusion if we consider the maxim on which telling the truth to the murderer depends. This would be the opposite maxim to M2, which can reasonably be formulated as "Whenever I sincerely believe that telling a lie is necessary to save someone's life, I will tell the truth to respect people's freedom to choose their own ends, thereby putting the life of an innocent person in danger" (M4). On this maxim, the circumstances are the same as in M2, but the conduct and end are somewhat different. The end of M4 is to respect the free choices of others at all times (even when this endangers another human being); it cannot be to put someone else's life in danger, since intentionally exposing someone to danger cannot be a reasonable end to adopt when formulating a maxim. Now, if M4 fails either the CC or CW tests, then, by the FUL, it is our duty to act on the maxim that is its opposite (in this case, M2). This would mean that we are morally *required* to lie to the murderer at the door.

It should be obvious that M4 does not fail the CC test. There is no contradiction involved in a world where everyone routinely tells the truth, even when doing so can endanger someone else's life. We can logically conceive of such a world[4], and there

[4] A recent film titled *The Invention of Lying* (2009), by Gervais and Robinson (Focus Features International Films) portrays a world in which everyone routinely tells the truth and nobody has ever lied. The writers are able to imagine and create such a fictional world on film, which shows that there is no logical impossibility precluding its existence.

are certainly no institutions such as promising that are undermined by adopting M4 as a universal law of nature.

Turning to the CW test, can we consistently will that M4 be adopted as a universal law of nature? As noted previously, the adoption of this maxim will most likely destroy bonds of trust between people (Singer, 1954), which is undesirable. Furthermore, adopting M4 will mean that the murderer is able to get *any* information from us that might assist his projects, since we are bound to answer any of his questions truthfully. He might ask, for example, "When is your friend at her most vulnerable?" or "When is she going to be alone here in the house?" Honest answers to questions such as these will certainly aid the murderer in achieving his goals, and it does not seem as though we could rationally will a world in which people routinely assist murderers through such extreme honesty.

If we follow Herman's (1993) suggestion of approaching the question from behind a veil of ignorance, we also see that we could not will M4 to be a universal law. As with M2, we must examine the situation from the point of view of all three parties concerned. It is hard to see how we could, as possible victims of murders, consistently will our friends to co-operate with our killers and give our location away. Instead, we would in all likelihood will that our friends lie on our behalf – in fact, the only person who could reasonably will the truth to be told in this case is the murderer. Therefore, willing M2 as a universal law of nature seems far more rational than willing M4, so M4 seems to fail the CW test. This means that, based on the FUL, lying to the murderer emerges as an imperfect duty to our friend. Thus Kant's moral theory does not require him to arrive at the strict conclusion that lying is *always* wrong, and the case of the murderer at the door appears to be an exception.

Two other cases are worth exploring here. What if the liar is under oath, and knows that telling the truth is likely to result in an innocent person's being convicted of a crime and executed because of circumstantial evidence? This case seems to be similar to that of the murderer at the door, as a lie can save the life of an innocent person. However, there is the crucial difference that the lie must be told under oath in a court of law. Does this additional feature of the case make the lie impermissible?

Here, Wood's (2009) emphasis on Kant's distinction between a "lie" and a "falsiloquium" in the *Lectures on Ethics* appears to be relevant. A "lie" is "the only kind of untruth... that directly infringes upon another's right" (MM, 6: 238, 394, footnote), whereas a "falsiloquium" (or falsification) is an untruth that "violates no duty of right" (Wood, 2009: 4). This means that a lie is properly understood as an untruth where the hearer has a right to the truth (such as in contracts); or, as Wood (2009: 4-5) describes it, as a false "declaration", that is, a false "statement that occurs in a context where others are authorized... to rely on the truth". A falsification, on the other hand, is a false statement where the speaker "is merely communicating his thoughts to" the audience and "it is entirely up to them whether they want to believe him or not" (MM, 6: 238, 394). The crucial distinction between the two is whether or not others "are authorized to rely upon" them (Wood, 2009: 5) – a lie occurs only in situations where we make a false declaration when others can reasonably expect us to tell the truth. If this requirement is not met, the statement is a mere falsification.

When the murderer at the door asks whether my friend is in, it seems as though he is not authorised to rely on my information. I am not under oath, there is no legal contract between us and I am not seeking to "deprive" him "of something that is rightfully his" (Wood, 2009: 5). This means that it is "up to him" whether he wants to believe me or not, and he has no inherent right to a declaration from me. My statement that my friend is not home is, therefore, a falsification rather than a lie and, as we have seen, Kant's moral theory permits us to tell an untruth to the murderer under these conditions.

A declaration in a court of law, on the other hand, is a situation where others (that is, the judge and jury) are authorised to hear the truth from us. As Wood (2009: 5) points out, such a declaration "makes the speaker liable by right, and thus often liable to criminal penalties or civil damages if what is said is knowingly false." Kant himself considers false declarations under oath to be impermissible, noting that a perjurer "by his deceit removes all credit and worth from the *instrumentum* of public trust, and commits a greater crime than any wrought by open force" (LE, 27: 701, 427). This quotation from Kant shows that a universal law permitting perjury would fail the CW

test, since the removal of "all credit and worth" from the justice system is surely not something that we could consistently will.

Thus it seems as though the FUL permits us to lie to a murderer to prevent the death of an innocent, but not to lie under oath in a court. The relevant difference between the cases is that the court requires us to make a declaration, or a truthful statement, while the murderer has no such claim on us. Falsifications are sometimes permissible, whereas lying declarations are not. This certainly seems to be a plausible conclusion.

Schapiro (2006) provides a counterexample to this case, where someone is required to make a declaration in Nazi Germany. Here, the person requiring the declaration is a *bona fide* government official who is searching for Jews to exterminate. He knocks on my door asking whether there are any Jews in the house and, since he is a representative of the government, he is authorised to rely on the information I provide. Am I required to tell him about the innocent Jew hiding in my cellar, when I know that he will murder her and when I do not support the Nazi government? In this situation, it seems as though the official is entitled to expect the truth from me, and a false statement on my behalf will amount to a lie (and not a mere falsification).

However, the key feature of this case is that the official represents a corrupt, evil government whose policy of killing Jews cannot be justified. As Schapiro (2006: 52) explains, the Nazi official's "end is blatantly at odds with" Kantian ideals, even though he is being honest about his intentions. Adopting a maxim such as "Whenever I am in a position where I have to make a declaration to an obviously evil government, and I sincerely believe that a lie is necessary to save someone's life, I will lie to prevent that person's death" (M5) does not seem to fail the CW test – if the government is indeed "obviously evil", it seems as though we could will a world in which people routinely commit perjury to save the lives of those whom the state persecutes.

This counterexample appears to provide a case where making even a false declaration might be permissible on the FUL. Even if this argument is not entirely convincing,

Schapiro (2006) shows that lying to the Nazi is permissible on the third formulation of the categorical imperative (FKE), an argument that I shall consider in the fourth section of this paper. However, it seems clear that there are at least some cases where the FUL permits lying, so Kant's conclusion that lies are never permissible appears to be too hasty.

One problem with this counterexample is that we might be required to tell the truth, even to an evil government, because we have a duty to obey the state at all times. Peter Nicholson (1976: 215) attributes this view to Kant: "it is my contention that he regards both not lying and not resisting the sovereign as absolute moral duties". This is the case even if the government is corrupt and unjust. As Kant explains, citizens of a state "cannot offer any resistance to the legislative head of state" since "a people has a duty to put up with even what is held to be an unbearable abuse of supreme authority" (MM, 6: 320, 463). On this view, both the duty to tell the truth and the duty not to resist the state permit of no exceptions whatsoever, and the counterexample of Nazi Germany fails.

There are, however, some difficulties with attributing this rigorous view to Kant. First, as Schwarz (1964: 127) explains, we may distinguish between "active" and "negative" resistance. Here, "active" resistance refers to deliberate acts taken to "compel the government to a certain procedure", such as an armed revolution to overthrow the state, while "negative" resistance is merely a "*refusal* of the people... to comply with the demands" of the government. Since lying to the Nazi officer constitutes a refusal to comply with his demands rather than coercing him to act in a certain way, it is a form of "negative" resistance.

Schwarz (1964: 130) claims that Kant would permit citizens to engage in negative resistance and shows that we "can quote passages that clearly exonerate Kant from any charge of having deprived the individual citizen of a right of resistance" and provides a number of citations from Kant's works to support this view. Perhaps, therefore, Kant's prohibition of resisting the state applies only to active resistance, so that we are "not justified in killing a tyrant in order to preserve the lives of... even millions of his subjects" (Beck, 1971: 420), but it may be permissible to lie to a

representative of an unjust government as this is a negative form of resistance involving non-compliance rather than aggressive rebellion.

A second reason for thinking that Kant might allow us to lie to the Nazi official is that, in Kant's times, citizens of unjust states could emigrate easily relative to those living under modern tyrants. As Schwarz (1964: 130) points out, a "terrorized individual" of Kant's age, "persecuted by a despotic ruler, usually had to travel only a few miles to enter the province of another sovereign, out of reach of the raging ruler". Since Kant thinks that any subject of a state "has the right to emigrate, for the state could not hold him back as its property" (MM, 6:338, 478), he would probably suggest that someone who felt persecuted by her ruler should simply move if she did not wish to obey the law, and this would not be as difficult as it was for those who wished to escape Nazi Germany.

Since Kant values human autonomy as an ideal, he would perhaps advocate resistance to governments that greatly limit freedom, such as the North Korean one, which did not exist in his times. Reiss (1956: 190) emphasises this point, explaining that Kant "could not have foreseen the modern totalitarian state which is much worse than anarchy" and that his call for us to obey the state, while relevant to the times in which he lived, is perhaps not applicable to our present-day world. Thus I feel that the counterexample is valid – it is not clear that Kant would condemn someone who lies to a Nazi judge and, even if he would, his commandment to obey the state at all times is perhaps outdated when dealing with some of the despotic governments of today.

A further objection to this counterexample and to the other aforementioned cases might be that they are invalid because they are all instances where the perfect duty of telling the truth conflicts with an imperfect duty to others. As Kant points out, "imperfect duties always succumb to perfect ones" (LE, 27: 537, 296) – we always have to fulfil our perfect duties, but "failure to fulfil[-]" our imperfect duties "is not in itself *culpability*... but rather mere *deficiency in moral worth*" (MM, 6:390, 521). Kant thinks that telling the truth is a perfect duty we have to ourselves ("the greatest violation of a human being's duty to himself... is the contrary of truthfulness, *lying*", MM, 6:429, 552). This is because, as explained earlier, the maxim of lying (in

general) fails the CC test because telling a lie successfully is impossible in a world where everyone routinely lies.

Thus it might be argued that we face a conflict between the perfect duty to tell the truth and the imperfect duty not to harm others in the case of the murderer at the door. Since the former takes precedence over the latter, we ought nevertheless to tell the truth despite the possible consequences of doing so. Here, it is regrettable that the situation requires us to breach an imperfect duty to others, but it remains obligatory to tell the truth to the murderer in order to respect his agency – the duty of truthfulness may never be violated. However, even if this consideration means that maxims M2 through M5 are not universalisable, we may provide one final counterexample that appears to show that the FUL permits instances of lying.

Schapiro's (2006) example may be modified so that it plausibly leads to a conflict between two perfect duties to oneself. Suppose that a Nazi judge asks me (a Jew) under oath what my religion is. I know that the judge will order my immediate execution if I tell the truth, but I also know that a lie will set me free since there is no evidence to prove that I am a Jew apart from my response to this question. The dilemma here is whether I should tell the truth and allow the Nazis to kill me, or to lie and prolong my life, while violating a perfect duty to myself to be truthful in all interactions with others.

