The Scope and Force of the Ideal World Objection

By Jonathan Fryer

UCL

MPhil Stud

Philosophy

Supervised by Dr Joe Horton

I, Jonathan Fryer, confirm that the work presented in my thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

This thesis is concerned with the Ideal World Objection (IWO), how it has force, and how widely it applies. This is an objection that affects moral theories which determine what one morally ought to do by evaluating possible worlds that differ from the actual world in more than what is up to us. There are many moral theories that determine what one morally ought to do in this way, including various versions of Rule Consequentialism, Kantian Ethics, and Moral Contractualism. These are otherwise attractive moral theories — each has strong explanatory power, capturing many of our moral intuitions.

I argue that the IWO's force is widely misinterpreted. I do so by using Abelard Podgorski's formulation of the IWO, which he calls the Distant World Objection (DWO). I argue that the IWO has a very wide scope — wider than even Podgorski believes. As well as providing an ordered exposition of various moral theories, this thesis makes three contributions to the existing literature on the IWO.

First, it provides greater explanation of why solutions to the IWO from proponents of various versions of Rule Consequentialism are unsuccessful (**3.3**). Second, it argues that Korsgaard's (1986) paper, *The right to lie: Kant on dealing with evil*, which attempts to resolve the IWO for Kantian Ethics, misinterprets the force of the objection (**4.5**-**4.6**). Korsgaard, and the proponents of Rule Consequentialism, have misunderstood the *force* of the IWO. Third, I argue that Podgorski's claim that Rawls' Theory of Justice is not affected by the DWO is false. In doing so, I show that the IWO has a wider *scope* than even Podgorski believes (**5.4**-**5.5**).

Impact Statement

As mentioned above, this thesis provides an ordered exposition of various moral theories and a detailed explanation of Podgorski's formulation of the IWO (the DWO). It also makes three contributions to the existing literature on the IWO.

First, it provides greater explanation of why solutions to the IWO from proponents of various versions of Rule Consequentialism are unsuccessful (**3.3**). Second, it argues that Korsgaard's (1986) paper, *The right to lie: Kant on dealing with evil*, which attempts to resolve the IWO for Kantian Ethics, misinterprets the force of the objection (**4.5**-**4.6**). Third, I argue that Podgorski's claim that Rawls' Theory of Justice is not affected by the DWO is false. In doing so, I show that the IWO has a wider scope than even Podgorski believes (**5.4**-**5.**5).

The last of these contributions is the one that could have the most novel impact for the existing literature. In the final Chapter, I argue that a response is required from Rawlsian theorists to show how Rawls' Theory of Justice does not suffer from the IWO. Accordingly, this thesis has the potential to open a dialogue between proponents and critics of Rawls upon whether it faces this objection. Thus, the IWO might be expanded into further areas of research.

Table of Contents

List of Abbreviations

IWO	Ideal World Objection
DWO	Distant World Objection
AC	Act Consequentialism
AU	Act Utilitarianism
AP	Pluralist Act Consequentialism
RC	Rule Consequentialism
RU	Rule Utilitarianism
RP	Pluralist Rule Consequentialism
UCRC	Universal Compliance Rule Consequentialism
UARC	Universal Acceptance Rule Consequentialism
PCRC	Partial Compliance Rule Consequentialism
PARC	Partial Acceptance Rule Consequentialism
FRRC	Fixed-Rate Rule Consequentialism
ORRC	Optimum-Rate Rule Consequentialism
ARRC	Average-Rate Rule Consequentialism
ERRC	Every-Rate Rule Consequentialism
KE	Kantian Ethics
FoUL	Formula of Universal Law
FoH	Formula of Humanity
KoE	Kingdom of Ends
HC	Hobbesian Contractarianism
KC	Kantian Contractualism

Introduction

This thesis is concerned with the Ideal World Objection (IWO), how much force it has, and how widely it applies. It is an objection that is said to affect moral theories which determine what one ought to do by evaluating possible worlds that differ from the actual world in more than what is up to us.[1] This objection is important because there are many moral theories that determine what one ought to do in this way. This group contains various versions of Rule Consequentialism, Kantian Ethics, and Moral Contractualism. These are moral theories that many people find attractive — each has strong explanatory power, capturing many of our intuitions about which acts are morally permissible or impermissible and why.

There are many formulations of the IWO,[2] but the one I take to be most informative comes from Abelard Podgorski. He calls his formulation the Distant World Objection (DWO). He writes that any moral theory which determines what one ought to do by evaluating possible worlds that differ from the actual world in more than what is up to us will cause us to either:

(1) Do something avoidably disastrous; or

(2) Avoid doing something wonderful for no good reason; or

(3) Allow irrelevant facts about distant worlds to determine what we ought to do.[3]

I think that the IWO is the most serious objection facing these moral theories. Any moral theory that implies you ought to do something avoidably disastrous — or avoid doing

[1] Podgorski (2018) p.279
[2] See Rumbold (2024) fn3
[3] Podgorski (2018) p.290

something wonderful without a very strong reason — is deeply flawed.[4] So, it is a problem for a moral theory if it faces the IWO.

This thesis is a critical analysis of the scope and force of the *Ideal World* Objection, couched primarily in the language of Podgorski's *Distant World* Objection. This is because, by using Podgorski's DWO, I am making a claim about the scope and force of the IWO. As we will see, Podgorski uses the DWO to argue that various attempts to resolve the IWO fail. He argues that these attempts fail because they misinterpret why the IWO affects moral theories — they misinterpret the *force* of the objection. He also uses the DWO to argue that the IWO affects a wide range of moral theories — this is a claim about the *scope* of the objection. So, I use the DWO as the formulation of the IWO which best shows its scope and force. The thesis proceeds as follows.

The first Chapter sets out the most influential version of Consequentialism: Act Consequentialism (AC). It does so to provide a foil for Rule Consequentialism (and the other moral theories affected by the Distant World Objection), introduced subsequently. I initially formulate AC as *an act is morally permissible if and only if there is no other available act that would make things go better* in (**1.1**). To know what AC implies, we need to know what is meant by 'best', so (**1.2**) clarifies how we determine what makes outcomes as good as possible, or 'best', under AC. (**1.3**) provides 3 reasons why one might accept AC. I also explain how AC deals with cases where one does not know what is 'best', either due to uncertainty regarding the outcomes of our act(s), or due to false beliefs regarding those outcomes (**1.4**). I then present the 3 objections to AC that I view as most challenging (**1.5**). I

[4] Whether an act is morally permissible is different to whether one ought to perform that act. However, when there is only one morally permissible act available to you, and every other available act (including inaction) is morally impermissible, it follows that you ought to act in that way.

do so to motivate the introduction of an alternative version of Consequentialism, Rule Consequentialism (RC), in Chapter 2.

In Chapter 2, I define RC. I initially formulate RC as *an act, A, is morally permissible if and only if it accords with a set of rules, S, such that if everyone followed S, the consequences would be at least as good as if everyone followed any set of rules other than S*. I again clarify what makes things go 'best', and I adapt RC to account for uncertainty and false beliefs (**2.1**). I then distinguish between Universal Compliance Rule Consequentialism and Universal Acceptance Rule Consequentialism (**2.2**), provide 3 reasons why one might accept RC (**2.3**), and show how RC resolves the issues faced by AC (**2.4**).

Having introduced both AC and RC, Chapter 3 introduces the focus of the thesis: the IWO. I first provide an example to illuminate the IWO (**3.1**). I introduce Podgorski's formulation, the DWO, and explain how it affects RC (**3.2**). The rest of the Chapter discusses two unsuccessful ways a proponent of RC might try to avoid the DWO (**3.3**). This Chapter uses the DWO to show that certain proponents of RC misinterpret the force of the IWO.

The next Chapter looks at how the DWO affects another moral theory, Kantian Ethics (KE). I do this for two reasons. First, to show that the IWO affects more moral theories (has a wider scope) than just RC. Second, to discuss another unsuccessful solution to the DWO, taken from Christine Korsgaard's (1986) paper, *The Right to Lie: Kant on Dealing with Evil*. Korsgaard, I argue, also misinterprets the force of the IWO. In (**4.1**), I explain Kantian Ethics. Since Kantian Ethics determines what one ought to do by asking which 'maxims' a person could 'rationally will', in (**4.2**) I explain what it means to 'rationally will' a 'maxim'. (**4.3**) provides 3 reasons why one might accept Kantian Ethics. I then show how the DWO affects

Kantian Ethics (**4.4**), before setting out Korsgaard's attempt to solve the IWO (**4.5**). (**4.6**) shows why this attempt fails to resolve Podgorski's formulation of the problem, the DWO.

In Chapter 5, I challenge one of Podgorski's claims about the scope of the IWO. He suggests that the DWO does not apply to Rawls' Theory of Justice. Since Rawls' Theory is a Contractualist one, I first define Contractualism, and explain the distinction Podgorski makes between Hobbesian and Kantian Contractualist Views (**5.1**), before providing three reasons why one might accept Contractualism (**5.2**). Then I show how the DWO affects Contractualist Views (**5.3**). Having done that, I explain why Podgorski thinks the DWO would not affect Rawls' Theory of Justice (**5.4**). I then argue that he is mistaken, and that his DWO would affect Rawls' Theory of Justice (**5.5**). In doing so, I show that the IWO has a wider scope than even Podgorski believes and show that a satisfying defence of Rawls' Theory of Justice requires a response to the DWO.

This thesis makes three main contributions to the existing literature. First, it provides greater explanation of why various solutions to Podgorski's formulation of the IWO are unsuccessful (**3.3**). Second, it argues that Korsgaard's (1986) paper, *The right to lie: Kant on dealing with evil*, misinterprets the force of the IWO (**4.5**-**4.6**). Third, I argue that Podgorski's claim that Rawls' Theory of Justice is not affected by the DWO is false. In doing so, I show that the IWO has a wider scope than even Podgorski believes (**5.4**-**5.5**).

Before proceeding, I would like to make two clarificatory points.

Firstly, the aim of this thesis is not to convince the reader that any particular moral theory is the one that we should follow. It would be a mistake to view this thesis as a defence of Rule

Consequentialism, for example. Nor do I look to claim that any of the moral theories discussed are better or worse than the others. So, when I ask why one might accept a given moral theory — as I do in (**1.3**), (**2.3**), (**4.3**), and (**5.2**) — I am not making an argument that we *should* accept that theory. I merely aim to show that there are plausible reasons that a person might use each theory to determine what they ought to do.