Kant claims that every person has a duty "to himself as an animal being... *to preserve himself*" (MM, 6: 421). In other words, one should take the necessary steps to stay alive, insofar as this is possible. It looks as though the *prima facie* duty to tell the truth is inconsistent with this *prima facie* duty to preserve my own existence, since I will effectively be signing my own death warrant unless I lie. I might choose to act on the following maxim: "Whenever I am in a position where I have to make a declaration to an obviously evil government, and I sincerely believe that a lie is necessary to save my own life, I will tell the truth in order to respect people's freedom to choose their own ends, thereby putting my life in danger" (M5) and tell the judge that I am a Jew.

It may be argued that this maxim fails the CC test. If my aim in telling the truth to the evil judge is to respect people's freedom to choose their own ends, and if the consequence is that I am executed or imprisoned, then I will fail to achieve my aim because my ability to choose my own ends must necessarily cease if I am no longer alive, or at least be significantly hampered if I am imprisoned. On the one hand, I promote the autonomy of others by telling the truth; on the other, I allow my *own* freedom to be constrained. The maxim of telling the truth at the same time upholds the autonomy of others while severely limiting or even destroying my own autonomy. It is surely plausible to argue that there is a contradiction in adopting, for the sake of autonomy, a maxim that is likely to result in the loss of one's own autonomy. Thus M5 appears to fail the CC test, as it threatens the very value that it seeks to preserve, which is a contradiction.

Furthermore, a premature death will prevent me from continuing to fulfil some of my "many other actual duties" identified by Kant (G, 424, 32), such as cultivating my talents or helping others in need. These are impossible to fulfil if I am no longer alive. Kant explains that "several... duties outweigh a single one" (LE, 27: 537, 296). This means that, if circumstances demand a choice between fulfilling only the *prima facie* perfect duty of telling the truth or fulfilling both the *prima facie* perfect duty of self-preservation and the imperfect duty of developing my talents, I should choose the action that fulfils the larger number of duties.

It therefore seems as though we are morally required to lie in this situation, even though the duty not to lie remains and it is deplorable that we are unable to fulfil it. However, in this unfortunate case, "it is absolutely impossible to fulfil both duties", so we ought to tell a lie "since the ground of [self-preservation] binds more strongly than that of" truthfulness (LE, 27: 537, 296-297). As Schwarz (1964: 129) explains, "commands which place a person in direct conflict with the law of freedom so that their observance would annihilate him as a person... must be resisted" even if they are the law of the country. From the FUL, then, it seems as though we are justified in telling a lie to the Nazi judge to save our own life under these circumstances, even if we break the law by doing so.

This conclusion, however, is not yet satisfactory. As Wood (1999) notes, there are several problems with the use of the categorical imperative in the FUL. He claims that we should see the FUL as a "merely provisional" formulation of Kant's ethical thought and that it is an intermediate step toward the fuller moral law expressed in the other two formulations (Wood, 1999: 97). According to Christine Korsgaard (1986: 347), the FH is a more sophisticated and rigorous principle than the FUL, and can lead to different conclusions. Thus, even though the FUL seems to lend support to the view that lying is permissible under some circumstances, we have to investigate further by considering the other two formulations of the categorical imperative.

IV. The FH and Lying

Both Korsgaard (1986) and Herman (1993) emphasise that, on the FH, lying is impermissible because telling a lie to another person means that we are not treating her with the respect due to all human beings; or, in Kant's terminology, we fail to treat her as an end. Korsgaard (1986: 331) explains the point clearly – when we deceive others, we interfere with their ability to choose their own ends and violate their "capacity for autonomous choice". We treat them simply as mere means to our own ends (which they cannot share as they are deceived regarding what they are) and this is disrespectful and impermissible.

One possible response to this view is that we are entitled to treat the murderer at the door as a mere means to an end because he has evil purposes, so that he does not have any right to hear the truth. This is the original objection to Kant's views from Benjamin Constant (1797) that prompted Kant to write the *Right to Lie* (Benton, 1982). On Constant's view, duties "correspond to the rights of others" and, in a place "where there are no rights, there are no duties" (Constant, 1797: 36, my own translation)[5]. Constant thinks that the murderer has given up any rights he has to hear the truth, so that we no longer have a duty to be truthful to him (although we still have other duties to him such as a duty not to harm him, for instance). However, this objection fails because it misunderstands the key point that we are required to tell the

[5] The original French passage is "L'idée de devoir est inséparable de celle de droits : un devoir est ce qui... correspond aux droits d'un autre. Là où il n'y a pas de droits, il n'y a pas de devoirs".

truth to the murderer not because of any rights *he* has, but rather because lying to him violates a perfect duty we have to ourselves.

Kant argues this point by saying that "the expression 'to have a right to the truth' is meaningless" (RL, 426, 64). The murderer at the door is not in a position to require a true declaration from us, because nobody has "objectively a right to truth"; this means that "by telling an untruth I do no wrong to him who unjustly compels me to make this statement" (RL, 426, 64). Instead, the wrong I do arises as a violation of *duty*, that is, I fail to treat other people in the way I ought to.

Consequently, the lie does not harm the murderer individually, but is instead "a wrong done to mankind in general" (RL, 426, 64). The purposes of the murderer are irrelevant to what we ought to do – treating him as an end requires that we allow him the freedom to set his own goals, no matter how evil they may be. Failure to do so violates the principle of humanity that is central to Kantian ethics, and on these grounds the lie is impermissible. Constant's objection to Kant's position therefore fails, because it is based on a misunderstanding of his theory.

A much better response is the argument developed by Korsgaard (1986) and Schapiro (2003; 2006). Korsgaard (1986: 341) distinguishes between "ideal and nonideal theory". Ideal theory assumes that "everyone will act justly" so that "the ideal" of a "just state of affairs" is possible (Korsgaard, 1986: 342). Under ideal conditions, then, it is assumed that others will be ethical - "people, nature and history will behave themselves so that the ideal can be realised" (Korsgaard, 1986: 342). As I have shown previously, lying would be abhorrent in such an ideal world since it would not permit others to choose their own ends, thereby disrespecting their humanity. As Korsgaard (1986: 333) explains, applying the FH under these ideal conditions means that lying is one of "the most fundamental forms of wrongdoing to others", such is its disrespect for humanity.

Non-ideal cases, on the other hand, arise when we can no longer make the assumption that others will be ethical. This means that we become faced with difficult moral choices so that doing what would ordinarily be the right thing appears undesirable.

Korsgaard (1986: 343) explains that, under non-ideal conditions, it is impossible to realise the ideal of justice – instead, the best we are able to do under these conditions is to follow whichever one "of our nonideal options is least bad, closest to ideal conduct".

On this analysis, then, the case of the murderer at the door is governed by non-ideal conditions. The intentions of the murderer are evil, and he is likely to deceive us as to what they are – Korsgaard (1986: 329) explains that he "must suppose that you do not know who he and what he has in mind", so that "there is probably already deception in the case". Thus, the ideal assumption that others will always act ethically no longer applies, and we are justified in lying to the murderer because he "tries to use [our] honesty as a tool" and we "do not have to passively submit to being used as a means" to the murderer's own, evil ends (Korsgaard, 1986: 338).

Korsgaard (1986: 338) notes that Kant seems to permit lying when conditions are not ideal, explaining that "this is the line that Kant takes". Kant says that "if, in all cases, we were to remain faithful to every detail of the truth, we might often expose ourselves to the wickedness of others who wanted to abuse our truthfulness" (LE, 27: 448, 204). We are required *always* to tell the truth only under the ideal conditions where everyone is "well disposed" – sadly, however, "men are malicious" and the world is non-ideal (LE, 27: 448, 204). This means that we are permitted to tell "a 'necessary lie'... where someone forcibly compels [us] to make a declaration of which [we] know they will make wrongful use" (Wood, 2009: 12). Here, the lie is justified because it is "a weapon of defence" and protects us from having our own humanity disrespected by others (LE, 27: 448, 204). Korsgaard (1986: 340) makes the even stronger point that we are obligated to tell a lie to the murderer as we "owe it to humanity not to allow [-]our honesty to be used as a resource for evil."

Schapiro (2003 and 2006: 49) argues along similar lines, claiming that there are cases where "mitigating circumstances" apply and where we may engage in "defensive deception" to prevent ourselves from being used as a mere means by others. Lies are also justified when we lie to "very young children or the mentally disabled", which Schapiro (2006: 38) calls "paternalistic deception". Since very young children and

mentally ill people are not rational, we may deceive them for their own good, as such deception does not interfere with "their rightful authority to govern" themselves (Schapiro, 2006: 38). Very young children or the mentally disabled are not able to make good choices, since they are not completely rational.
Thus, an act of deceiving such people for their own good does not violate the FH's directive to respect autonomy, since they lack autonomy to begin with. This argument would, therefore, permit lying to a small child about the Tooth Fairy to prevent its distress about losing a tooth.

On the face of it, this seems to be a plausible view to hold. Ordinarily, the FH commands us not to lie, but under the non-ideal or mitigating circumstances where others are either intending to use us as a mere means to their own ends, or where they are unable to make rational choices, a certain degree of deception is permissible. Circumstances such as these result in a "case of emergency" which "subverts the whole of morality" – in such non-ideal conditions "the moral rules are not certain" (LE, 27: 448, 204) and Kant seems to allow that we might be permitted to deviate from the rules that ordinarily bind us.

However, there is an important concern about this view that is raised by both Schapiro (2006) and Korsgaard (1986). This is that Kant regards deception as being "wrong in itself" because of its "manipulative character", which means that "no change in external circumstances could make it right" (Schapiro, 2006: 37). When I lie to someone else, I *always* violate the FH since I fail to allow her to choose her own ends, even if she "has adopted a blatantly immoral" one (Schapiro, 2006: 51). As Korsgaard (1996: 347) points out, lies are always "wrong in themselves, regardless of whether they are told with good intentions or bad."

I think that we may accept the force of this objection and concede that there is always *something* wrong about lying without having to conclude that we have a rigorous duty to tell the truth, even under non-ideal conditions. The key point to consider is that the case of the murderer is an exceptional one where we are faced with a real moral dilemma. This means that there is something undesirable about *any* choice we could make in these circumstances. We may tell the murderer the truth and risk the death of

our friend, thereby violating a *prima facie* duty not to allow others to be harmed when we could prevent their harm; or we lie to the murderer and disrespect his humanity. Assuming these are our only two options, neither of them seems attractive. However, lying is nevertheless permissible in this case since we have no choice but to violate some *prima facie* duty in this rare situation.

This point may be made more clearly using an analogy. I consider a case where there is a shipwreck, and one person has managed to reach the safety of a lifeboat. He sees two people floundering in the water – one of them is his wife, and the other is a doctor who has recently discovered the cure for cancer and is the only living person with this knowledge. I suppose also that both these people are unable to swim and that the person in the lifeboat knows about the doctor's breakthrough – perhaps she made her discovery on board the ship and shared it with the passengers, without divulging any of the details as to what the cure was. Furthermore, the man is unable to reach both his wife and the doctor in time, since they are both drowning and a sufficient distance apart – essentially, the situation is such that saving one will result in the inevitable death of the other.

Clearly, the person in the lifeboat is in an unenviable position. If he saves the doctor, he allows his wife to die, bringing grief to himself and his family and friends; if he saves his wife, he deprives thousands of cancer patients of a cure. This is a true moral dilemma because neither option is attractive. The person in the lifeboat has no choice but to violate some *prima facie* duty, because the circumstances force him to do so.