Secondly, Podgorski uses quite abstract terminology and examples to set out his DWO. One might complain that Podgorski uses such abstract language and cases because his objection is so specific that it does not bear on the real world. This is misguided. Podgorski uses abstract terminology and examples for two reasons: to prevent his reader getting distracted by insignificant details in his examples, and to show that the problem with RC (and, subsequently, other moral theories) is structural, fundamental to its design. The same objection can be pressed with examples that use more concrete, familiar terms. I often attempt to include such cases to supplement Podgorski's abstract examples. Further, even in exclusively hypothetical cases, a moral theory should not produce highly counter-intuitive action-guidance. So, even if the DWO did not apply to many real-world cases, it would still require a response.

1. Act Consequentialism

Consequentialism is a moral theory whose central claim is that there is one ultimate moral aim: that outcomes be as good as possible.[5] There are many different versions of Consequentialism. In this Chapter, I explain the most influential version: Act Consequentialism (AC). I define AC (**1.1**), clarify how we determine what makes outcomes as good as possible, or 'best', under AC (**1.2**), and provide 3 reasons why one might accept AC (**1.3**). I also explain how AC deals with cases where one does not know what is 'best', either due to uncertainty regarding the outcomes of our act(s), or due to false beliefs regarding those outcomes (**1.4**). I then present the 3 objections to AC that I view as most challenging (**1.5**). I do so to motivate the introduction of an alternative version of Consequentialism, Rule Consequentialism (RC), in Chapter 2.

1.1 What is Act Consequentialism?

Act Consequentialism (AC) is a form of Consequentialism that focusses on evaluating the moral status of acts based upon their real-world consequences. It states:

> *(AC1): An act is morally permissible if and only if there is no other available act that would make things go better.*

An Act Consequentialist morally assesses an act by determining how well that act will make things go if performed, and comparing this to how well every other act the agent might perform would make things go. The only morally permissible acts are those that make things

[5] Parfit (1986) p.24

go best. If there are two 'best' outcomes that make things go equally well (i.e., two available acts for which there is no other available act that would make things go better), then both acts are morally permissible.

To know what AC implies, we need to know what is meant by 'best'.

1.2 What Makes Things Go 'Best'?

Deciding what is 'best' requires the adoption of an axiological view — a view on what has value. Different views on what has moral value will lead to AC having different implications regarding which acts are morally permissible.[6] Thus, there are various versions of AC, depending upon which axiology one adopts.

One axiology is Hedonism. According to this view, what would be best is what would create the most happiness, or wellbeing. Combining the Hedonistic axiology with (AC1) gives us a variant of AC known as Act Utilitarianism:

> *(AU): An act is morally permissible if and only if there is no other available act that would produce greater total wellbeing.*

Another axiology might be a Pluralist Axiology. Hedonism is concerned only with what would create the most happiness, or wellbeing. The Pluralist Axiology, meanwhile, holds that other things can make outcomes better, such as knowledge, justice, or the equal distribution of wellbeing. According to this view, what would be best is what would create the greatest

[6] Parfit (1986) p.4

sum of all of these values. Combining the Pluralist axiology with (AC1) gives us a different variant of AC:

> *(AP): An act is morally permissible if and only if there is no other available act that would realise our various values to a greater overall degree.*

Both of these views — AU and AP — are versions of Act Consequentialism. Since they differ in their definition of 'best', there are cases where they produce different moral guidance. For example:

> Case 1: It is Friday night. You have a choice between going to a party or reading philosophy at home. The party will be very fun, increasing your wellbeing significantly, but you will gain very little knowledge. Reading philosophy at home is also fun, but suppose it will increase your wellbeing slightly less than going to the party. You will, however, gain considerably more knowledge reading philosophy than you would going to the party.

AU and AP each provide you with different action-guidance. AU implies that the only morally permissible act available is to go to the party — this will maximise wellbeing. AP implies that the only morally permissible act available is to stay at home and read philosophy. The combined amount of wellbeing and knowledge produced by reading philosophy is greater than the combined amount of wellbeing and knowledge produced by going to the party. Thus, AU and AP imply that different acts are morally permissible.

Although the different versions of AC differ in specific cases, all plausible versions agree that happiness and pleasure are a part of what makes things go 'better', and misery and pain are part of what makes things go 'worse'. As Parfit notes, on all plausible versions of AC, the Hedonistic Theory is *at least part of the truth*.[7] Whenever total wellbeing is the only thing at stake, versions of AC will tend to agree. Therefore, we can assume that total wellbeing is one of the things that makes outcomes better or worse — the examples in this Chapter that discuss total wellbeing, and the assessment of AC built upon them, will therefore apply to AC when combined with most plausible axiologies.

1.3 Why Accept Act Consequentialism?

Prima facie, Act Consequentialism is incredibly compelling. All else being equal, one ought to make things go best. I will now outline 3 reasons why we might accept AC.

First, it is plausible that morality is the impersonal analogue of rationality. It also seems that *a rational agent* cares about making things go best *for them*. If both these statements are true, then morality must be about making things go best *impersonally*, or best *simpliciter*. AC is about making things go best *simpliciter*. This suggests we ought to accept AC.

Second, to deny AC would be to say that sometimes morality permits you to make things go worse. It is very counterintuitive to suggest that it is morally permissible to make things go worse, or morally impermissible to make things go better. This intuition supports AC.

Third, AC provides clear and intuitive action-guidance in a range of cases. For example:

[7] Parfit (1986) p.4

> Case 2: Betty sees an old lady carrying groceries across the road. Betty can either: (a) do nothing; (b) carry the groceries at very little cost to herself; or (c) carry the groceries, and have a 5-minute conversation, at very little cost to herself.

It is intuitively morally right that Betty ought to help, and briefly talk with, the old lady. AC produces this result. Clearly, (a) would produce lesser total wellbeing than (b). And (b) presumably produces less wellbeing than (c). So, (c) is the only morally permissible act available to Betty upon AC. AC is good at providing clear and intuitive guidance in cases where one must choose between multiple good acts. It is also good at providing guidance in cases where one must choose between multiple bad acts:

> Case 3: You are a rescue-helicopter dispatcher. You have two emergencies requiring assistance, but only one helicopter. In the first, one person needs saving. In the second, five people need saving. There is not time to attend to both, and whichever you do not save will surely die. You have 2 options: (d) dispatch the helicopter to emergency one, saving one person but meaning five die; or (e) dispatch the helicopter to emergency two, saving five people but meaning one dies.

Whilst any death is a bad result, most of us would agree that, intuitively, the morally permissible act is (e), to dispatch the helicopter to save five people. AC implies that the only morally permissible act available is (e). The explanation it provides is also fair: the cost of one person's life is outweighed, axiologically speaking, by the benefit of saving five lives. AC can also provide clear and intuitive action-guidance in *non-ideal* cases — cases where there is already wrongdoing present:

> Case 4: A known murderer knocks on your door and asks if you know where your neighbour is. This neighbour is hiding in your garden. You could either (f) tell the truth and watch your neighbour be killed; or (g) lie and save your neighbours life.[8] Your neighbour's not-being-murdered makes things go better than their being murdered.

This Case is one of Kantian Ethics' most controversial.[9] Most people strongly believe that one ought to lie to a murderer at your door. Chapter 3 will show that many moral theories struggle to provide clear, intuitive action-guidance in these *non-ideal* cases. AC, though, provides clear and intuitive guidance: one ought to lie to the murderer at the door — doing so makes things go better than telling the truth.

I have provided 3 reasons why we might accept AC: it cares about what makes things go best *simpliciter,* so is plausibly the impartial analogue of rationality; it is incredibly counter-intuitive to deny; and it provides clear and intuitive action-guidance in a wide range of cases, including *non-ideal* cases. Let us, though, consider another question AC needs to answer.

1.4 What If You Do Not Know What Makes Things Go Best?

Even with an axiology in hand, we still might not know what makes things go best. This can happen for two reasons: first, because you are uncertain about what will make things go best.

[8] Case taken from Kant (1797) 'On a Supposed Right To Lie From Altruistic Motives' 427/348
[9] See Korsgaard (1986); Varden (2010)

Second, because you have false beliefs about what will make things go best. Let us look first at uncertainty.

1.4.1 Uncertainty

Case 5: A stranger will die unless you correctly predict the outcome of a coin toss. Living maximises total wellbeing, dying does not.

(AC1) is a claim about what one ought to do when one knows the consequences of the acts available to them — *an act is morally permissible if and only if there is no other available act that would (in fact) make things go better*. This view is not helpful in Case 5: suppose that in Case 5 the coin will land on Tails — (AC1) would imply that it is morally impermissible to call 'Heads'. This seems wrong. We often do not know what the effects of our acts will be. We need to be able to make claims about the likelihood that a given act will produce a particular result. In this case, one justifiably believes that there is a 50% chance of guessing correctly. Accordingly, we want AC to imply that calling 'Heads' and calling 'Tails' are morally equivalent, permissible acts.

To do this, we can update (AC1) to incorporate *expected value* in our determination of which of our available acts are morally permissible. Expected value is the sum of the utility produced by each possible outcome of an act, multiplied by that outcome's probability of occurring. What one ought to do is the available act whose outcome has the greatest expected value.[10] We can formulate this as:

[10] Parfit (1986) p.25

(AC2): An act is morally permissible if and only if there is no other available act that has greater expected value.

In Case 5, for example, we believe that a fair coin has a 50% chance of landing 'Heads', and a 50% chance of landing 'Tails'. Given some arbitrary units, imagine that the stranger living produces 100 wellbeing, and their dying produces -100. Expected total wellbeing can be calculated by multiplying the probability of each outcome with the wellbeing it would produce, and adding those values together. For example, calling 'Heads' has a 50% chance of producing 100 wellbeing (0.5 x 100) and a 50% chance of producing -100 wellbeing (0.5 x -100). Calling 'Tails' is the same. Each option, then, has an expected total wellbeing value of ((0.5 x 100) + (0.5 x -100)) = 0. For each available act, there is no other available act that has greater expected value. Therefore, (AC2) implies that calling 'Heads' and calling 'Tails' are both morally equivalent, permissible acts. Thus, (AC2) allows AC to deal with cases where we are uncertain about the outcome of our acts.

1.4.2 False Beliefs

As well as being able to incorporate uncertainty, it is also important that AC can accommodate for false beliefs. Let us use an example to see why:

Case 6: A stranger will die unless you correctly predict the outcome of a coin toss. Living maximises total wellbeing, dying does not. You believe the coin to be fair. Your belief is false: this coin is not fair; it lands on tails 99% of the time.