I think that the presence of this moral dilemma means that we should not blame this person if he chose to save his wife over the doctor, or vice versa. It is entirely impossible for him to avoid doing wrong to *someone*, and his circumstances are, indeed, most regrettable. He must make a choice about whom to harm, and perhaps saving the doctor carries greater moral weight. Since Kant thinks that we should choose the action that fulfils the greater number of imperfect duties, there may be a case for saying that he should save the doctor, since allowing her to die violates the imperfect duty to help more people (the doctor and all the cancer sufferers in the world) than saving his wife would.

However, I think it is at least plausible that we would not condemn him if he chooses to save his wife instead, although I will not argue that here. The point is that saving the doctor, assuming it is the better alternative, still means violating a *prima facie* duty, and that no action that violates no such duty is possible for this person. Thus, in this case, he may be excused for violating such a duty. Of course, if he chooses to save *nobody* and simply rows to safety, then we have a good reason to hold him morally culpable, since he is in a position to fulfil *some prima facie* duty but fails to do so. It is nevertheless plausible to claim that saving one of the two people absolves him from blame for allowing the other to drown.

This example shows how non-ideal conditions permit one to perform an action that would be considered morally wrong under ideal circumstances. It is easy to think of other, similar moral dilemmas where we have no option but to violate a *prima facie* duty to *someone* or something, such as where we can save the life of an endangered plant only by killing the endangered animal that is eating it, or perhaps having to harm an intruder to protect our families. Everyday life sometimes throws up cases such as these, and under these circumstances morality requires that we do the best we can. Fortunately, though, these situations appear to be comparatively rare.

Thus, the FH applies chiefly to ideal circumstances, and we may depart from the rule of respecting humanity if there is no alternative. As Korsgaard (1986: 346) explains, the FH "provide[s] the ideal which governs our daily conduct" but "[w]hen dealing with evil circumstances we may depart from this ideal", since the FH "is inapplicable" in these cases "because it is not designed for use when dealing with evil." Anyone who tries to apply the FH rigorously under all conditions misses this fundamental point – if circumstances make it impossible to choose a course of action that respects the humanity of all the parties involved, we are permitted to act contrary to the prescriptions of this principle.

Another feature of the case of the murderer at the door is that the moral dilemma is actually brought about by the murderer's own actions and not by circumstances alone. It is the murderer who forces us to make a choice about which *prima facie* duty we

are going to violate. This might provide us with an additional reason to lie to him, since the moral dilemma arises solely from his actions; because he has created the non-ideal conditions, we might be justified in disrespecting his autonomy by lying to him.

Furthermore, as Singer (1954: 588) points out, telling the truth to the murderer might "be to treat the *victim* merely as a means to the end of the murderer, ends [s]he, as a victim, cannot... share" (my italics). We might argue that divulging the victim's whereabouts to the murderer amounts to allowing her to become a tool for his evil intentions. If this is the correct interpretation of the case, the moral dilemma is even more severe – both telling the truth and lying violate someone's autonomy, and there is no way of fulfilling the demands of the FH. Once again, I feel that these exceptional circumstances permit us to lie to the murderer, especially because his actions bring about this dilemma in the first place.

One might argue, however, that there are ways of dealing with the murderer that do not require treating him as a mere means. For instance, we could tell him the truth but then stop him from killing the victim by pushing him down the stairs; we could remain silent even if he tortures us to try to extract the information; or we could answer his questions honestly, but then quickly lock the door to prevent his entering the house. None of these actions involve violating the *prima facie* perfect duty of truthfulness, so it might be that they are all preferable than lying to the murderer.

This argument fails because the case of the murderer at the door assumes that we have *no alternative* but to answer the murderer's question – if we had other options that did not involve violating any *prima facie* duties, then we would surely choose one of these. As Singer (1954: 586) explains, the case must be framed in such a way "so that it cannot be said that one has the alternative of refusing to speak at all". The circumstances may be such that our *only* two options are to tell a lie or to expose an innocent person to danger. However, even if we must answer the murderer's question, it might be argued that we are nevertheless able to answer truthfully without jeopardising the victim's safety.

Jonathan Adler (1997) provides an example of how this may be done. He supposes that the murderer's intended victim "spends a good deal of time... at the local diner, the Nevada" (Adler, 1997: 437). When the murderer asks where the victim is, we may reply, truthfully, that the victim has "been hanging around the Nevada a lot" (Adler, 1997: 438), thereby leading the murderer to believe that the victim might be at the Nevada at that time. Adler explains that, since the murderer thinks that we are co-operating with him and that our response is intended to be an answer to his question, he will be likely to act on the information we have given him and seek out his victim at the Nevada. In this way, we may save our friend's life without violating our perfect duty not to lie – our response to the murderer's question is true, so we have not lied.

On the face of it, this looks like it could be a plausible way of dealing with the dilemma of the murderer at the door. We prevent the death of our friend, and, as Adler (1997: 438) explains, we do not lie to the murderer – instead, "he will take [us] as having conversationally implicated that [the victim] is at the Nevada now." Since the murderer draws his own conclusions from our *true* statement, it may be argued that we do not commit the "fundamental wrong" (Adler, 1997: 439) of lying to him.

The murderer is free to ask further questions if he is not satisfied with our answer, and perhaps he could formulate a question that is so specific that we cannot escape either lying to him outright (by making a false statement) or giving away the whereabouts of the victim. However, if this strategy succeeds in misleading the murderer, it seems as though there is "an admirable alternative to lying" that "is to be recommended" (Adler, 1997: 439) in dealing with non-ideal circumstances. We can escape lying by "formulating such devious assertions" (Adler, 1997: 438) – all we have to do is think of an assertion that is true, but that is likely to mislead the person who hears it.

Kant himself takes such a strategy into consideration. He points out that "[o]ne could merely seem to give an answer" (MM, 6: 431, 554) to a question where a truthful response might harm someone – in Kant's example, if an author asks his audience whether they like his work, they might reply by "joking about the impropriety of such a question" (MM, 6: 431, 554) instead of admitting that they dislike it and offending

him. From my answer to the murderer he "may deduce what I want him to" and, even if his conclusion is false, "I have told him no lie" (LE, 27: 446-7, 202).

There are two important objections to this strategy, however. The first is that it is perhaps not as likely to succeed as an outright lie. As Adler (1997: 438) points out, there "is less certainty that the murderer will be misled" since the response is not a direct answer to his question. When I say that the victim is often to be found at the Nevada diner, he might ask me "Are you sure she is there now?" requiring me to make yet another assertion. This questioning process might conceivably last quite a long time, and the murderer could eventually come up with a question that is impossible to answer in this way – "no claim is made that there are always feasible constructions" to handle every possible question that we may be asked (Adler, 1997: 438).

As Kant points out, we might not be able to think of an answer that would both be true and satisfy the questioner: "who has his wit always ready?" (MM, 6: 431, 554). The "slightest hesitation in answering" (MM, 6: 431, 554) could cause the murderer to become suspicious and this would cause our strategy to fail. Furthermore, the murderer might force me to answer only "yes" or "no" by threatening to kill me if I give any other answer. The circumstances may be such that an outright lie is the only way of foiling the murderer's evil intentions, and if a true, but deceptive answer is unlikely to succeed, then this way of escaping the moral dilemma is not available to us.

However, even if a misleading answer is an effective means of preventing the murder, there is another, more serious objection to this strategy. This is that we will need to provide some account of why the misleading answer is a morally preferable way of dealing with the murderer than lying to him, and it is far from obvious that this is so. The only relevant difference between the lie and the deceptive truth is that the latter statement is true, while the former is false. It may, however, be argued that the intention behind each of the statements is the same – that is, to deceive the murderer – and that it is this *intention* of deceiving someone that is morally abhorrent, on the FH, since any act of deception prevents him from freely choosing his own ends and,

therefore, disrespects his humanity. If this is so, then the misleading truth is no better than the lie.

This line of argument may be developed by using an example from Siegler (1966), which again shows that telling a lie does not necessarily require the statement that is made to be false. He considers Jean-Paul Sartre's story *The Wall*, in which a character called Pablo is asked by the authorities where his friend, Ramon Gris, is hiding. Pablo does not wish Gris to be found, so he answers that his friend is in the cemetery, while being "quite certain that [he is] hiding somewhere else" (Siegler, 1966: 130). This is clearly "an attempt to deceive the authorities" since Pablo tells them "what he believe[s] to be false" in the hope that they will act on it and fail to find Gris (Siegler, 1966: 130). Thus it appears as though Pablo lies to the authorities in this story.

However, Ramon Gris is *really* hiding in the cemetery, quite "by chance, and completely unknown to Pablo" (Siegler, 1966: 130). This means that the statement that Pablo makes is *true*, even though he intends it to deceive the authorities. If the falsity of the statement made is a necessary condition for telling a lie (contrary to what I have argued earlier), it seems as though Pablo does not lie to the authorities after all, since he tells them the truth. Yet, as Siegler (1966: 130-1) points out, "[i]f Pablo were asked immediately afterward whether he had told a lie, and he were to answer honestly, he would say that he had". Furthermore, if people regularly made statements to others that they believed to be false, "[s]uch behavio[u]r would be excellent grounds for mistrusting [their] honesty" since "it is just that sort of thing that liars do" (Siegler, 1966: 130), even if some of those statements turned out to be true afterwards.

In this example, therefore, we might plausibly say that someone can tell a lie even if the statement that he makes is true. Kant (RL, 427, 65) considers a similar case, where we tell the murderer that our friend is not at home while she "has gone out... with the result that" the murderer encounters her and kills her. Even though this statement is true, Kant (RL, 427, 65) thinks that we have "told a lie" by saying that

she was "not in the house" and that we may be "justly accused as having caused [her] death".

This statement may reasonably be regarded as a lie as it contains the intention to deceive, which is "necessarily involved in lying" (Siegler, 1966: 130). For Kant, it is our intent to deceive others and thereby prevent them from attaining their own ends that is impermissible, and even a true statement may be made with the intent to deceive, as the preceding examples show. Deceptive statements "endeavo[u]r to contribute causally toward [another person's] believing" a particular proposition, according to Roderick Chisholm and Thomas Feehan (1977: 146), and, when we make an utterance that causes someone to acquire a belief that we believe to false, we deliberately interfere with his capacity to make free choices, since we try to impose a false belief on him. Heimo Hofmeister (1972: 264) makes this point clearly: "falsifications of facts are not immoral by themselves, but only by being apprehended... in a certain way" – thus it is not the falsehood of the proposition that makes it morally wrong, but the effect it has (or is intended to have) on the autonomy of the victim of the lie.

From this analysis it is reasonable to claim that a misleading truth is not an acceptable alternative to an outright lie after all, since it also involves deception. Thus the strategy of telling a deceptive truth is not morally preferable to that of lying outright – they both violate the prescription of the FH never to treat others as a mere means, and there does not seem to be a relevant moral difference between them. In genuine moral dilemmas, however, we may be permitted to deceive others (by lying or otherwise) simply because the conditions are non-ideal; here, we have no option but to violate some or other *prima facie* duty. This means that, although the FH prescribes absolute honesty under ideal conditions, it does not forbid lying under non-ideal circumstances. I will now turn to the third and final formulation of the categorical imperative and show that it allows us to derive similar conclusions.

V. The FKE and Lying

Most commentators focus on one of the first two formulations of the categorical imperative when they criticise or defend Kant's views on lying (Wood, 2008).

However, as I have explained previously, it is also possible to argue that lying is impermissible based on the FKE. Legislators making laws for a kingdom where everyone shares ends and where everyone is a co-legislator would outlaw lying, because lies make shared ends – the foundation on which the kingdom of ends is based – impossible.