Calling ‘Heads’ has a 1% chance of producing 100 wellbeing (0.01 x 100), and 99% chance of producing -100 wellbeing (0.99 x -100). This option has the expected value of -98. Contrarily, calling ‘Tails’ has a 99% chance of producing 100 wellbeing (0.99 x 100), and a 1% chance of producing -100 wellbeing (0.01 x -100). This option has the expected value of 98. So, in this case, (AC2) implies that it is morally impermissible to call ‘Heads’. This, too, seems wrong. If it is reasonable for you to believe the coin is fair, your treating ‘Heads’ and ‘Tails’ as morally equivalent, permissible acts should be justified. We want to be able to make claims such as: ‘*if someone does what they believe will make the outcome better, they are acting permissibly, even if it happens to make things go worse*’ (and vice versa).[11]

To do this, we can introduce the idea of credences. Rational agents form reasonable predictions about the effects of their acts based upon their credences — their beliefs about what is true. One predicts that both ‘Heads’ and ‘Tails’ are morally equivalent, permissible acts based upon the credence, ‘the coin is fair’. In Case 6, this credence is inaccurate. The coin will, in fact, most likely land on ‘Tails’. But although the credence is false, it is reasonable to hold. False belief *can* explain why calling ‘Heads’ is subjectively permissible, even when it is objectively wrong. We can say that it is morally permissible to act upon the credences that it is rational, based upon the evidence available to you, to hold. Even if you make things go worse because those credences turn out to be false.

We can alter (AC2) to accommodate this:

[11] Parfit (1986) p.24

(AC3): An act is morally permissible if and only if no alternative has greater expected value relative to the credences that are rationally required given the evidence available to you.

Importantly, this alteration does not mean that it is always morally permissible to make things go worse because one has mistaken credences. That depends upon *why* one holds the belief that 'the coin is fair'. One's credences ought to be based upon the available evidence. Holding the belief 'the coin is fair' because you have no evidence to the contrary seems reasonable — most coins are fair; we do not often check them. However, if you hold the belief 'the coin is fair' despite having just purchased the coin and pulled it out of a container labelled '*Trick Coin: Lands Tails Every Time'*, then your credences are not rational.

It seems that an agent is rationally required to hold certain credences based upon the evidence available to them. To make things go worse, based upon an expected value calculation that was itself based upon false (but rationally required) credences, is morally permissible. Note: this does not commit us to saying that it is morally wrong to fail to hold these rationally required credences, *despite* the evidence available to you — we might want to call this irrational, rather than immoral.

(AC3) allows AC to deal with cases of false beliefs: if one holds a false belief, and this false belief is both rationally required by the available evidence, and causes one to make things go worse, one does not act wrongly. (AC3) answers the question, '*what if you do not know what makes things go best?'* AC is now sufficiently formulated. We can now turn to look at some objections that it faces.

1.5 Objections to Act Consequentialism

This Section will state the three objections to AC that I take to be most concerning.

1.5.1 Ignores Rights and Duties

The first major objection to AC is that it ignores the right and duties of agents — or at least the rights and duties that we intuitively feel agents have. All AC cares about is making things go best *simpliciter*; there are no limits to the kind of acts AC will prescribe to achieve this. In (**1.3**), I noted that denying AC counterintuitively involves saying that it is sometimes permissible to make things go worse. It is also counterintuitive for a moral theory to make permissible seemingly morally repugnant acts. For example:

> Case 7: David is a transplant surgeon. Five of his patients need different organs, and all are of the same blood-type. David has a sixth, healthy patient of the same blood-type. David can kill the healthy patient to harvest their organs, saving the others, and nobody will find out. Or he can not kill his healthy patient, and let his other patients die.[12]

This Case seems to be structurally similar to Case 3, where the choice was between the lives of 5 people needing rescuing, or one person needing rescuing. Since the proponent of AC must say that saving the five makes things go best in that case, it seems they must say the same here. AC will then imply that the only morally permissible act is to kill the healthy patient and harvest his organs. Yet, Cases 3 and 7 also seem importantly different. Most of us

[12] Adapted from Thomson (1976) p.206

agree we ought to save the five people in Case 3, but if a surgeon were found to kill a healthy patient to secretly harvest their organs in real life, they would be morally condemned (and arrested). Why? Duties and rights seem to be a large part of the explanation.

Duties are moral obligations that agents have to one another. For example, having promised to help you move, I have a moral obligation (a duty) to keep my word. Since they are moral obligations, it should be morally impermissible to violate one's duties (at least without a very strong reason). Duties can also be generated by rights. Rights are — broadly speaking — things that an agent possesses that limit what others may do to pursue their own ends.[13] Similarly, it should be morally impermissible to violate another's rights (at least without a very strong reason). At least, a moral theory should give some weight to rights and duties.

Duties and rights provide a plausible explanation of why David acts wrongly, whilst the rescue helicopter dispatcher in Case 3 does not. David has a duty to his patients not to secretly murder them to harvest their organs. There is no equivalent duty that the person in control of the helicopters would be violating by choosing to save the five people over the one. Every patient has a right against being secretly murdered for organ harvesting that is stronger than five patients' rights to organs. The one person in Case 3 still has a right to ask for rescue, but so do the 5 people at the second emergency — their rights combined can reasonably be said to outweigh the one's.

AC ignores rights and duties. The Act Consequentialist only adheres to rights and duties when it happens to be the case that doing so makes things go better than violating them. Even in these cases, we can only say that AC happens to coincide with rights and duties — even

[13] Nagel (2007) p.103

when AC can recommend the correct policies, there is something profoundly wrong with how it produces them.[14] When asked what David should do, proponents of AC do not immediately refuse to secretly murder people; they instead ask whether doing so would have good consequences.

This is a serious problem for AC. But the Act Consequentialist might respond to this objection as follows: 'on my axiological view, killing is one of the things that makes things go worse. So, while it is better for only one person to die than for five people to die, it is not necessarily better for one person to be killed than for five people to die.'

However, this response fails, for two reasons.

First, it remains true with this axiology that AC merely coincides with rights and duties in David's Case. There is still something profoundly wrong with needing to do a utility calculation to determine whether it is wrong or not to secretly murder a healthy patient for organ harvesting.

Second, this Act Consequentialist claims that on their axiology, one killing make things worse than five deaths. Such an axiology would produce very counterintuitive cases. For example, imagine that you have a choice between saving five people from dying from a natural disaster, or one person from being killed during a robbery. It seems that we ought to save the five from dying. And yet this version of Act Consequentialism would imply that one ought to save the one person from being killed during the robbery. This is deeply counterintuitive.

[14] Rawls (1997) as found in Arneson (2005) p.26

This remains a serious problem for AC. Earlier, I said that it seemed counterintuitive that morality permits that you make things worse. It is also deeply counterintuitive for morality to ignore rights and duties. A plausible moral theory must give at least some weight to rights and duties.

1.5.2 Demandingness

Another major objection to AC is that, in its pursuit of the best outcome *simpliciter*, it is far too demanding of agents. For example:

> Case 8: Elena earns a modest salary. She wants to enjoy her life and help people where she can. She knows, though, that the act available to her that would produce the greatest total expected wellbeing would be to donate her entire salary to charity.

Even if Elena were to lose her home and starve to death, donating her entire salary to a competent charity could still produce greater total wellbeing than not doing so. The only permissible acts for an Act Consequentialist are those where there is no other available act that would make things go better. This means that the class of morally permissible acts for an Act Consequentialist often consists only of extreme acts. AC cannot incorporate a morally supererogatory class of acts — acts that are morally good but go beyond what is morally required of an agent.[15]

[15] See Urmson (1958)

Donating your entire salary to help others, at your own detriment, should be morally permissible. But it should not be the *only* morally permissible act available to an agent. Nor should it be expected of anyone. A plausible moral theory should not be too demanding. AC is far too demanding to be a plausible moral theory.

1.5.3 Ignores Special Relationships

We seem to have special obligations to certain people based on our relationships with them. These obligations go beyond our obligations to people in general. One has moral obligations to their mother that they do not have to a stranger. Many people believe that these relationships ought to play a role in our moral decision-making. AC, though, cannot give priority to special relationships. For example:

> Case 9: Felicity comes across two people in trouble: her mother and a stranger. Felicity can only help one. Watching one's mother drown makes things go worse *for that person.* Suppose that Felicity knows that, despite the fact things would go better *for her* if she saved her own mother, saving the stranger would make things go just as well *impersonally* as saving her mother.

Common sense suggests that in this case where her two choices make things go equally well, there would be something morally abnormal about Felicity if she did not save her mother over the stranger. Even if saving her mother produced 100 expected total wellbeing and saving the stranger produces 100.01 expected total wellbeing, most of us believe that it is morally permissible for Felicity to help their mother. We believe that a plausible moral theory should give at least some weight to special relationships.

AC ignores special relationships. Saving her own mother produces greater expected wellbeing *for Felicity*. However, in terms of total expected wellbeing, since the options are indiscernible, AC cannot explain why we feel it is morally impermissible to help the stranger. When special relationships are upheld upon AC, it is in the same accidental way that rights and duties were adhered to earlier. It is also deeply counterintuitive for morality to ignore special relationships.

A similar response can be made by an Act Consequentialist here as with rights and duties: 'on my axiological view, special relationships are one of the things that make things go better. So, there is greater value in saving someone that you love than in saving a stranger.' This response fails for the same two reasons: AC continues to merely coincide with special relationships, and the proposed axiology leads to counterintuitive cases. For example: imagine that you can either save two strangers, or you can enable a stranger to save someone that they love. So, you can either save two strangers, or help someone save one stranger. If there is greater value in saving someone that you love than in saving a stranger, this version of AC could imply that one ought to let two strangers drown to allow a different stranger to save someone that they love. This is counterintuitive. This too remains a serious problem for AC.

Conclusion

AC is *prima facie* very plausible. This is because it is a plausible impartial analogue of rationality; it is counterintuitive to deny; and it provides clear, intuitive action-guidance in a wide range of cases. However, it also faces objections of real concern: it ignores rights and

duties, it is too demanding, and it ignores special relationships. At least, a moral theory should give some weight to rights and duties. A moral theory should not be too demanding. At least, a moral theory should give some weight to special relationships. AC fails on all three counts.

The next Section introduces a different version of Consequentialism: Rule Consequentialism (RC). I will argue that it is also *prima facie* very plausible, that it bears a plausible relation with rationality, that it is counterintuitive to deny, and that it provides intuitive action-guidance in a wide range of cases. Unlike AC, though, I will show that it can give at least some weight to rights and duties, it is not too demanding, and it can give at least some weight to special relationships. Despite facing its own objections, this makes it a very attractive alternative to AC.