In this section I will argue that the FKE does not lead to an absolute prohibition on lying, for two main reasons. First, as I will show, there are some cases in which the universal co-legislators would make laws permitting (or even requiring) the members of the kingdom of ends to lie. Second, I will claim (with Schapiro, 2003; 2006) that the murderer's evil intentions disqualify him from membership of the kingdom of ends, which means that we are entitled to act differently towards him than towards other people. I will consider Schapiro's argument in detail and try to defend it against an important objection. From these two lines of argument I will conclude that we may, indeed, lie to the murderer at the door, even in the kingdom of ends.

As noted in the previous section, Herman (1993) suggests taking a Rawlsian approach to the categorical imperative by employing his veil of ignorance in moral judgements. This means that the universal co-legislators should consider the position of the worst-off person in the kingdom of ends when deciding what laws to pass – in other words, they should "direct[-] [their] attention to the worst that can happen under any proposed course of action" (Rawls, 1971: 154) and make laws that secure the best possible outcome for that person.

It seems as though the Rawlsian "original position" is quite similar to the conception of the kingdom of ends. Rawls (1971: 141) claims that the "right course of action", or the right laws to pass, are those that "best advance[-] social aims as these would be formulated by reflective agreement given that the parties... are moved by a benevolent concern for one another's interests". This description of social aims and mutual benevolence seems to match what Kant has in mind with his kingdom of ends, where rational co-legislators must make laws for the good of the society as a whole. It is therefore a reasonable strategy to consider the laws that the Kantian legislators would make in terms of the original position.

As discussed in the preceding section, it seems as though the co-legislators would not pass a law that requires strict truthfulness at all times. The mere possibility of a situation where a lie could save a life or prevent harm would be sufficient to permit it, on Rawlsian grounds, because the intended victim of the murder clearly stands to lose far more than the victim of the lie. It is surely reasonable to suggest that the co-legislators would agree, after reflection, that lying should be allowed in the kingdom of ends in such extreme cases, since they would wish the lie to be told if they were in the position of the murder victim. Certainly, killing someone has a far more detrimental effect on her ability to choose her own ends than lying to her, since it constitutes a permanent frustration of all her current and possible future ends, while the lie only frustrates one end. Faced with this choice, one might think that the co-legislators would prefer a law that allows the lie to one that requires a rigorous adherence to the truth.

It is possible to think of some other cases of lying that might rationally be permitted by the universal co-legislators. Kant (AP, 180, 73) points to the "harmless lying that is *always* met with in children and *now* and *then* in adults"; that is, the tendency to embellish one's stories with exaggerations or invented facts in order to make them more interesting, and, as I will argue, this type of lie would not necessarily be forbidden on the FKE.

It is easy to think of an example of Kant's "harmless lying". Suppose someone who, disturbed by some felines caterwauling outside her bedroom at night, makes the following statement (S) to her colleagues the next morning: "Last night there were dozens of cats outside my window; they made the worst racket you can think of, all night long, and I didn't get a wink of sleep."

Clearly, S is a lie, as it contains at least four blatant untruths. First, there were surely not *dozens* of cats outside – at most there would have been two or three (and not *at least* twenty-four, as the word "dozens" implies). Second, the sound of a few cats fighting is certainly not *the worst racket* that the speaker's colleagues *can think of.* Most people would be able to think of sounds that are far more displeasing to the ear.

Third, the noise certainly did not last *all night long* – even the most energetic cats are unlikely to meow incessantly for an entire night without respite. And fourth, it is improbable that the speaker had no sleep whatsoever – perhaps she was able to sleep for one or two hours when the noise died down. Yet, even though S is a lie, it nevertheless appears to be "harmless", and it is difficult to see how the co-legislators would forbid it, as I will explain.

The law in the kingdom of ends is supposed to be universally binding on all its members at all times. As Campbell Garnett (1964: 299) points out, a moral rule is a "rule of general applicability that is required for the welfare of the interacting group as a whole." If the co-legislators make a generally applicable law that outlaws statements such as S, they would have to say that exaggerations are always morally wrong. This is unsatisfactory, for at least two reasons.

The first reason is that nobody is deceived by a statement such as S – it is, indeed, harmless. When the speaker says S, we commonly draw the (true) conclusion that there were some cats that made a noise and disturbed her sleep, even though none of the facts she asserts in S are strictly true. She also does not have an intention to deceive us – instead, she is simply trying to make her story more interesting to the hearer. Nobody's ends depend on S, so she is not frustrating anyone's goals. However, if the law in the kingdom of ends is formulated so that it forbids *all* forms of lying, we have to say that the speaker has done something morally wrong; that she has harmed humanity or prevented social co-operation; and that she deserves censure and reproach for her actions. It looks as though a law that leads to these conclusions is unreasonably strict.

The second reason is that the universal co-legislators could rationally *choose* laws that permit statements such as S to be made instead of laws that permit only entirely true statements such as "I didn't sleep well last night because of some noisy cats" (S'). If we apply the veil of ignorance to this situation, we may argue that the scenario where the speaker says S' causes the person who is worst off to hear a story that is significantly less interesting and enjoyable than what he would hear if the speaker

says S. Thus, laws permitting S are, *ceteris paribus*, preferable to those do not, since S causes the greater degree of pleasure.

Furthermore, in the kingdom where exaggeration is forbidden, people would have to take the utmost care in formulating statements. The prohibition on lying would make them count their words carefully to avoid any inaccuracies, since these would be morally impermissible. This is not the case when statements such as S are allowed, and narrators of stories are allowed some poetic licence (which increases the satisfaction of those who listen to the stories). Perhaps the co-legislators would, therefore, choose laws that permit lies that make stories more entertaining.

This argument may be extended to other types of lies, such as those that are told to enable surprise parties or to placate children – these lies seem to be at least permissible, if not desirable. Players of card games such as poker often engage in deception within the game, and it may be argued that permitting deception during certain games greatly enhances the enjoyment of the players. Similarly, one might contend, with Alan Strudler (1995) that deception is an essential part of negotiation and that, under certain circumstances, it is an important "device that... people... can use to work their way to a reasonable and mutually advantageous agreement" in commerce.

Even if one does not find all these arguments convincing, it seems reasonable to say that there are some situations in which lying would usually render the worst-off person better off than absolute truthfulness. The example of the murderer at the door is a case in point. This means that a kingdom where certain harmless lies and lies told to prevent harm are permitted could be said to be superior to one where lies are absolutely forbidden. A blanket ban on lying therefore seems to be undesirable in the kingdom of ends, as the lawmakers need to take these extreme cases in account before reaching reflective agreement. Thus, it is plausible to claim that the co-legislators in the kingdom of ends would perhaps allow certain kinds of lies in their realm, and these types of lies are, therefore, consistent with the FKE.

There is one important objection to this analysis, however. One might argue, rightly, that the Rawlsian method of maximising the position of the worst-off person is a consequentialist one, since it looks only at the likely effects of the lies. This is inconsistent with the Kantian view that lies are impermissible *independent* of their consequences – there is something inherently wrong about lying, and the results of an act of lying are irrelevant to their wrongness. Thus we might contend that the universal co-legislators would forbid lying because it is *always* wrong *in itself*, despite the existence of certain cases where lying has better consequences than truthfulness.

However, there is a second line of reasoning that escapes this objection. Even if the universal legislators would not choose laws that permit these lies, it may still be argued that the FKE does not forbid us from lying to people who have evil purposes. This is because such people show, by their actions, that they are unwilling to be members of the kingdom of ends. We might claim that the moral rules bind only those people who are part of the realm, and that they do not apply when we deal with those who have chosen to leave it of their own volition, such as the murderer at the door. This is the argument developed by Schapiro (2003; 2006).

To understand Schapiro's view, we have to refer to Rawls' (1955: 23-4) emphasis on the distinction between "rules defining a practice" and "general" or "summary" rules, which Schapiro (2003: 334) calls "rules of thumb". Summary rules are "generali[s]ed reports of the results of applying some or other rule directly to the cases at hand" (Schapiro, 2003: 334). This means that, when we apply a summary rule, we do not consider the features of the individual case – instead, we simply use a rule of thumb that "by and large... will give the correct decision" (Rawls, 1955: 23) in moral deliberations.

This point may be clarified by means of an example. Usually, the rule to be truthful "may be relied upon to express the correct" moral judgement (Rawls, 1955: 23) since there are many cases in which honesty seems to be, intuitively, the right thing to do. The extreme cases in which it looks as though we might be better off telling a lie are comparatively rare. Therefore, as Rawls (1955: 23) points out, we "would be justified in urging [the] adoption" of being honest at all times "as a general rule". If we follow

this rule of thumb, we are likely to make many more correct moral decisions than wrong ones, so we are entitled to adopt it as a rule of conduct, even though it may be "laid aside in extraordinary cases where there is no assurance that the generali[s]ation will hold and the case must... be treated on its merits" (Rawls, 1955: 24).

Practice rules, on the other hand, "define procedures compliance with which constitutes participation in some new activity" (Schapiro, 2003: 334). In other words, practice rules are set up to govern a particular institution or activity and, if someone wishes to participate in that activity, he or she must follow the rules. Rawls (1955: 25) explains the point clearly by using a sporting analogy – the game of baseball is governed by certain rules that describe our actions, and we may perform actions, *within baseball*, such as "[s]triking out, stealing a base" and so on.

These actions are specific to the game of baseball because, even though one may perform actions such as swinging a baseball bat outside a game, one cannot "be described as... striking out... unless [one] could also be described as playing baseball" (Rawls, 1955: 25). Thus only the practice rules properly describe our behaviour within the boundaries of that practice (Schapiro, 2006). Moreover, the rules of baseball define what one is legally allowed to do in a game – non-compliance with these rules (for example, throwing the bat instead of the ball) means that we are not playing baseball. As Rawls (1955: 26) explains, "[i]f one wants to do an action which a certain practice specifies then there is no way to do it except to follow the rules which define it." On Schapiro's (2003: 336) view, "the notion of a practice rule can be invoked... to build the right kind of flexibility into a two-level Kantian theory", making it permissible to lie to the murderer at the door.

This can be done by regarding the kingdom of ends as a practice, or an institution, that has been established for the good of its members. The universal co-legislators all agree on what the rules of the kingdom are, and these rules govern and describe the actions of all those participating in the kingdom. On Rawls' analysis, one needs to comply with the rules of the kingdom in order to participate in the practice that they govern. It may be argued that the murderer at the door has, by his actions, revealed

himself as unwilling to participate in the practice and that this may entitle us to lie to him.

Schapiro (2003: 339) argues that, when someone fails to comply with the practice rules, the practice of respecting humanity becomes a "sham". This occurs when "the non-compliant party... tacitly or implicitly claim[s] the protections and prerogatives attached to his role" within the practice "while at the same time failing to live up to its demands" which renders any attempt by others to follow the practice rules futile and meaningless. It seems as though the murderer may be regarded as such a non-compliant party, and that his actions cause the practice of respecting humanity to become a sham, as I shall explain.

When the murderer enquires about our friend's whereabouts, he is indeed "implicitly claiming the prerogatives attached to his role" within the kingdom of ends. He seeks the truth from us and expects us to be honest, and he is entitled, as a participant in the kingdom, to receive truthful answers to his questions as honesty is a practice rule governing expected and appropriate behaviour within the kingdom of ends. However, he "fails to live up to the demands" of the practice because he intends to use the information to break one of the rules of the kingdom (that is, not to kill any of its members).

Thus he expects me to be governed by the very same practice rules that he intends to break. He wishes to disrespect the humanity of his victim, while expecting me to respect his humanity. When I tell the truth to him, this act of honesty "is no longer what it ought to be" (Schapiro, 2003: 339) because the murderer is no longer a participant in the practice, and he has turned the practice of respecting humanity into a sham. In essence, the murderer "abandon[s] [his] role" as a member of the kingdom and the ordinary rules governing the practice therefore do not apply to our dealings with him (Schapiro, 2003: 340).