2. Rule Consequentialism

The previous Chapter set out AC — an initially attractive but also problematic version of the moral theory, Consequentialism. In this Chapter, I introduce another version of Consequentialism, RC. I hope to present RC as an attractive moral theory which solves the problems faced by AC. To do so, in (**2.1**), I define RC, I again clarify the role of axiology, and I adapt RC to account for uncertainty and false beliefs. I then distinguish between Universal Compliance Rule Consequentialism and Universal Acceptance Rule Consequentialism (**2.2**), provide 3 reasons why one might accept RC (**2.3**), and show how RC resolves the issues faced by AC (**2.4**). In the next Chapter, I will then set out what I take to be RC's most pressing problem: the Distant World Objection.

2.1 What is Rule Consequentialism?

RC is another version of Consequentialism. The simplest form of Rule Consequentialism states:

> *(RC1): An act, A, is morally permissible if and only if it accords with a set of rules, S, such that if everyone followed S, the consequences would be at least as good as if everyone followed any set of rules other than S.*[16][17]

A Rule Consequentialist morally assesses an act by determining how well things would go if every agent acted that way in those circumstances (i.e., if a rule requiring the act in these

[16] Podgorski (2018) p.280
[17] RC is not exclusively act-focussed. It can also provide guidance on which dispositions and beliefs one ought to have. In this thesis, though, I focus only on acts.

circumstances were universally followed, or 'universalised'), and then compares this to how well every other act the agent might perform would make things go if universalised. The only morally permissible acts are those that follow rules that — if universalised — would make things go best. If there are two sets of rules, S and T, for which there are no other sets of rules that would make things go better if universalised, then acts that adhere to either S or T are morally permissible.

RC also requires an axiology to understand what is meant by 'best'. Just like AC, we can combine RC with the Hedonistic or the Pluralist axiology to produce different versions of RC:

> *(RU): An act, A, is morally permissible if and only if it accords with a set of rules, S, such that if everyone followed S, the total sum of wellbeing produced would be at least as great as if everyone followed any set of rules other than S.*

> (*RP): An act, A, is morally permissible if and only if it accords with a set of rules, S, such that if everyone followed S, our various values would be realised to a greater overall degree than if everyone followed any set of rules other than S.*

Both of these views — RU and RA — are versions of RC. Since they differ in their definition of 'best', there are cases where they produce different moral guidance. However, as with the corresponding versions of AC discussed in (**1.2**), all plausible versions of RC agree that happiness and pleasure are a part of what makes things go 'better', and misery and pain are part of what makes things go 'worse'. The Hedonistic Theory is again *at least part of the*

truth.[18] Whenever total wellbeing is the only thing at stake, versions of RC will tend to agree. Therefore, we can assume that total wellbeing is one of the things that makes outcomes better or worse — the examples in this Chapter that discuss total wellbeing, and the assessment of RC built upon them, will therefore apply to RC when combined with most plausible axiologies.

Just as with AC, we must answer the question '*what if you do not know what makes things go best?'* And equally, we can alter RC to introduce expected value and credences, in order to account for uncertainty and false beliefs:

> *(RC2): An act, A, is morally permissible if and only if it accords with a set of rules, S, such that one believes (relative to the credences that are rationally required given the evidence available to you) that the expected value of everyone following S would be at least as good as if everyone followed any set of rules other than S.*

RC still requires further formulation before we assess whether it ought to be accepted, because there are various ways that the idea of following a rule can be interpreted. We need to distinguish between two versions of RC.

2.2 Universal Compliance vs Universal Acceptance[19]

To be able to speak about the consequences of a set of rules, we need to understand the relation that set of rules has to the world. There are two ways that rules can be related to the

[18] Parfit (1986) p.4
[19] This distinction as made by Podgorski (2018) p.280

world worth discussing: rules can be *complied with*, or they can be *accepted*. If a rule is *complied with* by an agent, this means the agent merely acts according to the rule. If a rule is *accepted* by an agent, that means that as well as acting according to the rule, the agent psychologically internalises the rule in some way. They are aware of it and agree with it. Using this distinction, we can formulate two different versions of RC:

> *Universal Compliance RC (UCRC): An act, A, is morally permissible if and only if it accords with a set of rules, S, such that one believes (relative to the credences that are rationally required given the evidence available to you) that if everyone complied with S, the consequences would be at least as good as if everyone complied with any set of rules other than S.*

> *Universal Acceptance RC (UARC): An act, A, is morally permissible if and only if it accords with a set of rules, S, such that one believes (relative to the credences that are rationally required given the evidence available to you) that if everyone accepted S, the consequences would be at least as good as if everyone accepted any set of rules other than S.*

Universal acceptance of a rule can have consequences that mere compliance with a rule does not. For example, the universal internalization of a rule can come with a cost that mere compliance does not. These consequences can change the total sum of wellbeing produced by the set of rules. Accordingly, there are cases where an act is permissible under UCRC, but not under UARC (and vice versa).[20]

[20] Whether one can accept a moral rule without complying with it is unclear. Podgorski (2018) follows Hooker (2000) in saying that acceptance does imply compliance (see Podgorski (2018) fn2; Hooker (2000) p.76); I will do the same. But it does not affect my argument if acceptance does *not* imply compliance — the examples used throughout can be adjusted, and the Ideal World Objection will still apply.

We will see later an example of UCRC and UARC producing different sets of morally permissible acts (**2.4.1**). It is not vital to this Chapter to decide which of UCRC and UARC is the more plausible variation of RC — both will be shown to be problematic in the next Chapter. (**2.4.1**), though, will provide a reason why we might consider UARC to be more plausible than UCRC. I accept this reason. Before that, though, let us consider why one might accept any form of RC.

2.3 Why Accept Rule Consequentialism?

Prima facie, RC is also incredibly compelling. It makes sense to think of rules to govern moral conduct. I will now outline 3 reasons why we might accept RC.

First, RC captures some of our powerful moral intuitions about what makes an act morally impermissible. RC asks one to think, "*what if everyone thought or acted like you do?*" This is a phrase often used to explain why a person's act is morally wrong. By asking a perceived wrongdoer this question, you imply that the world would be worse off if everybody acted like them. The wrongdoer acts in a way that a reasonable person hopes others would not; they make an exception of themselves. For that reason, their act was wrong. RC has great explanatory power when it comes to *why* an act is morally impermissible.

Second, it is plausible that morality consists of a set of rules that ought to govern our behaviour. It is likely that — if this were true — the set of rules that makes things go best when universally complied with and/or accepted is the only set of rules that we could all rationally agree to. It is also plausible that we ought to follow the set of rules that everyone

could rationally agree to. Thus, the Rule Consequentialist project seems well motivated, and bears a plausible relation to rationality.

Third, RC also provides clear and intuitive action-guidance in a range of cases. It does so in many cases where AC struggled to provide such guidance. As such, the next Section will kill two birds with one stone: it provides cases where RC provides clear and intuitive action-guidance; it also shows how RC solves the most pressing objections faced by AC.

2.4 How does Rule Consequentialism avoid Act Consequentialism's problems?

In (**1.5**), I set out some concerns for AC: a moral theory should give at least some weight to rights and duties (**1.5.1**), and special relationships (**1.5.3**); it should also not be too demanding (**1.5.2**). I showed that AC fails to meet these requirements. This Section will show how RC succeeds at meeting these requirements. In doing so, I also give evidence that RC provides clear and intuitive action-guidance in a wide range of cases.

2.4.1 Rights and Duties

The first objection I presented to AC was that it cannot accommodate rights and duties. Recall:

> Case 7: David is a transplant surgeon. Five of his patients need different organs, and all are of the same blood-type. David has a sixth, healthy patient of the same blood-

> type. David can kill the healthy patient to harvest their organs, saving the others, and nobody will find out. Or he not kill his healthy patient, and let his other patients die.[21]

Under AC, the only morally permissible act available to David is to murder the healthy patient to harvest his organs — this result is deeply counterintuitive and shows AC's inability to respect rights and duties. UCRC provides the same guidance — the only morally permissible act available to David is to murder the healthy patient. This is why I take UARC to be more plausible, as I mentioned in (**2.2**).

Consider the rule 'never kill your healthy patients, unless it makes things better.' In a world where this rule were universally complied with, countless examples of healthy people being murdered by their doctor would cause people to fear going to the doctor. This would presumably bring about the collapse of the healthcare system, and greatly worsen total wellbeing. The same would happen if this rule were universally accepted. There are many other rules that would make things go better if universally complied with or accepted — for example, 'never kill your healthy patients'. It appears proponents of both UCRC and UARC would condemn this rule.

But now consider the rule 'never kill your healthy patients, unless it makes things go better and you can do so in complete secrecy.' Here, UCRC and UARC seem to imply different action guidance.

It seems that UCRC would permit compliance with this rule. If everyone complied with the rule, the breakdown of trust in the healthcare industry would not occur — the only cases in

[21] Adapted from Thomson (1976) p.206

which a doctor would kill their healthy patients are those where they can do so without anybody finding out. So, nobody would know that their doctor was complying with the rule. Nobody would be afraid of their doctor. In fact, the rule 'never kill your healthy patients, unless it makes things go better and you can do so in complete secrecy' would make things go better than the rule 'never kill your healthy patients'. This is certain, because the only time a doctor would kill their healthy patient is when it would make things go better. So, UCRC implies that if David, the transplant surgeon, was certain he could harvest the healthy patient's organs with nobody finding out, then to do so is morally permissible. If it is the only permissible act available to him, UCRC implies that it is morally required. This is a counterintuitive result.

Contrarily, it seems that UARC would not permit acceptance of this rule. If everyone accepted the rule 'never kill your healthy patients, unless it makes things go better and you can do so in complete secrecy', they would be accepting that it is morally permissible to kill patients in such circumstances. This would provide healthy people with a reason to fear going to their doctor — they might be putting themselves in a situation in which it is morally permissible for their doctor to kill them. So, universal acceptance of the rule has significant, negative effects on total wellbeing that mere universal compliance would not. Presumably, universal acceptance of an alternative rule, such as 'never kill your healthy patients', would make things go better. There would be no reason to fear being murdered by your doctor. Accordingly, UARC would not permit acceptance of the rule 'never kill your healthy patients, unless it makes things go better and you can do so in complete secrecy.'

UCRC and UARC imply different action guidance in such cases. I take UARC as the most plausible version of RC discussed so far. It can give at least some weight to rights and duties.

UCRC would imply that the world in which it is permissible to secretly violate rights and duties is best. Contrarily, we can reasonably predict that a world where rights and duties are universally accepted as inviolable (without a very good reason) is one in which things go better than a world where it is universally accepted that rights and duties can be violated if it makes things go better and nobody ever finds out. UARC can better accommodate rights and duties.

This does instrumentalise right and duties — they do not have innate value (value in themselves). Instead, their value is derivative from a more fundamental value specified by the axiology one combines with RC (for example, wellbeing, in the Hedonistic axiology).[22] For some, this provides insufficient weight to rights and duties: they are valuable only because they happen to accord with what produces the greatest total wellbeing. Nevertheless, they have some value. Therefore, RC fulfils a condition that AC could not: it gives at least some weight to rights and duties.

2.4.2 Demandingness

The next criticism of AC was that, often, the only morally permissible acts are too demanding. Recall Case 8:

> Case 8: Elena earns a modest salary. She wants to enjoy her life and help people where she can. She knows, though, that the act available to her that would produce the greatest total expected wellbeing would be to donate her entire salary to charity.

[22] Nagel (2007) p.103-4

Under RC, Elena's morally permissible act is less demanding. The rule 'donate one's entire salary to charity' — if universally complied with and/or accepted — would result in a surplus of charitable money, which would likely lead to waste of resources; this world would likely not be the one with the greatest total expected wellbeing. The rule 'donate nothing to charity' — if universally complied with and/or accepted — would likely not produce the greatest total expected wellbeing either. No charity would occur. The rule 'donate to charity whatever one would like' is essentially the rule we follow now. Universal compliance with and/or acceptance of 'donate 5% of one's salary to charity' would likely produce greater total expected wellbeing. Presumably, 5% of everyone's salary would be enough money to fund all necessary charitable endeavours. We can reasonably believe that this rule (or the set of rules it belongs to) would produce the greatest total wellbeing.

We have fulfilled another condition that AC could not: a moral theory should not be too demanding. RC asks Elena only to do her 'fair share' and donate a modest amount of her salary to charity. That is a manageable demand. RC is less demanding than AC.

2.4.3 Special Relationships

The third objection was that AC gives no weight to special relationships. Recall:

> Case 9: Felicity comes across two people in trouble: her mother and a stranger. Felicity can only help one. Watching one's mother drown makes things go worse *for that person*. Suppose that Felicity knows that, despite the fact things would go better *for her* if she saved her own mother, saving the stranger would make things go just as well *impersonally* as saving her mother.

RC can value special relationships. Consider the rule 'ignore your special relationships'. If this rule were universally complied with, it would likely decrease overall wellbeing. Whilst in Felicity's case, there is no difference in total wellbeing between saving her mother and saving a stranger, the overall consequence of universal compliance with the rule 'ignore your special relationships' would be negative. Special relationships bring value to our lives, and we are disposed to be partial to those we hold special relationships with. It would cause great suffering if people had to ignore their special relationships when determining what they ought to do.

If this rule were universally accepted, we would likely lose the notion of a special relationship altogether. If everyone accepted that it was morally impermissible to be partial to those you are close to, then everyone would be accepting that special relationships are morally impermissible. A special relationship *just is* a feeling of partiality one holds towards another because of their closeness. Universal acceptance of this rule would curtail an individual's ability to foster loving, caring relationships with other people. This would make our lives much worse.

Contrastingly, consider the rule 'value your special relationships a little more than your ordinary relationships'. If universally complied with, we would not suffer the difficulty of ignoring the special relationships that we are strongly disposed to hold. If universally accepted, it remains morally permissible to hold special relationships. We can reasonably expect that each of these scenarios makes things go better than the rule 'ignore your special relationships.' So, compliance with or acceptance of the rule 'ignore your special relationships' is morally impermissible. Therefore, the only morally permissible acts

available to Felicity are those that adhere to compliance with or acceptance of the rule 'value your special relationships a little more than your ordinary relationships'; Felicity is morally permitted only to save their mother over the stranger.

Whilst AC could not explain why Felicity intuitively ought to save their mother, RC can. Like with rights and duties, special relationships under RC have instrumental value — they are valuable insofar as the world in which they exist produces greater total expected wellbeing as the world in which they do not. Like with rights and duties, some view this as insufficient. Nevertheless, this is sufficient to fulfil the condition that AC could not: RC gives at least some weight to special relationships.

Conclusion

In this Chapter, I have introduced RC, another version of Consequentialism. I have shown that there are good reasons to accept RC as an independent moral theory: it is *prima facie* plausible; it has explanatory power regarding why some acts are morally impermissible; it is plausible that morality is about finding the best set of rules to govern our behaviour, and so RC bears a plausible relation to rationality; and RC provides clear and intuitive action-guidance in a wide range of cases. I have also shown that RC solves the problems that I took to be most pressing for AC in Chapter One: it gives at least some weight to rights, duties, and special relationships, and is not too demanding. I hope to have presented RC as an attractive moral theory. The next Chapter will introduce what I take to be RC's most pressing objection, and the focus of this thesis: the Distant World Objection.

3. The Distant World Objection

The Distant World Objection (DWO) is Podgorski's formulation of the Ideal World Objection. Podgorski's formulation is particularly informative because it exposes the flaws of various solutions offered to the Ideal World Objection by proponents of the moral theories it affects. By using the DWO, I am making a claim about both the force and the scope of the Ideal World Objection.

In this Section, I will first provide an example to illuminate the IWO (**3.1**). I then introduce Podgorski's formulation, the DWO, and show how it affects RC (**3.2**). The rest of the Chapter discusses two ways that proponents of RC have tried to avoid the Ideal World Objection. Each fails to avoid Podgorski's formulation, the DWO. Thus, they misinterpret the force of the IWO. The next Chapter looks at how the DWO affects other moral theories and another unsuccessful attempt to avoid the objection.

3.1 Setting Up the Problem

Imagine the following situation:

> Case 10: You and Greg each have two buttons in front of you, A and B. If you both press A, you each receive £1,000,000. If you both press B, nothing happens. But if one presses A and one presses B, the world ends. If one or both of you fail to press a button, or press multiple buttons, the world ends. Suppose that you are 100% certain that Greg will press button B.

What does RC imply in this Case? The rule '*press button A'* — if universally complied with or accepted — would cause everyone in such cases to receive £1,000,000. The rule '*press button B'* — if universally complied with or accepted — would cause everyone to receive nothing. Presumably, receiving £1,000,000 would make things go better than receiving nothing. So, RC implies that the only morally permissible actions are those that accord with the rule *'press button A'*; the only morally permissible action available to you is to press A.[23]

But you know that Greg will press B. So, if you follow RC and press A, the world ends. This is a disastrous result.

Greg is not following the rule *'press button A'.* And, intuitively, this fact should be relevant when determining what we ought to do. But RC is insensitive to this fact. Instead, RC determines what we ought to do based on facts about a world where you, Greg, and everyone else follows a particular rule. In doing so, RC implies that the only morally permissible act available to you is to knowingly destroy the world. And, as we will see, there are many real-life cases where RC implies that the only morally permissible act available to you is to do something avoidably disastrous *because* of certain facts about what happens only in a faraway world where everyone follows a certain rule. That is the problem.

Let us now look at Podgorski's formulation of the Ideal World Objection, the DWO.

3.2 What is the DWO?

[23] One might object: '*'press button A'* and *'press button B'* are overly simplistic rules that nobody would follow'. In (**3.2-3.3**), we will see that more complicated rules also face the DWO. Regardless, the rule '*press button A'* is optimal according to RC, so you are permitted to comply with it.

Podgorski's DWO states that *any view which determines what we ought to do by evaluating possible worlds that differ from the actual world in more than what is up to us* will cause us to either:

(i) Do something avoidably disastrous; or

(ii) Avoid doing something wonderful without good reason; or

(iii) Allow irrelevant facts about distant worlds to determine what we ought to do.[24]

This seems a problem, as it is incredibly plausible to accept the following:

(1) We should not do anything avoidably disastrous.

(2) We should not avoid doing something wonderful without a very strong reason.

(3) Certain facts about what happens only in faraway possible worlds are irrelevant to what we ought to do.[25]

In Case 10 with Greg and the buttons, RC implies that the only morally permissible act available to you is to do something avoidably disastrous (knowingly destroy the world), and it is the only morally permissible available act *because* of certain facts about what happens only in a faraway world where everyone follows a certain rule. RC causes you to violate (1) and (3).

It will be helpful to focus on a generalised version of the DWO presented by Podgorski. Podgorski introduces the idea of utility landmines — abstract wellbeing bombs that are inert

[24] Podgorski (2018) p.290
[25] As assumed by Podgorski (2018) p.285

until specific trigger conditions are met. He uses an abstract example to avoid distracting features of real-life cases, and to allow us to look at the problem with RC more structurally.[26] Utility landmines can be good (i.e., a 'good-mine') and release an unimaginable amount of wellbeing when triggered; bad (i.e., a 'bad-mine') and release an unimaginable amount of suffering when triggered; or have trigger conditions for both. He also proposes a dud factory — a factory that produces utility landmines with trigger conditions in worlds very distant to ours, such that nothing we could do in the actual world could ever trigger the mine.[27]

Podgorski asks us to consider the rule, *R*, that says 'do a jumping jack at noon every day'. Imagine further that you know that, in your world, there exist two mines:

1. A good-mine that triggers when there is universal adherence to *R*; and
2. A bad-mine that triggers when 1 or more people adhere to *R*, but there is not universal adherence to R.[28]

Suppose that you know both mines exist. Our two versions of RC (UCRC and UARC) both imply that the only morally permissible act available in this world is to go outside at noon and do a jumping jack. If everyone complied with R, then the good-mine would be triggered, releasing an unimaginable amount of wellbeing. If everyone accepted R, the result would be the same. So, on both versions of RC, going outside and doing a jumping jack is the only morally permissible action available to you, as it is your only available action that adheres to R. Realistically, though, there would not be universal adherence to *R*, so when you go outside to do a jumping jack, you do so knowing that your action will most likely trigger the bad-

[26] Podgorski (2018) p.284
[27] Podgorski (2018) p.284-285
[28] Here, I use 'adherence' as a term that is deliberately ambiguous between compliance and acceptance.

mine and release unimaginable suffering. Thus, in this Case, we violate (1) and (3) — we do something avoidably disastrous *because* we allow irrelevant facts about distant worlds to determine what we ought to do.