The analogy with baseball may serve to clarify this point. If I am a pitcher in a baseball game, I am ordinarily expected to throw pitches at the batter. However, if a streaker who wishes to disrupt the match runs over and stands on the batter's plate, it

is futile for me to throw a pitch at her and expect her to hit it, since she is simply not playing the game. This would certainly be a strange way of dealing with a streaker – she is not a participant in the game and should not be treated as such.

Of course, the streaker was not a participant in the game to begin with, and might therefore not be subject to any of its rules, but we may modify the example by supposing that a batter from the opposing team chooses to campaign for some or other cause by stripping off all his clothes on the batter's plate. Once again, throwing a pitch at him seems to be futile, since he indicates that he does not wish to play baseball. As Schapiro (2003: 340) explains, "because you are a participant [in the practice] you have to play... and the only way to play is to play by the rules".

It seems reasonable to assert that we are justified in breaking the rules of a practice in order to deal with someone who has already done so or who intends to do so. Even though the rules of football ordinarily forbid picking up the ball and throwing it at the referee, we may be justified in doing so if he abandons his role as referee and instead runs towards a fan to attack her. We will not be committing a football foul if we stop the attack in this way, since, at that time, we are simply not playing football. Furthermore, we are breaking a rule of football only because the referee's actions have caused the game to stop – he has turned the football match into a sham and it is futile to continue playing under these circumstances.

This seems to be a promising line of argument as it accounts for the view that we may lie to the murderer because he is not entitled to the truth. As Schapiro (2003: 343) explains, "[w]hat is essential to the practice is that it is a system of social cooperation". Clearly, the murderer is not willing to promote the ends of social cooperation – his choice of an evil end "makes it appropriate for [us] to regard him as having refused to participate in the shared activity of which honesty is a part" (Schapiro, 2006: 52). In this case, "it is impossible... to be honest... in the spirit proper to honesty" (Schapiro, 2006: 52) as the murderer's ends conflict with the reciprocity presupposed by the kingdom of ends. On these grounds, we are justified in lying to him.

Schapiro's argument has the further advantage in that we can extend the analysis to cases involving what she calls "paternalistic deception" (Schapiro, 2006: 52); that is, lying to people who are incapable of making rational choices. We might think that it is permissible to lie to someone who has a psychological disorder, or is clinically insane, if this would prevent her from making a very bad choice. People who are not fully rational are simply unable to participate in the institution of a kingdom of ends, "due to conditions of disease or immaturity" (Schapiro, 2006: 52) and we are not bound by the usual practice rules governing the kingdom in our dealings with such individuals. As Schapiro (2006: 52) explains, we may be justified in breaking the rules when dealing with people who "either 'can't' or 'won't'" participate in the kingdom of ends. In these cases, lies are not inconsistent with the FKE, since the rules made by the universal lawmakers apply only to those who are willing participants in the kingdom and not to those who "refuse[-] to play the colegislation game" (Schapiro, 2006: 52).

This is an attractive conclusion since it accounts for "both the rigidity and the flexibility of moral rules" (Schapiro, 2006: 32) – it explains why universal co-legislators would permit exceptions to their otherwise binding prohibition on lying, since the rules are rigid only for those who adhere to the practice in question. Moreover, this account cannot be accused of being a consequentialist one. The lawmakers choose the rules because they consider the actions they forbid to be wrong intrinsically, not because of their possible consequences, and the explanation of when and why we may depart from the rules is coherent and reasonable. However, this argument is vulnerable to a seemingly telling objection, that is, that it proves too much, as I shall now explain.

The conclusion of Schapiro's analysis is that we are permitted to break the otherwise binding moral rules when dealing with someone who is no longer a participant in the kingdom of ends. This explains why lying to him is not inconsistent with the FKE. However, this seems to entail that we would be justified in taking *any* action against the murderer (even killing, maiming or torturing him) if this would prevent the crime, since his voluntary non-participation in the kingdom of ends renders us free to treat him in any way we like. This objection poses a significant challenge to Schapiro's

argument – if we can condone lying to the murderer, why can we not use the same considerations to justify torturing or killing him? If we accept Schapiro's conclusions, then we appear to be on a slippery slope indeed, as ordinary morality does not seem to apply to the murderer at the door.

I think that Schapiro could adequately respond to this objection by appealing to the degree to which we are allowed to break the rules in order to deal with the murderer. It is true that telling a lie is immoral, but the effect of lying is far less harmful to the murderer than killing or torturing him would be, even if these methods would serve the same purpose of preventing the crime. We might be able to argue that, since the moral rules have intrinsic value, and that violating them is always regrettable, we should choose the course of action that *least* harms the person whose actions and intentions leave us with no choice but to violate the rules, even if he has chosen not to participate in the kingdom of ends.

If we are faced with the choice of either lying to the murderer or killing him to try to save the victim's life, then, it seems as though we have some good reasons for opting to lie. I will argue that, while the lie is a permissible way of dealing with the situation, murdering him is excessive and unwarranted in this situation. This is because, while a threatening situation "allows [us] to use a variety of means" in our defence (Strudler, 1995: 811), we should nonetheless choose the method that causes minimal harm to our assailant. I will defend this position by considering the two possible ways of protecting our friend against the murderer – lying to him or killing him - in turn, assuming that both of them are guaranteed to succeed and that, therefore, we have no reason to choose one over the other on the basis of its efficacy.

If we lie to the murderer, it is clear that we disrespect his autonomy and freedom of choice, on Kant's view. This is a deviation from the ideal moral laws made by the universal co-legislators in the kingdom of ends, and we should therefore not tell a lie under ordinary circumstances. However, since the murderer intends to use our honest answer to violate the credo of mutual respect prescribed by the FKE, we are apparently justified in using the lie in self-defence. Here, no force is involved and the effect on the murderer's ends is minimal – all we do is frustrate *one* of his purposes.

We do not prevent him from attaining any of his other goals, such as earning a living (assuming he is not a hit-man who will be paid if he kills our friend), developing his talents, taking care of his family or keeping his promises. All we do is thwart his plan to kill his victim.

Murdering him, on the other hand, clearly constitutes a far greater interference with his ability to set his own goals. Since we end his life, we prevent him from attaining *all* of his aims. We contribute towards his inability to fulfil any of his other *prima facie* duties, and it seems reasonable to assume that he intends to fulfil some of these – perhaps he wishes to take care of his family or develop his talents. It may, therefore, be argued that the murderer intends to break only *one* of the practice rules defining the kingdom of ends – that is, the only rule he intends to break is the decree not to kill. We may, therefore, break the rules in order to prevent him from achieving this aim; but it seems unreasonable to say that we should act so that he is no longer able to achieve *any* of his goals.

Strudler (1995: 811) makes a similar point when he notes that "[i]t is excessive to kill somebody to prevent him from stealing carrots from your garden". It hardly seems warranted to use deadly force in such a situation if there are other means to stop the theft. Kant's moral theory assigns a great deal of importance to autonomy and it seems reasonable to think that, since autonomy is so valuable, we should try to minimise our interference with the goals of others. It is always regrettable when someone's autonomy is violated, even if they have chosen to stop participating in the kingdom of ends, and we should, therefore, find ways of dealing with them that respects their autonomy to the greatest degree possible. Since the lie prevents only *one* free choice, while the murder prevents *all* future choices, the lie is clearly preferable. This seems to be a plausible way of responding to the objection that Schapiro's analysis proves too much – even though we are free to choose from a number of methods of defending ourselves against evil-doers, we should nevertheless try to respect their autonomy insofar as this is possible.

However, it is possible to conceive of a situation where killing the murderer and interfering permanently with his autonomy would, after all, be justified. Suppose that, instead of finding the murderer on our doorstep, we come home and encounter him about to kill his victim. Perhaps he has a knife at her throat and it is clear that he intends to use it. If we have a gun and can protect an innocent life by shooting the murderer, it seems as though we may be justified in doing so. It seems reasonable to think that we may use even deadly force or perform certain otherwise impermissible acts in cases such as this one, or in self-defence, when this is the only way to save a victim's life.

This case differs from Kant's case in the important respect that killing the murderer appears to be the *only* way in which to save the victim's life. Clearly, a lie would be of little use in this situation – I cannot protect my friend by telling the murderer a blatant untruth, such as "Paris is the capital of Germany". Of course, in certain circumstances a lie might prevent the murder; for example, if I tell the murderer that I just won a million rand in the lottery and will give it to him if he drops the knife and lets my friend go, he might reconsider. However, if shooting the murderer is truly the only way to prevent the death of his intended victim, it seems as though we may be entitled to do it (although one might argue that we should try to injure him instead of killing him if this is at all possible).

We might, therefore, have a response to the objection that Schapiro's solution leads to a slippery slope if we claim that different types of non-ideal conditions require varying degrees of interference with the autonomy of those who have chosen to break the rules of the kingdom of ends. Any way of dealing with these circumstances will involve violating someone else's autonomy to a certain extent, but we should, where possible, minimise this violation. If the lie is all that is required to prevent a murder, we choose it over shooting the perpetrator because it is a one-time interference with his autonomy over a permanent one. Similarly, if we can stop the murderer by shooting at his arm instead of at his head, we should choose the arm, since that action interferes only with the choices he could make regarding the use of a limb instead of *all* of his future choices. Of course, any interference with someone else's freedom of choice is deplorable, but in non-ideal circumstances we have no option but to do so.

This response gives us a way of working out precisely what we are entitled to do to a wrong-doer who breaks the rules of the kingdom of ends – perhaps there is some rank-order of levels of interference with other people's autonomy. Every possible response could be evaluated in terms of the degree to which it disrespects someone else. Killing someone ranks near the top of the list, since death renders all our ends impossible, whereas lying is probably somewhere near the bottom; and a once-off lie, which interferes with someone's autonomy only on a particular occasion, ranks lower than an elaborate ruse or charade designed to deceive a person over time (such as someone lying to her spouse about her infidelity for an extended period). There might be certain acts at the top of the list that disrespect autonomy to such an extent that they may *never* be done, under any circumstances, since there is always a preferable alternative – biological or nuclear warfare on an entire nation possibly meets this condition. On the other hand, some actions near the bottom of the list may be minimal violations of autonomy, such as standing in a doorway for a short period of time to prevent someone from entering it freely.

This could mean that interference with autonomy is a matter of degree, and different situations warrant different levels of intrusion. I feel that this is a plausible view to hold and we could perhaps appeal to such a rank-order in deciding what we are entitled to do in non-ideal conditions – some actions may be justified in certain circumstances, but not in others, and the rank-order may be helpful in determining what is permissible.

Thomas Nagel (1972: 141) provides some support for this view in claiming that there are some actions that can *never* be justified, even in warfare, such as "weapons designed to maim or torture or disfigure the opponent rather than merely to stop him". Nagel argues that, in warfare, we are entitled to perform actions that stop combatants in their capacity *as soldiers*, but not as human beings. If we use weapons that cause more harm than necessary to prevent a soldier from carrying out acts of war, we "attack the men, not the soldiers" (Nagel, 1972: 141) and this is not justifiable since it is extreme. We might extend this conclusion to the case of the murderer at the door as follows: if we lie to him, we stop him *as a murderer* because his plan is thwarted. He

can, however, still engage in a variety of free choices that are consistent with his rational nature as a human being and our lie does not attack him as a person.

Of course, it may be argued that killing the murderer *does* stop him as a human being since, as I have explained, all his future choices are rendered impossible. However, if this is the only way to prevent the murder of the victim, we may – even on Nagel's analysis - nevertheless be entitled to do so. Nagel (1972: 138) makes the key point that one may use deadly force against someone else but still treat her with respect: "to fire a machine gun at someone who is throwing hand grenades at your emplacement is to treat him as a human being". This is because the "attack is aimed specifically against the threat presented by a dangerous adversary" (Nagel, 1972: 138). Firing the machine gun is the appropriate response to someone who chooses to throw hand grenades simply because he could foresee that someone would respond to him in that way and because such a response is justified in meeting that threat. A soldier expects that his opponents will try to deal with his actions by force – in doing so, they are not disrespecting him as a human being, but reacting appropriately to him as a soldier.

In the same way, lying to the murderer at the door is a way of meeting his threat, as is shooting him when he holds a knife to his victim's throat. It may be argued that there is no disrespect to him as a person, but that we are merely dealing with him in the way we ordinarily deal with murderers – that is, to employ the means necessary to save the lives of their victims or to punish them for their wrongdoing. Just as a human being acquires special status as a soldier when she joins the army, so a human being acquires special status when he chooses to become a murderer. On Nagel's view, then, it seems as though we would not disrespect the murderer by dealing appropriately with him; we are justified in lying to him or killing him "on the basis of [his] immediate threat or harmfulness" (Nagel, 1972: 140). This view is not uncontroversial and it certainly needs more development, but at the very least it provides some support for the position that there are levels of interference with the autonomy of others.

A further consequence of this response is that we may be required to tell the murderer the truth if he asks us any other questions that are not relevant to his evil purposes,

since failure to do so would disrespect him as a human being because he is not asking those questions *qua* murderer. Suppose that he is satisfied when we tell him that our friend is not home, and then asks us where the nearest Italian restaurant is as he is hungry. Are we permitted to lie to him about its location? I contend that, if he seeks nothing further by asking this question than to find a place to eat dinner, we are bound to answer honestly. We have successfully thwarted his ends and this question is innocent – he may be said to be participating in the kingdom of ends once again.

We may return to the baseball analogy to support this conclusion. Suppose that the protesting batter has stripped naked on the batter's plate and made his point, gaining the attention needed to demonstrate for his cause. He now gets dressed, fetches his bat and is ready to play again. Should I pitch the ball to him? It seems as though the game can now restart, and his momentary departure from the rules governing the practice is over. Since we are once again playing baseball, it seems as though I should throw my pitch. In the same way, I should answer the murderer honestly, since he has abandoned his evil plan.

To summarise, then, the conclusion of this analysis is as follows: Members of the kingdom of ends are usually bound by its rules, although exceptions may be made when dealing with those who have voluntarily refused to participate in the kingdom. However, this does not mean that we may treat the non-participants in any way we like. We are entitled to break the rules in our dealings with them, but only insofar as their actions render the kingdom a sham. Furthermore, when faced with alternative ways of responding to their actions, we should try to minimise our interference with their autonomy. Therefore, even though the FKE usually prohibits lying, there are certain situations in which it is permissible.

In the preceding discussion I have tried to show that all three of the formulations of the categorical imperative permit lying under some circumstances. Some maxims requiring lies are universalisable according to the FUL; the FH allows lies in non-ideal conditions; and lying to non-participants in the kingdom of ends may be justified on the FKE. Kant's moral principle therefore does not seem at odds with our intuitions that lying is sometimes not morally wrong. So far, however, I have

assumed that the proper way to understand the categorical imperative is to see it as a moral decision-making rule that guides our actions – however, there is an alternative interpretation that regards it as a principle grounding an objective moral law that underlies all right actions. I will discuss this interpretation in the next section and consider whether or not it leads to the same conclusions.

VI. The Categorical Imperative as an Objective Moral Law

In this section, I consider the categorical imperative not as a normative rule for judging whether a particular action is morally permissible, but as the grounding of an objective moral law. On this view, the categorical imperative is not a *method of deciding* whether or not we may perform a given action, but rather as an *explanation* of why the action is morally permissible or not. As Philip Stratton-Lake (2000: 68) explains, the categorical imperative "may be understood as grounding particular obligations not in the sense that it justifies them, but in the sense that it acts as the condition of their possibility."

Jens Timmermann (2007: 112) points out that people generally "know full well what they ought to do if only they pay due attention to their own moral judgement"; they do not need to have read Kant's works in order to tell right from wrong. We therefore do not need to use the categorical imperative as a decision-making rule in practice. Instead, it grounds an objective standard that may be used to justify our moral judgements – we say actions are wrong *because* they do not correspond to any of the three formulations of the imperative. What all morally impermissible actions have in common is that they fail to meet the objective standard of the categorical imperative.

The difference between these two interpretations of Kant's moral principle can be seen clearly if we consider the murderer-at-the-door scenario. If the categorical imperative is a decision-making rule, we consider our action in terms of any (or all) of the three formulations of the principle. We conclude that the action is morally wrong if the maxim on which it depends cannot be universalised (violating the FUL); if it treats others as ends in themselves (violating the FH); or if it is incompatible with the laws that would be made by co-legislators in a kingdom of ends. On the other hand, if

the action passes these three tests, we may deduce that the action is permissible. Clearly, this procedure treats the categorical imperative as a test for whether an action is right or wrong.

If the categorical imperative is an underlying principle that makes right action possible, however, and if it grounds an objective moral law, then the analysis is quite different. To understand this view, we must first note that Kant thinks that moral laws (and specifically the categorical imperative) are known *a priori*, that is, independently of all experience. He points out that moral laws "hold as laws only insofar as they can be *seen* to have an *a priori* basis" (MM, 6: 215, 370) and that "reason commands how we are to act, even though no example of this could be found" (MM, 6: 216, 371). This means that all moral principles and duties "could be spelled out entirely independently of any empirical knowledge" (Wood, 2002: 2). From reason alone, we are able to derive moral judgements such as "It is wrong to lie", without drawing on any empirical facts.

Another key element of this interpretation is that it considers moral principles to be hierarchical. As Hanna (2009) explains, we need to distinguish three levels of principles. The highest level contains "absolute moral meta-principles, which are strictly and unconditionally universal *a priori* normative rules binding on all rational beings" (Hanna, 2009: 6) and the categorical imperative falls into this category. These meta-principles underlie all our moral principles and we are, therefore, required to obey the categorical imperative at all times, since it is the "law of morality" and an "absolute command" (G, 420, 29).

The second level in the hierarchy contains elements that Hanna (2009: 6) calls "first order *ceteris paribus* moral principles". These are objective moral principles that prescribe what we ought to do in ideal circumstances. Kant refers to these principles as "grounds of obligation" and explains that these are rules that an agent "prescribes to himself" (MM, 6: 224, 329). The tenet not to lie falls into this category – it is a principle that establishes how we should, in general, conduct ourselves.

The third and final category of the hierarchy of moral principles contains "*moral duties*, which are first-order moral principles that are also *agent-centred obligations*" (Hanna, 2009: 6). Moral duties are, simply put, what a particular agent ought to do in her specific set of circumstances. Grounds of obligation tell us what we should do, "other things being equal" (Hanna, 2009: 6), while moral duties tell us what we should do *in our specific situation*, where other things are not equal. Now that we have explained the key features of this interpretation of the categorical imperative, we may consider how it functions as an explanation for the permissibility or impermissibility of a particular action.

Instead of subjecting the action in question to a test involving the three formulations of the categorical imperative, we intuitively recognise the rightness or wrongness of the action on Hanna's interpretation. This is because the categorical imperative is known *a priori* – it "is an *a priori* synthetic practical proposition" (G, 420, 29). Because we know what our grounds of obligation are, we are able to determine the rightness or wrongness of an action for ourselves through reason alone. Having established that the action is morally impermissible, then, we are able to appeal to the categorical imperative as an explanation as to why it is wrong – it fails to conform to the objective moral standard that is grounded by this principle.

When we encounter the murderer at the door, on this view, we are immediately able to see that there are two conflicting grounds of obligation involved. On the one hand, if we tell the truth to the murderer and allow him to kill our friend, we violate the "first-order substantive *ceteris paribus* moral principle requiring us to benefit others and prevent harm to them" (Hanna, 2009: 37); on the other hand, lying to the murderer goes against the tenet to be truthful, which is also ordinarily a ground of obligation. In this case, then, it is clear that the two "grounds conflict with each other" (MM, 6: 224, 379), and this results in a moral dilemma.

It is because of this conflict between two first-order principles that the moral dilemma occurs. If the person at the door just asked us where the nearest Italian restaurant was, we might not even think twice about giving him an honest answer, because telling the truth would not violate any other ground of obligation. In this case, however, the fact

that the murderer has evil intentions causes us to question whether, in this specific instance, there is a reason to override a moral principle that we already know to be true *a priori*, and whether we are permitted to lie to him to save our friend's life.

We already see that it would be wrong to lie to the murderer under ordinary circumstances, since we know that lying is wrong – if it were not, there would be no moral dilemma to begin with and we would not have to think about what the best course of action is. Typically, we are able to judge quite easily what we should do – we do not spend hours deliberating over our actions when performing everyday acts since the grounds of obligation are able to guide us under normal circumstances. In the case of the murderer at the door, however, there are exceptional circumstances where the grounds of obligation conflict. Here, the moral dilemma arises precisely because we know that both alternatives are *prima facie* wrong, and we must make a choice.

How, then, ought we to deal with this moral dilemma? Hanna (2009: 24, 29) provides a possible solution by making use of three key principles, called the No-Global-Violation Constraint (NGV), the Excluded Middle Constraint (EMC) and the Lesser Evil Principle (LEP) respectively. I will explain these principles in turn before considering how Hanna proposes resolving the moral dilemma.

The NGV simply states that an act cannot be morally permissible if it violates any global moral principles; that is, if it goes against any of the absolute moral meta-principles at the top level of Hanna's hierarchy. From this principle, Hanna (2009: 24) distinguishes between "global" and "local" moral transgressions – a global transgression violates a meta-principle such as the categorical imperative and is "strictly" forbidden, while a local one goes against our grounds of obligation and may be permissible "in some act-contexts". Applying this analysis to Kant's theory, any act that violates the categorical imperative is impermissible (since the imperative is a meta-principle), while grounds of obligation such as the injunction not to lie may be violated in some situations.

The EMC claims that "if an agent has a moral duty in an act-context, then there is always one and only one moral duty for an agent in that act context" (Hanna, 2009: 24). In all situations where we have a moral duty, there is, therefore, only one action that properly fulfils that duty, "no matter how difficult it is for the agent herself to discern it" (Hanna, 2009: 24). This constraint is in line with Kant's discussion of duty, since he claims that "a collision of duties... is inconceivable" (MM, 6: 224, 379), which follows from the very concept of duty. Because a duty "express[es] the objective practical necessity of certain actions" (MM, 6: 224, 378), it follows that we either act in accordance with our duty or violate it in a given situation. There is an "excluded middle" since we cannot *partially* fulfil our duty – what we do is either right or wrong. The word "duty" *means* what we ought to do in a certain situation, and any agent "only ever has *one* moral duty" in a particular context (Hanna, 2009: 24). Furthermore, acting on any principle contrary to our duty "is morally impermissible in that context" (Hanna, 2009: 24).

The immediate consequence of the ECM is that any conflicts of *prima facie* duties are not real conflicts, since duty simply describes the right act in the particular context. This means that, when we encounter a moral dilemma such as the one involving the murderer at the door, the dilemma is only an "apparent or *prima facie*" one (Hanna, 2009: 25) since there *is* a right action that would not violate any meta-principles and be morally permissible on the LGV. The dilemma arises simply because we cannot discern what this action would be, because the grounds of obligation that would normally guide our actions are in conflict.