This is a serious objection. Any moral theory that implies you ought to do something avoidably disastrous — or avoid doing something wonderful without a very strong reason — is deeply flawed. And there are many real-life phenomena that are comparable to utility mines. For example, take the rule '*never use violence'*.[29] If everyone complied with (or accepted) this rule, the result would be world peace. Presumably, the universalisation of this rule would make things go at least as well as the universalisation of any other rule about violence. So, actions that accord with this rule would be permissible upon both UARC and UCRC. But in the real world, following this rule would make things go badly — it would prevent you from being able to defend yourself or protect others from wrongdoers.[30] RC misses the important fact that wrongdoers exist. As Podgorski notes, *"[t]he fact that sometimes people are violent matters"*.[31] RC seems unequipped to deal with the real world.

I think that the DWO is the most serious objection facing RC. The rest of this Chapter looks at two ways a proponent of RC might try to avoid the DWO, and it explains why each attempt fails.

3.3 Unsuccessful Solutions

[29] Parfit (2011) p.321-20 as found in Podgorski (2018) p.281
[30] As noted by Podgorski (2018) p.281
[31] Podgorski (2018) p.281-2

As mentioned, this Section will look at 2 unsuccessful attempts to resolve or avoid the DWO. First, changing the level of compliance or acceptance at which you evaluate how well rules make things go (**3.1**). Second, by using conditional rules of the form '*follow R, unless X, in which case…'* (**3.2**).

3.3.1 Change the Level of Compliance

As noted, the DWO is a specific formulation of the Ideal World Objection. It accuses RC of basing what we ought to do on things that are too "*far away"*, modally speaking.[32] The reason that rules like '*never be violent'* have disastrous real-world consequences is that some people are violent. There will never be universal compliance with or acceptance of *'never be violent'*, and this should be relevant to our moral decision-making. So, some proponents of RC have suggested that the source of the problem is that RC looks at *ideal* worlds, where there is an unrealistic level of adherence to rules. The solution, they propose, is to change RC's *degree of ideality* — to evaluate rules by looking at worlds with more realistic levels of rule-adherence.[33]

I will now argue that this proposed solution fails because it misdiagnoses the problem; RC's proponents misunderstand the force of the IWO. The problem is not the degree of ideality. It is that a core feature of RC is analysing worlds that we can never realise. Any view that does this will be sensitive to irrelevant facts about those worlds and insensitive to important facts in the real world.[34] Following Podgorski, I will show that changing RC's degree of ideality fails to resolve the DWO. I do so by introducing 4 versions of RC that change its degree of

[32] Arneson (2005); Portmore (2009) as found in Podgorski (2018) p.289
[33] Podgorski (2018) p.282
[34] Podgorski (2018) p.286-7

ideality. Then, I use small changes to the utility landmine case to show that these views will cause you to violate at least one of (1)-(3).[35] In doing this, I follow Podgorski's paper, but provide greater detail as to why each version of RC fails to resolve the problem.

Since they change the degree of ideality in RC, each of these versions of RC can be described as concerned with *partial* adherence to rules rather than *universal* adherence to rules. Here, like in (**2.2**), there are multiple ways that rules can relate to the world — they can be partially complied with, or partially accepted. Accordingly, we can introduce some new versions of RC: Partial Compliance Rule Consequentialism (PCRC) and Partial Acceptance Rule Consequentialism (PARC). All 4 *partial* versions of RC described in this Section can be formulated as either a version of PCRC or as a version of PARC. The distinction between PCRC and PARC is not particularly relevant for (**3.3.1**) — all 4 partial versions of RC will fail to avoid the DWO regardless of whether they are complied with or accepted. So, for brevity, I use the neutral term, *adherence*, in this Section, rather than *compliance* or *acceptance*. This allows us to discuss both compliance and acceptance versions of RC at once.

(i) *Fixed-Rate Rule Consequentialism*

Some philosophers have responded to the IWO by proposing what we can call Fixed-Rate Rule Consequentialism (FRRC). Richard Brandt and Brad Hooker have proposed such views.[36] FRRC evaluates actions by comparing rules that are adhered to not perfectly, but rather at a fixed rate under 100%. No specific rate is required, but Hooker recommends 90%

[35] As done by Podgorski (2018) p.287-8
[36] Brandt (1992); Hooker (2000); as found in Podgorski (2018) p.283

as the fixed rate.[37] Taking 90% as our fixed rate, we can formulate FRRC as: *an act, A, is morally permissible if and only if it accords with a set of rules, S, such that if 90% of people adhered to S, the consequences would be at least as good as if 90% of people adhered to any set of rules other than S.*

This view avoids the original utility mines example — if 90% of people went outside to do a jumping jack at midday a bad-mine would be triggered. Thus, according to FRRC, we ought not to follow *R*. We can also see that such a view would not imply that we ought to follow the rule '*never use violence'* — if 90% of people adhered to that rule, and 10% did not, then that would presumably lead to disastrous results.

However, the DWO persists. For we can simply change the trigger conditions for our utility mine. Consider again the rule, *R*, that says 'do a jumping jack at noon every day', and two new utility mines:

1. A good-mine that triggers when there is 90% adherence to *R*; and
2. A bad-mine that triggers when 1 or more people adhere to *R*, but there is not 90% adherence.

Suppose you know that less than 90% of people will do a jumping jack at noon today. Should you nevertheless follow R? Of course not, for that would lead to disaster. But FRRC implies that you should! This is because, in a world where 90% of people adhered to this rule, the good mine would be triggered. In this Case, FRRC implies that you ought to follow R, and that the only morally permissible action available to you is to go outside and do a jumping

[37] As noted by Podgorski (2018) p.283

jack. Irrelevant facts about other possible worlds continue to determine what is morally permissible in our world. FRRC thus violates at least (1) and (3). So, this solution fails.

(ii) Optimum-Rate Rule Consequentialism

Some philosophers have responded to the IWO by proposing what we can call Optimum-Rate Rule Consequentialism (ORRC). Holly Smith has proposed (but not endorsed) this view.[38] ORRC is harder to provide a concise formulation for. Loosely, we can formulate it as: *an act, A, is morally permissible if and only if it accords with a set of rules, S, such that the consequences of adherence to S at S's 'best' level of acceptance would be at least as good as the consequences of adherence to any other set of rules at the 'best' level of adherence for these alternative rules.* In other words, imagine: Rule A makes things go better at 80% adherence than at any other rate of adherence; Rule B makes things go better at 95% adherence than at any other rate of adherence; Rule A at 80% adherence makes things go better than Rule B at 95% adherence. If this were the case, then ORRC implies we ought to pick Rule A over Rule B.

But ORRC fails in even our original formulation of the utility mine case. 100% adherence to *R* triggers a good-mine; it would make things go better than any rate of adherence to any other rule. Thus, ORRC implies that we ought to accept R — the only morally permissible action available to us is to go outside and do a jumping jack. Again, irrelevant facts about other worlds continue to determine the permissibility of actions in the actual world; ORRC will cause you violate one of (1)-(3). This solution also fails.

[38] Smith (2010), as found in Podgorski (2018) p.283

(iii) Average-Rate Rule Consequentialism

Other philosophers have responded to the IWO by proposing what we can call Average-Rate Rule Consequentialism (ARRC). Michael Ridge endorses this view.[39] Upon ARRC, we look at multiple rates of adherence and follow the rules that do best on average across adherence rates. We can formulate it as: *an act, A, is morally permissible if and only if it accords with a set of rules, S, such that — if you take an average of how well S makes things go at every adherence rate — the average would be at least as high as the corresponding average for any other set of rules.*

This view avoids our original counterexample — at nearly every possible level of compliance, and thus in nearly all possible worlds, following *R* results in unimaginable suffering. So, if we take an average of how well S makes things go at every adherence rate, the average will be very low. RC thus implies that one ought not to follow *R*.

However, this solution to the problem fails if we change the case. Imagine there is a weak bad-mine — rather than releasing an unimaginable amount of suffering, it releases a serious but not astronomical amount. By changing the trigger conditions and nature of the mines, the issue with ARRC is exposed. Imagine:

1. A strong good-mine that triggers when there is 80% or more adherence to *R*; and
2. A weak bad-mine that triggers when there is under 80% adherence to *R.*

[39] Ridge (2006), as found in Podgorski (2018) p.283

On average, the unimaginable amount of good produced at every rate of adherence over 80% would massively outweigh the serious, but not astronomical, amount of bad produced at every rate of adherence below 80%. R will thus perform better, on average, than any alternative rule. ARRC thus implies that you ought to follow *R*, even though you know this will lead to seriously bad consequences in the real world. This is a third view where irrelevant facts about other worlds continue to determine what is morally permissible; ARRC thus violates at least (1) and (3).

(iv) Every-Rate Rule Consequentialism (ERRC)

The final version of RC to look at is Every-Rate Rule Consequentialism (ERRC). Derek Parfit has proposed this view.[40] ERRC states that we ought to act according to those rules that do best at every level of adherence. We can formulate it as: *an act, A, is morally permissible if and only if it accords with a set of rules, S, such that at every rate of adherence, the consequences of S would be at least as good as any set of rules other than S.*

ERRC also avoids the original utility mine example: at every level of adherence other than 100% and 0%, rule *R* has disastrous consequences. So, ERRC implies that you ought not go outside and do a jumping jack at noon.

This view, though, is very unlikely to provide any moral guidance at all: if there is no set of rules that is best at all levels of adherence, this view fails to recommend or condemn any actions at all. Placing good-mines for incompatible rules at different levels of adherence guarantees this result. And, in the real world, there are few rules that would do best at every

[40] Parfit (2011) p.317 as found in Podgorski p.283

level of adherence. For example, *'never use violence'* has disastrous consequences at many levels of adherence. *'Be kind to people'* would presumably have good consequences at many levels of adherence, but worse consequences at 0.0001% adherence than '*be unkind to people'*. A moral theory that cannot provide action guidance is clearly unsatisfactory. Further, ERRC still allows irrelevant facts about other worlds to determine actions; it causes you to violate (3).

FRRC, ORRC, ARRC, and ERRC all fail to avoid the DWO. As Podgorski emphasises, the degree of ideality in RC is not the problem.[41] To suggest so misunderstands the force of the objection. The problem is evaluating worlds that differ from ours in more than what is up to us. And this is a *core*, *indispensable* feature of RC.

More generally, as Podgorski emphasises, *any view which determines what we ought to do by evaluating worlds that differ from ours in more than what is up to us* will cause us to violate at least one of (1)-(3).[42]

Let us now turn to another attempt to escape the IWO, and again see why it fails.