This situation is described by Hanna (2009: 29) as a "real local moral dilemma" since the conflict arises between grounds of obligation and not meta-principles. We have to violate one of our grounds of obligation, but this would be only a local transgression in the terminology of the EGV. This point, combined with the fact that there is only one duty that would be morally right, means we are able to act morally after all under the "desperately nonideal" conditions that characterise apparent moral dilemmas (Hanna, 2009: 2), and there is a third principle that we may appeal to in order to determine the right course of action.

This principle is the LEP, which Hanna (2009: 29) defines as follows: "Given a real local moral dilemma between first-order substantive *ceteris paribus* moral principles, you ought to choose [the one] which in that context is the lesser of several evils, in the sense that acting on it most keeps rational faith with the Categorical Imperative." In other words, when two of our grounds of obligation conflict, we should choose the *lesser evil*, that is, the action that best matches the directives of the categorical imperative, or which "*adequately expresses* it" (Hanna, 2009: 30). This seems to square with Kant's own solution to moral dilemmas, as he says that, when "two such grounds conflict with each other... the stronger *ground of obligation* prevails." Of course, we need to understand what is meant by "the stronger ground of obligation" and what it is to "keep rational faith" with the categorical imperative.

Hanna's position is that keeping rational faith with the categorical imperative means seeking to uphold its ideals. This means that we must consider our action as a local means towards a global end, and choose the action that aims to attain the ends prescribed by the categorical imperative, in any of its formulations.

Thus, when we consider what to do in the case of the murderer at the door, we should not ask ourselves which action can be universalised, or best treats others as ends in themselves, or would be chosen by a universal law-maker in a kingdom of ends: instead, we have to ask ourselves whether, through our action, we aim to uphold or betray the ideals of the categorical imperative (Hanna, 2009: 38). In situations of moral dilemmas, there is *some* evil in both alternatives – "each of your two options is an evil" (Hanna, 2009: 37). Since we cannot avoid doing evil, then, we must choose the *lesser* evil according to the LEP.

We can see what it means to uphold the ideals of the categorical imperative when considering the case of the murderer at the door. If we lie to him, our lie is a means to achieve the "end of preventing murder" (Hanna, 2009: 38). This squares with the categorical imperative's directive to prevent harm to others and, as it may be argued, our *end* in lying to the murderer is to prevent such harm, even if the *means* we use are evil. Telling the truth, on the other hand, "would be to accede to or condone the murderer's intention to harm and treat the victim as a mere thing", which is a far

greater evil than treating only the murderer as a mere thing without also harming someone else (Hanna, 2009: 38).

Since the situation is framed in such a way that these are our only two possible choices, we must, according to the LEP, choose the lesser evil – this means that it is our duty to lie to the murderer under these conditions because, in Hanna's (2009: 38) terminology, "lying in this context is not *globally* wrong; on the contrary, it is *locally* obligatory and only *ceteris paribus* wrong." Furthermore, the lie is not told in an attempt to violate the categorical imperative – instead, it is told "for the sake of the categorical imperative" (Hanna, 2009: 39), in an attempt to preserve one of its rules. Thus, even though the lie is regrettable, and even though we must "take full responsibility for" it (Hanna, 2009: 39), it is our best option under the circumstances and constitutes our duty under these conditions. Moreover it is consistent with the categorical imperative on the NGV, since it is only *ceteris paribus* wrong.

Hanna's analysis seems to be a plausible way of interpreting the categorical imperative, and it seems to square with the way in which we ordinarily view morality. We have certain intuitions about what is right and what is wrong, and these intuitions arise because moral principles "can be apprehended completely *a priori*" (LE, 27: 254, 49). Apparent moral dilemmas occur when these grounds of obligation are in conflict with one another. When we encounter these dilemmas, we usually try to choose the lesser evil, and we may sometimes regret the circumstances that force us to act in ways that we consider to be morally wrong.

For example, if someone chooses to run away from a burning building in order to save her own life and that of her child instead of helping others trapped inside, she will, in all likelihood, experience regret and might even be subject to Hanna's (2009: 39) undesirable consequences of "moral criticism, blame[-] or punishment" from others. It is the fact that she recognises the grounds of obligation that bind her, and feels their force, that she acknowledges that there would ordinarily be something morally wrong in allowing innocent people to die when one could have saved them – however, under the circumstances, she may still be said to be doing her duty by rescuing her baby. We usually experience pangs of guilt or regret when we violate a moral law, even if it

is the best we could do, and Hanna's analysis provides an explanation for this – the act remains wrong to some extent, but it is only locally wrong.

This interpretation of the categorical imperative leads to the conclusion that lying *is*, after all, always morally wrong – but only if the word "wrong" is interpreted to mean "locally wrong", in Hanna's terms. The categorical imperative does permit us to lie, but only when this is absolutely necessary in order to prevent a greater evil; that is, to prevent something that is globally wrong. There is no obvious contradiction in saying that, although it is always wrong to lie, there are cases in which we must nevertheless do so since the alternative involves an even greater wrong. According to Hanna, anyone who would condemn a person who lies to the murderer at the door (or in similar situations) would be a "moral idiot and a rule monger... a moral martinette or prig" (Hanna, 2009: 39) since he commits the "disastrous Flatlander Fallacy" (Hanna, 2009: 36) of mistaking a local wrong for a global one.

Hanna's interpretation, therefore, allows Kant to maintain that lying is always morally wrong without requiring us to tell the truth at all times. Thus Kant's ethical theory is not inconsistent with the view that some lies are justifiable, because moral dilemmas permit us to tell lies in exceptional cases (even though it is deplorable that we must do so, and even though the ground of obligation that underlies our duty to tell the truth is still binding). On this view, Kant simply draws the wrong conclusion in his essay about lying when he claims that we ought to tell the murderer at the door the truth. He merely gives priority to the *prima facie* duty to be truthful when he should prefer the principle of preventing harm to others, so that he is mistaken about which course of action best upholds the ideals of the categorical imperative. As Hanna (2009, 40) notes, it is possible to make "errors in moral judgment", and we may argue that this is exactly what Kant does.

Of course, one might contend that it is Hanna who errs, and that allowing our friend to be harmed is only locally wrong, while the lie is the greater evil in this case. This would mean that we *ought* to tell the murderer the truth, since, on the EMC, there is only one action that corresponds to our duty. Regardless of which conclusion is the correct one to draw, however, this interpretation of the categorical imperative

provides us with a solution to the problem that Kant's theory is too rigorous to be taken seriously.

If Hanna is correct to say that there is some action that corresponds to our duty in situations where two different grounds of obligation conflict with one another, it seems as though we need some way of determining what the right course of action is. There needs to be some sort of means of deciding when we are able to override grounds of obligation, particularly in difficult cases. The scenario of the murderer at the door does not appear to be a difficult case, as it seems clear that we should lie to him to prevent the murder and override our obligation of answering honestly, but it is obvious that there may be some situations where it is not so obvious what our correct course of action is. For instance, we might imagine a situation in which a security guard, trusted by her employer to watch over a vast amount of money in a vault, is tempted to steal some of it in order to pay for an expensive back operation that will make her physically disabled daughter more mobile.

Here, it seems as though it is less obvious what the right course of action is. The guard is faced with a conflict between two grounds of obligation, that is, between keeping her promise to her employer to guard the money and helping her sick daughter. How can she tell which of her two grounds of obligation is more binding? Which possible action best upholds the spirit of the categorical imperative? It seems as though there are good reasons to say that she should keep the money safe (since we are normally required to keep our promises to people who trust us), but one could also justify the theft, saying that the guard has an obligation to end her daughter's suffering.

It is beyond the scope of this paper to develop a full theory of how we may decide moral dilemmas such as these while maintaining the spirit of Kant's categorical imperative, but it is possible to make some tentative suggestions regarding what such a theory would entail. I propose that such a theory would need to refer primarily to autonomy and respect for persons in order to solve moral dilemmas, since these are the core values propounded in Kantian ethics. Specifically, we must evaluate the

extent to which our actions disrespect the autonomy of others in order to determine what our duty is when we are faced with a moral dilemma.

As I have argued previously, lying to the murderer is a lesser violation of Kant's commandment to promote autonomy than telling him the truth would be – lying thwarts only *one* of the murderer's goals, while honesty permits him to use his victim as a means to his own ends, thereby also permanently frustrating *all* her future projects. Thus it may be argued that lying is the action that best promotes the categorical imperative – *any* action violates autonomy to some degree, and we are therefore bound to choose the lesser evil. Since it is less disrespectful to lie to the murderer than to permit him to kill the victim, it is clear that we should lie. We might also appeal to autonomy in deciding exactly what we are allowed to do to the murderer. As I explained previously, it is plausible to argue that we may lie to him, block his path or push him down the stairs, but perhaps not cut off his arm or kill him – the last two responses represent a far greater degree of interference with his autonomy than the first two and are, therefore, impermissible. A fully developed theory of when we are entitled to override some or other grounds of obligation will certainly need to take autonomy into account if it is to be Kantian in nature, and it may be argued that these considerations get the right answer in the case of the murderer at the door.

In the case of the security guard, however, it is not at all obvious that one action better keeps faith with the categorical imperative than the other. Stealing the money would be disrespectful to the employer and to the company as a whole, as well as possibly undermining the institution of promise-making (just as making a lying promise to get money would). The theft also interferes with the employer's autonomy in using the money for whatever purposes he chooses. Not taking it, on the other hand, means that many of the employee's daughter's goals continue to be frustrated, since the disability does not allow her to do several things that she would otherwise choose to do. It seems as though it is not so easy to determine what should be done under these conditions, and the solution does not present itself as easily as in the example of the murderer at the door.

A further point worth noting is that the grounds of obligation remain binding even if we have a duty to override them on particular occasions. For example, if a moral dilemma requires me to lie on one occasion in order to keep faith with the categorical imperative, I am still required to tell the truth in future interactions with people where there is no conflict between the grounds of obligation. Moreover, I will have to acknowledge that I have done *something* wrong in violating the autonomy of the person to whom I told the lie. The grounds of obligation requiring me to tell the truth still stand, but I have no choice but to violate them in the presence of a moral dilemma.

Furthermore, a moral theory that keeps faith with the categorical imperative should not look at the consequences of a particular action, since, on Kant's analysis, the action is intrinsically right (or wrong) independently of its consequences. As Nagel (1972: 124) points out, an absolutist position like Kant's does not "give[-] primacy to a concern with what will *happen*", but rather "to a concern with what one is *doing*." Thus we cannot simply say that stealing the money is likely to have better consequences than keeping the promise, since Kant's theory is *not* a consequentialist one; instead, we must examine the action itself and the degree to which it violates someone's autonomy, irrespective of the results of the action. In a genuine moral dilemma, then, we must provide some reason for saying that one action is preferable to the other, and that one ground of obligation may be overridden more readily than another, without any appeal to its possible consequences.

Considering a conflict between two grounds of obligation may be seen to be a consequentialist method of deciding what to do in a moral dilemma. For instance, if I am faced with the choice of lying to the murderer or telling the truth to endanger my friend, it may be argued that I appeal to consequences when I choose to lie. I foresee that the consequences of telling the truth are less desirable than those of lying, so I choose the latter option.

Properly understood, however, the consideration of what should be done when there is a moral dilemma regards the action itself instead of its consequences. Instead of asking ourselves, "What are the consequences of lying to the murderer?" we should

ask “To what extent do I disrespect the autonomy of the murderer if I lie to him?” This means that we do not consider the possible future effects of our actions in deciding what to do, but rather the degree to which they violate Kant’s tenet that we should respect humanity. Even if lying to the murderer has tragic results, we might still say that we make the right choice in doing so, because we do the best we can to respect our friend’s humanity (despite violating the autonomy of the murderer). The grounds of obligation are valid regardless of consequences, since they are part of an objective moral law, and we need some way of determining when they may be overridden without appealing to what will happen.