3.3.2 'Follow R, unless X, in which case…'

Presumably, the rule *'never use violence'* will make things go best in cases where nobody else is using violence. But it will make things go badly when others use violence. So, a proponent of RC might suggest the rule *'never use violence, unless someone else does, in*

[41] Podgorski (2018) p.284
[42] Podgorski (2018) p.290

which case use as little violence as possible in pursuit of your ends' However, the use of conditional rules (i.e., rules of the form *'follow R, unless X, in which case…'*) does not avoid the DWO. In the rest of this Section, I go through each of the versions of RC introduced in this Chapter and show how conditional rules cannot help any to avoid the DWO.

I have discussed four versions of RC. In (**2.2**)**,** I introduced Universal Compliance and Universal Acceptance versions (UCRC and UARC). I will first show that each of these versions of RC cannot use conditional rules to avoid the DWO. Podgorski does not discuss it, but a proponent of RC might look for a solution to the DWO by using conditional rules on one of the partial versions of RC introduced in (**3.3.1**): Partial Compliance Rule Consequentialism (PCRC) and Partial Acceptance Rule Consequentialism (PARC). I will show that using conditional rules cannot help PCRC or PARC avoid the DWO either.

(i) Universal Compliance Rule Consequentialism

Consider these two conditional rules: *'never use violence, unless someone else does, in which case use as little violence as possible in pursuit of your ends'*, and *'never use violence, unless someone else does, in which case kill as many people as you can in pursuit of your ends.'*[43] This second rule is morally repugnant. But UCRC implies that each of these rules has the same moral status as *'never use violence'*.[44] This is because, when everyone complies with each of these rules, nobody ever uses violence. UCRC thus implies that actions which adhere

[43] Podgorski (2018) p.286; Parfit (2011) p.315

[44] 'Does universal acceptance imply universal compliance?' becomes relevant here (see fn20). If universal acceptance does not imply universal compliance, then the two rules are equivalent upon universal compliance. But if universal acceptance implies universal compliance, then the two rules are not equivalent: when we assess all possible results of universal compliance with a rule, some of those results will include the universal acceptance of that rule. Universal acceptance of these two rules is not equivalent. This does not save RC from the Distant World Objection, though — we can alter the trigger conditions of the utility mines so that the mines trigger only if there is universal compliance and *not* universal acceptance of the rule.

to *'never use violence, unless someone else does, in which case kill as many people as you can in pursuit of your ends'* are morally permissible. If this rule was universally complied with, the result would be the same as the results of universal compliance with the rule *'never use violence'* — the second part of the rule is inert.[45] In the real world, though, on this view, when confronted with violence, it becomes morally permissible to kill as many people as you can in pursuit of your ends. A moral theory that makes it morally permissible to kill as many people as you can is deeply flawed. UCRC cannot use conditional rules to escape the DWO.

This problem only affects UCRC. Universal (or partial) acceptance of *'never use violence', 'never use violence, unless someone else does, in which case use as little violence as possible in pursuit of your ends'* and *'never use violence, unless someone else does, in which case kill as many people as you can in pursuit of your ends'* are *not* equivalent — the acceptance of each rule might have different psychological effects for agents. And PCRC can also avoid this problem. In PCRC, it is built into the assessment of a rule that some people will not comply with it; the second part of the rule can come into effect. For each of these views, though, the DWO persists, even when using conditional rules. Let us look next at UARC, before looking at the partial views.

(ii) Universal Acceptance Rule Consequentialism

To see that UARC cannot use conditional rules to avoid the DWO either, consider again the rule, *R*, that says 'do a jumping jack at noon every day', and two new utility mines:

[45] Where 'unless X' is used to refer to noncompliance with R. If the rule was 'follow X, unless it is raining', then the second clause would not be inert (but fairly arbitrary).

1. A good-mine that triggers if and only if there is universal acceptance of the *unconditional* rule, *R*; and
2. A bad-mine that triggers if and only if 1 or more people accept the *unconditional* rule, *R*, but there is not universal acceptance of this rule.

In this scenario, the unconditional rule *R* will trigger a good-mine when universally accepted. Any conditional rule, such as *R**, which says 'do a jumping jack at noon every day, unless doing so would trigger a bad-mine, in which case do not', would not trigger a good-mine when universally accepted. Universal acceptance of *R* makes things go better than *R**. Accordingly, UARC would imply that the only morally permissible acts available to you are those that accord with the unconditional *R*. Acts which accord with the conditional *R** would be morally impermissible.[46] So, we can see that UARC cannot avoid the DWO by using conditional rules — the utility mine case makes it such that unconditional rules make things go better than conditional ones, and thus acts which accord with conditional rules will not be permissible. The only available act available to you under UARC is still to do a jumping jack in accordance with *R* and, when others in the actual world fail to do a jumping jack, trigger the bad-mine.

(iii) Partial Compliance Rule Consequentialism

In (**3.3.1**), we went through four different partial adherence versions of RC, and I showed how each fails to avoid the DWO. In a similar manner, all 4 versions of PCRC from (**3.3.1**) fail to avoid the DWO by using conditional rules. Rather than go through all 4, let us just look at the first, Fixed-Rate Rule Consequentialism (FRRC) to see how the combination of

[46] Unless that action also adheres with R

partial compliance and conditional rules still fails to avoid the DWO — doing so should make clear that the other 3 versions of PCRC will also fail to avoid the DWO.

Recall that FRRC evaluates actions by comparing rules that are not followed perfectly, but at a fixed rate under 100%. No specific rate is required, but Hooker recommends 90% as the fixed rate.[47] Taking 90% as our fixed rate, we formulated the compliance version of FRRC (or, Fixed-Rate Partial Compliance Rule Consequentialism) as: *an act, A, is morally permissible if and only if it accords with a set of rules, S, such that if 90% of people complied with S, the consequences would be at least as good as if 90% of people complied with any set of rules other than S.*

Consider again the rule, *R*, that says 'do a jumping jack at noon every day'. Consider, also, the conditional rule, R*, that says 'do a jumping jack at noon every day, unless someone else does not do a jumping jack at noon, in which case do nothing'. Unlike the universal compliance version of RC, the partial compliance version does not imply that these two rules are equivalent. Recall that in (**3.3.2 (i)**) we saw that the second half of a conditional rule like 'follow R, unless X' is inert under UCRC. When there is universal compliance, nobody breaks the rules — the 'unless X' is never used. So, *R* and *R** make things go equally well according to UCRC. That is not the case for PCRC, as imperfect compliance with a rule is built into the assessment of that rule. 'Do a jumping jack at noon every day' and 'do a jumping jack at noon every day, unless someone else does not do a jumping jack at noon, in which case do nothing' will produce difference results when partially complied with.[48]

[47] Hooker (2000), as noted by Podgorski (2018) p.283

[48] There is a strange world in which R and R* produce the same results. If the 10% of non-jumpers were isolated away from the 90% of jumpers, no jumper would see someone not jumping. So, the second half of R* would be inert, and partial compliance with R* would be extensionally equivalent to partial compliance with R. We do not, though, live in that world.

But this does not help PCRC avoid the DWO. Imagine there are two utility mines:

1. A good-mine that triggers when 90% of people comply with *R*; and
2. A bad-mine that triggers when 1 or more people comply with *R*, but there is not 90% compliance.

In a world where 90% of people complied with *R*, the good mine would be triggered. Thus, there is no rule which makes things go better than *R*. So, acts that accord with *R* will be morally permissible. PCRC also implies that acts that accord with *R** are morally impermissible — 90% compliance with *R** will not trigger a good-mine, so will not make things go as well as 90% compliance with *R*. Thus, PCRC implies that only acts which accord with *R* are morally permissible. In the actual world, 90% of people would not go outside and do a jumping jack. Thus, following PCRC would cause you to trigger the bad-mine.

So, using conditional rules cannot help PCRC escape the DWO. Combining conditional rules with our other partial compliance views (ORRC, ARRC, ERRC) would also not help PCRC avoid the DWO.

(iv) Partial Acceptance Rule Consequentialism

The partial acceptance version of FRRC is very similar to the partial compliance version: *an act, A, is morally permissible if and only if it accords with a set of rules, S, such that if 90% of people accepted S, the consequences would be at least as good as if 90% of people accepted any set of rules other than S.*

Again, by introducing two utility mines we can see that it becomes morally impermissible to accept R* over R:

1. A good-mine that triggers when there is 90% acceptance of the unconditional rule *R*; and
2. A bad-mine that triggers when 1 or more people accept the unconditional rule *R*, but there is not 90% acceptance.

In this scenario, the acceptance version of FRRC implies that it is morally permissible to accept R, and morally impermissible to accept R* — 90% acceptance of R would make things go better than 90% acceptance of R*. 90% acceptance of R would trigger a good-mine. 90% acceptance of R* would not trigger a good-mine, as R* is a conditional rule. Accordingly, the acceptance version of FRRC would imply that the only morally permissible acts are those that accord with the unconditional rule, R, and that it is morally impermissible to act in a way that accords with any conditional rule like 'Follow R, unless others violate R, in which case…'. So, the acceptance version of FRRC cannot avoid the DWO by using conditional rules. Hopefully it is again clear that our other partial acceptance views (ORRC, ARRC, ERRC) would also not avoid the DWO by using conditional rules.

None of the versions of RC discussed in this Chapter — UCRC, UARC, PCRC, or PARC — can avoid the DWO by using conditional rules. These attempts to avoid the IWO fail because they misunderstand the force of the objection. Even when you change the degree of ideality, or introduce conditional rules, RC will still cause you to either: do something avoidably

disastrous; or avoid doing something wonderful without good reason; or allow irrelevant facts about distant worlds to determine what we ought to do.

Conclusion

The DWO seems a serious problem for RC which cannot easily be avoided. A moral theory presumably should not permit you to do something avoidably disastrous or permit you to avoid doing something wonderful without a very strong reason. It should not allow irrelevant facts about faraway possible worlds to determine what you ought to do. I have shown that two different attempts to avoid the DWO fail. I have argued that this is because they misinterpret the *force* of the objection. The DWO, though, affects more views than just RC. It affects *any view which determines what we ought to do by evaluating worlds that differ from ours in more than what is up to us*.

In the next Chapter, I look at how the DWO affects Kantian Ethics. In doing so, I show that — just like with RC — sophisticated versions of Kantian Ethics that aim to avoid the IWO fail to do so. I do this by focussing on Korsgaard's (1986) paper, *"The Right to Lie: Kant on Dealing with Evil."*

4. Kantian Ethics

In this Chapter, I discuss how the DWO affects Kantian Ethics. I do this for two reasons. First, to show that the DWO affects more moral theories than just RC — this begins to show the scope of the IWO. Second, to discuss another attempted solution to the IWO. Korsgaard proposes a version of Kantian Ethics which purports to deal with its problem of idealism. This can reasonably be viewed as an attempt to resolve the IWO. I show that this attempt fails, because Korsgaard has misunderstood the force of the IWO.