Clearly, an explanation for why certain grounds of obligation may be overridden will have to appeal to autonomy, and there is a great deal of scope for further research regarding how we are to determine our duty in these cases. For the purposes of this paper, however, it seems reasonable to say that there are at least some cases where lying could plausibly be said to uphold the principles of the categorical imperative better than other courses of action, and that we may sometimes override our obligation to tell the truth for the sake of autonomy.

VII. Conclusion

In this paper I have examined the three formulations of Kant’s categorical imperative and applied them to the case of lying to the murderer at the door (as well as to other cases of lies that we commonly regard to be morally permissible). On the FUL, it seems as though we may justify some instances of lying, since the maxims on which they are based can plausibly be seen to pass the CC and CW tests. We also see that it is essential to specify the maxim with exactly the right level of generality, to avoid the problems of false positives and false negatives, and that a maxim such as “I will always tell the truth” is much too general. Thus we have to consider more specific maxims that include features that are relevant to the case at hand. It seems reasonable to claim that the maxim of lying to save someone’s life meets these conditions; and it appears to be universalisable, since its opposite maxim (telling the truth to expose someone else to unnecessary danger) is not. Thus the FUL does not entail that we may never tell lies.

The FH abhors lying because it treats other people as mere means to ends that they cannot share, since they are deceived as to what these ends are, thereby disrespecting their autonomy. However, the FH should be taken to apply only under ideal conditions. When there are mitigating circumstances – that is, when someone intends to use *us* as a mere means to *her* ends, we may be permitted to lie in order to defend ourselves, even though this involves violating the autonomy of the person we lie to. The strategy of deceiving someone without lying outright, by telling a misleading truth to her instead of a blatant falsehood, does not allow us to escape disrespecting her autonomy, since it is our intention to deceive her that is objectionable and not the truth or falsehood of the statement we make. Under non-ideal conditions, though, lying may be an acceptable means of self-defence and, on the FH, it is not always morally impermissible.

In the kingdom of ends, it is plausible to suggest that there are certain lies – such as Kant's "harmless lies" or lies to save a life – that may be chosen by the lawmakers to apply to their realm. Even if the universal co-legislators would outlaw all forms of lies, though, we may be permitted to lie to people who have, by their actions, shown themselves to be unable or unwilling to participate in the practice that is the kingdom of ends. Since practice rules apply only to those who contribute to that practice, we are entitled to break the rules when dealing with non-participants, as the sporting analogies show. However, there are limits as to how we may treat these non-participants, and what rules we are allowed to break and to what extent we are allowed to break them depends on various features of the situation. However, it is possible to think of some situations in which we may lie, even in the kingdom of ends, so the FKE also does not lead to the view that lying is always morally impermissible.

Finally, it is possible to regard the categorical imperative as grounding an objective moral law rather than as a decision-making rule. If we do so, we can explain how apparent moral dilemmas sometimes arise when some of our grounds of obligation are in conflict, as is the case with the murderer at the door. There are principles that obligate us to be truthful in the absence of moral dilemmas, but, when a dilemma arises, we may sometimes be required to lie since this is a lesser evil and the action that best expresses the ideals of the categorical imperative.

The grounds of obligation requiring us to tell the truth stand, even in the presence of a moral dilemma, since lying *always* violates the spirit of the categorical imperative. This is because it disrespects the autonomy of those to whom the lie is told. On this interpretation, there is always *something* morally wrong about lying, *ceteris paribus*. However, we are nevertheless required to lie in certain circumstances. It is not so easy to determine *when* the grounds of obligation to tell the truth may be overridden, but it is clear that they must come into conflict with another, more pressing ground of obligation (such as to save a life) before we can reasonably say that we may violate their commandments. In a moral dilemma, all possible courses of action go against *some or other* ground of obligation, and we have no choice but to override one of them, but they all, nevertheless, stand under ideal conditions.

The preceding analysis shows that Kant's moral theory does not necessarily commit him to the view that we are always obliged to tell the truth. His categorical imperative, in all of its forms, does not conclusively lead us to adopt his rigorous conclusion. While it is likely that lying will turn out to be morally wrong in the majority of cases, it seems plausible to say that there will be at least some cases in which a lie will be justified or even obligatory. Morality is not a discipline in which strict, universal laws can be applied rigidly in all and every circumstance, and there are likely to be situations in which we must depart from the grounds of obligation that would ordinarily bind us. The categorical imperative is not infallible, and it might not even be a decision-making rule, but is nevertheless a useful tool in guiding us as to what we ought to do in a particular case.

BIBLIOGRAPHY

Kant's Works (with abbreviations)

AP: *Anthropology from a Pragmatic Point of View* (1798) (translated and edited by Louden, R. B., 2006). Cambridge: Cambridge University Press.

G: *Grounding for the Metaphysics of Morals* (1785) (translated by Ellington, J.W., 1993). Indianapolis: Hackett Publishing.

LE: *Lectures on Ethics.* (18th Century). In: Heath, P. and Schneewind, J. B. (1997). *The Cambridge Edition of the Works Of Immanuel Kant: Lectures on Ethics.* New York: Cambridge University Press.

MM: *The Metaphysics of Morals.* (1797) (translated by Gregor, M.J.). In: Gregor, M. J. (ed.). (1996). *The Cambridge Edition of the Works Of Immanuel Kant: Practical Philosophy.* New York: Cambridge University Press.

RL: "On a Supposed Right to Lie because of Philanthropic Concerns" (1787), translated by Ellington, J.W., 1993). Indianapolis: Hackett Publishing.

Other Sources

Adler, J. E. (1997). "Lying, Deceiving or Falsely Implicating." In: *The Journal of Philosophy,* Vol. 94, No. 9 (September 1997). 435-52.

Audi, R. (ed.) (1999). *The Cambridge Dictionary of Philosophy* (2nd edition). Cambridge University Press: Cambridge.

Benton, R. J. (1982). "Political Expediency and Lying: Kant vs. Benjamin Constant." In: *Journal of the History of Ideas,* Vol. 43, No. 1 (January 1982). 135-44.

Beck, L. W. (1971). "Kant and the Right of Revolution." In: *Journal of the History of Ideas,* Vol. 32, No. 3 (July – September 1971). 411-22.

Carson, T. L. (2006). "The Definition of Lying". In: *Noûs*, Vol. 40, No. 2 (2006): 284-306.

Chisholm, R. M. and Feehan, T. D. (1977). "The Intent to Deceive". In: *The Journal of Philosophy*, Vol. LXXIV, No. 3 (March 1977). 143-59.

Constant, B. (1787). *Des Réactions Politiques*. Available online: "Les Classiques des Sciences Sociales" [http://pages.infinit.net/sociojmt.]. Accessed 5 December 2009.

Garnett, A. C. (1964). "A New Look at the Categorical Imperative." In: *Ethics*, Vol. 74, No. 4 (July 1964). 295-9.

Guyer, P. (2007). *Kant's Groundwork for the Metaphysics of Morals.* London: Continuum Books.

Hanna, R. (2009). "Living with Contradictions: The Logic of Kantian Ethics in a Nonideal World." Unpublished. Available online: [http://www.colorado.edu/philsophy/paper_hanna_living_with_contradictions_dec09.pdf]. Accessed 17 February 2010.

Herman, B. (1993). *The Practice of Moral Judgment.* London: Harvard University Press.

Hofmeister, H. E. M. (1972). "Truth and Truthfulness: A Reply to Dr. Schwarz". In: *Ethics*, Vol. 82, No. 3 (April 1972). 262-7.

Korsgaard, C. M. (1986). "The Right to Lie: Kant on Dealing with Evil." In: *Philosophy and Public Affairs*, Vol. 15, No. 4 (Autumn 1986). 325-49.

Korsgaard, C. M. (1996). "Two Arguments against Lying." In: Korsgaard, C. M. (1996). *Creating the Kingdom of Ends.* Cambridge: Cambridge University Press. 335-63.

Mahon, J. E. (2009). "The Truth About Kant on Lies". In: Martin, C. (ed.) (2009). *The Philosophy of Deception*. Oxford: Oxford University Press. 201-224.

Margolis, J. (1963). "'Lying is Wrong' and 'Lying is not Always Wrong'". In: *Philosophy and Phenomenological Research,* Vol. 23, No. 3 (March 1963). 414-8.

Nagel, T. (1972). "War and Massacre". In: *Philosophy and Public Affairs*, Vol. 1, No. 2 (Winter 1972). 123-44.

Nicholson, P. (1976). "Kant on the Duty Never to Resist the Soverign." In: *Ethics,* Vol. 86, No. 3 (April 1976). 214-30.

O'Neill, O. (2002). "Instituting Principles: Between Duty and Action." In: Timmons, M. (ed). (2002). *Kant's Metaphysics of Morals: Interpretative Essays.* Oxford: Oxford University Press. 331-347.

Pogge, T. W. (1998). "The Categorical Imperative". In Guyer, P. (1998) (ed). *Kant's* Groundwork of the Metaphysics of Morals: *Critical Essays.* New York: Rowman and Littlefield Publishers, Inc.

Rawls, J. (1955). "Two Concepts of Rules." In: *Philosophical Review,* Vol. 64, No. 1 (January 1955). 3-32.

Rawls, J. (1971). *A Theory of Justice.* Cambridge, MA: The Belknap Press of Harvard University Press.

Reiss, H. S. (1956). "Kant and the Right of Rebellion." In: *Journal of the History of Ideas,* Vol. 17, No. 2 (April 1956). 179-92.

Ross, W. D. (1930). *The Right and the Good*. Oxford: Oxford University Press.

Schapiro, T. (1999). "What is a Child?" In: *Ethics*, Vol. 109, No. 4 (July 1999). 715-38.

Schapiro, T. (2003). "Compliance, Complicity and the Nature of Nonideal Conditions." In: *The Journal of Philosophy*, Vol. 100, No. 7 (July 2003). 329-55.

Schapiro, T. (2006). "Kantian Rigorism and Mitigating Circumstances." In: *Ethics,* Vol. 117 (October 2006): 32-57.

Schwarz, W. (1964). "The Right of Resistance". In: *Ethics*, Vol. 74, No. 2 (January 1964). 126-34.

Schwarz, W. (1970). "Kant's Refutation of Charitable Lies." In: *Ethics*, Vol. 81, No. 1 (October 1970). 62-7.

Siegler, F. A. (1966). "Lying." In: *American Philosophical Quarterly,* Vol. 3, No. 2 (April 1966). 128-36.

Singer, M. G. (1954). "The Categorical Imperative." In: *The Philosophical Review*, Vol. 63, No. 4 (October 1954): 577-91.

Stratton-Lake, P. (2000). *Kant, Duty and Moral Worth.* London: Routledge.

Strudler, A. (1995). "On the Ethics of Deception in Negotiation". In: *Business Ethics Quarterly,* Vol. 5, No. 4, The Environment (October 1995). 805-22.

Timmermann, J. (2007). *Kant's Groundwork of the Metaphysics of Morals: A Commentary.* Cape Town: Cambridge University Press.

Wood, A. (1999). *Kant's Ethical Thought*. Cambridge: Cambridge University Press.

Wood, A. (2002). "The Final Form of Kant's Practical Philosophy". In: Timmons, M. (ed). (2002). *Kant's Metaphysics of Morals: Interpretative Essays.* Oxford: Oxford University Press. 1-21.

Wood, A. (2009). "Kant and the Right to Lie". Unpublished.

Wood, A. and O'Neill, O. (1998). "Kant on Duties Regarding Nonrational Nature". In: *Proceedings of the Aristotelian Society, Supplementary Volumes,* Vol. 72 (1998). 189-228.