In (**4.1**), I explain Kantian Ethics. In (**4.2**), I explain when you can (and cannot) rationally will the universal acceptance of a principle. (**4.3**) provides 3 reasons why one might accept Kantian Ethics. I then show how the DWO affects Kantian Ethics (**4.4**), before setting out Korsgaard's attempt to solve Kantian Ethics' idealism (**4.5**). (**4.6**) shows why this attempt fails to resolve the DWO.

4.1 What is Kantian Ethics?

Kantian Ethics (KE) is a deontological moral theory in which one must act according to the Categorical Imperative. The Categorical Imperative is a synthetic, a priori, moral law.[49] It is a priori because it can be known through reason alone, but synthetic because it is prescriptive — it provides normative instructions, not analytic truths.[50] Kant's first formulation of the Categorical Imperative is the Formula of Universal Law (FoUL): *act only according to that maxim through which you can at the same time will that it become a universal law".*[51]

[49] Korsgaard (2012a) xii
[50] DePaul & Hicks (2021)
[51] Groundwork 4:421 / 34.

The correct interpretation of the FoUL is heavily debated.[52] One plausible way to restate it is as follows:

> (KE1): *an act is morally permissible if and only if such acts are permitted by some principle whose universal acceptance everyone could rationally will*.[53]

There are two significant differences between the FoUL and (KE1) that I need to explain before we can proceed. First, the FoUL appeals to what one 'can… will' to become a universal law, whereas (KE1) appeals to what one can '*rationally* will' to become universal law. I assume that this is what Kant intended, rather than, say, what one could physically or psychologically will to become universal law. (**4.2**) explains this further when explaining what it means to 'rationally will' something.

Second, the FoUL appeals to what *you* could will to become a universal law, whereas (KE1) appeals to what *everyone* could will to become a universal law. I am following Parfit in modifying Kant's FoUL in this way.[54] This is an important difference — whilst I might be able to rationally will the maxim 'eat vanilla ice cream over chocolate ice cream', chocolate-lovers could not. This modification also allows us to see the parallels between the way Kantian Ethics and Rule Consequentialism face the DWO. Ultimately, though, the DWO would still apply to the version of Kantian Ethics which appeals only to that which *you* could will to become universal law.

---

[52] Podgorski (2018) p.291
[53] This is also the formulation of Kantian Ethics that Parfit finds most plausible: Parfit (2011) p.407; Rosen (2009) p.78
[54] See Parfit (2011) p.407

Kant provides further formulations of the Categorical Imperative which will be discussed in (**4.4**). But he believed those formulations to be equivalent with the FoUL — he believed that you could derive each formulation from the others.[55] So, we can ask which acts Kantian Ethics might permit without needing to state these other formulations, by using (KE1). If an act accords with a maxim — a "subjective principle of acting"[56] — which one can will to be a universal law, then the act is morally permissible. Acts which accord with maxims which one cannot will to be a universal law are morally impermissible.[57]

So, Kant rejects the foundational claim of AC: the moral permissibility of an act is *not* determined by how well that act would make things go if performed.[58] Instead, the moral value of an act is determined by the maxim upon which one acts.[59]

(KE1) does somewhat resemble RC. Unlike RC, though, KE is committed to universal acceptance. There could not be a universal compliance version of KE: Kant notes explicitly that — if an act is to have moral worth — it must be done *because* it is one's moral duty.[60] There is no moral worth in merely complying with one's duty. So, for an act to have moral worth, it must be that the agent accepts the maxim, and acts from it. There could also not be a partial compliance or partial acceptance version of KE. The moral law applies to every rational agent.

[55] Korsgaard (2012a) fn11
[56] Groundwork 4:401 / 16; McCarty (2009) p.3
[57] Korsgaard (2012b) p.79. Acts which fail this universalisation test are not morally permissible *because* they fail it. Instead, they fail because they are morally impermissible.
[58] Groundwork 4:394 / 10; Korsgaard (2012a) xiv; Larry & Moore (2021)
[59] Groundwork 4:400 / 15; Korsgaard (2012a) xiv. Note that moral value here is a term that tracks rightness, rather than permissibility.
[60] Groundwork 4:399 / 14

There is another important difference between Universal Acceptance RC (UARC) and KE. Upon UARC, only acts which accord with the set of rules that would make things go best if universally accepted are morally permissible. Contrarily, upon (KE1), only acts which are done from those maxims which one could rationally will to be universally accepted are morally permissible. There is a difference between the optimal set of universal rules and the set of universal rules that can be rationally willed. The next Section will discuss this difference further.

I appreciate that some might contest this construction of Kantian Ethics.[61] To do so would not undermine the argument of this Chapter. Whatever the 'correct' interpretation, (**4.4**) will show that the DWO applies to Kantian Ethics in general.[62]

To know what (KE1) implies, we need to know when everyone can (and cannot) rationally will that a principle be universally accepted.

4.2 What is it to 'Rationally Will'?

That the Categorical Imperative is synthetic and a priori limits what can be willed as a universal law. A maxim that is based upon personal inclination (for example, *choose cookies over ice cream*) could not be rationally willed by everyone — not every person would be able to arrive, a priori, to that conclusion.[63] Instead, a maxim can be rationally willed if one can will that it become a universal law *without contradiction.*[64]

[61] See Ebells-Duggan (2011) for a summary of differing interpretations of Kantian Ethics
[62] As noted by Podgorski (2018) p.291
[63] Korsgaard (2012b) p.77
[64] Groundwork 4:424 / 37; Korsgaard (2012b) p.77

Here, then, we can further differentiate (KE1) from UARC — being able to will that something become a universal law without contradiction does not require that its universal acceptance is optimal. It just requires that it does not lead to a contradiction. (KE1) and UARC are clearly distinct. But in what sense can the universal acceptance of a rule lead to contradiction?

For Kant, there are two ways in which willing a maxim to be a universal law can lead to contradiction. First, contradiction in conception: there is some form of *"internal impossibility"* found in the willing of the maxim.[65] The universalisation of the maxim leads to contradiction because we cannot conceive of a world in which everyone accepted that maxim. For example, the maxim 'make false promises' leads to a contradiction in conception — if everyone accepted this maxim, then nobody would ever believe a promise. If promises cannot be believed, the practice of promising loses its purpose. So, promises would not exist. So, there cannot be a practice of promising for one to violate in the first place. Kant calls this a contradiction in conception.

Second, contradiction in the will: universal acceptance of the maxim would frustrate the maxim's very aim.[66] For example, the maxim, 'lie when it benefits you'. If this maxim were universally accepted, then your friend who is moving house this weekend would never believe your lie that you are attending a cousin's dog's wedding. The universalisation of the maxim frustrates the purpose for which the maxim was employed.

[65] Groundwork 4:424 / 37; Korsgaard (2012b) p.78
[66] Groundwork 4:424 / 37; Korsgaard (2012b) p.78

So, to be able to rationally will a principle, that principle's universal acceptance must not lead either to a contradiction in conception, or a contradiction in the will. Within Kantian moral scholarship, there is further discussion over how to interpret the notion of contradiction.[67] For our purposes, though, this understanding of contradiction is sufficient — Section (**4.4**) will show that there are cases in which an act will avoid any form of contradiction, whilst causing an agent to either:

(i) Do something avoidably disastrous; or

(ii) Avoid doing something wonderful without good reason; or

(iii)Allow irrelevant facts about distant worlds to determine what we ought to do.

So, we can show that the DWO affects Kantian Ethics without needing to further discuss interpretations of contradiction. Before looking at how the FoUL will cause one to do at least one of (i), (ii), or (iii), let me provide some reasons why one might accept Kantian Ethics.

4.3 Why Accept Kantian Ethics?

KE is another *prima facie* plausible moral theory. It shares many of the strengths that RC had. Here I outline 3.

In much the same way as RC, KE captures some of our powerful moral intuitions about what makes an act morally impermissible. KE also asks one to think, *"what if everyone thought or acted like you do?"* So, KE also has great explanatory power when it comes to *why* an act is morally impermissible.

[67] As discussed in Korsgaard (2012b)

KE also plausibly suggests that morality consists of a set of rules (or maxims) that ought to govern our behaviour. But KE's rules are found through reason, whilst RC's rules are found in an assessment of acts' universal acceptance or compliance. It is plausible that we ought to follow the set of rules that everyone could rationally will. Thus, like RC, the Kantian project seems well motivated and bears a plausible relation to rationality.

Again, like RC, KE also provides clear and intuitive action-guidance in a range of cases where AC struggles to provide such guidance. Recall our case from (**1.5.1**), of David the organ-stealing surgeon — although the act would make things go best, we believe stealing the organs of healthy patients to be morally impermissible. KE can provide an explanation: the universal acceptance of the maxim 'never kill your healthy patients to harvest their organs, unless it makes things go better and no one ever finds out' would lead to a contradiction in the will — by willing this to be the case, you would frustrate your aim, as, if everyone accepted this maxim as a universal law, nobody would go to their doctor.

KE also has strengths which RC does not. Whilst RC can at best give some instrumental weight to rights and duties — as discussed in (**2.4.1**) — Kantian Ethics can provide them with fundamental value. Rights and duties are one of the things which a person can realise has value a priori. Similarly, Kant can provide fundamental value to special relationships: 'give reasonable partiality to those you love over strangers' is a maxim that could presumably be rationally willed by all. Also, it is highly plausible that the moral value of an act is separable from the utility it produces. Due to this separation, Kant can avoid the issues raised by luck and uncertainty discussed in the context of Consequentialism in (**1.4**) and (**2.1**). This is because the moral permissibility of an act does not come from its expected results, but instead

from the fact that acting in such a way is required by a principle one could accept as their duty. So, Kant's Ethic has various independent strengths too.

There are various objections to KE that are deserving of their own thesis. Our focus, though, will remain on the DWO.

4.4 How Does the DWO Affect Kantian Ethics?

As a reminder, the DWO is the following objection: any view which determines what we ought to do by evaluating possible worlds that differ from the actual world in more than what is up to us will cause us to:

(i) Do something avoidably disastrous; or

(ii) Avoid doing something wonderful without good reason; or

(iii) Allow irrelevant facts about distant worlds to determine what we ought to do.[68]

(KE1), as stated in (**4.1**), goes as follows: *an act is morally permissible only if such acts are permitted by some principle whose universal acceptance everyone could rationally will.* It clearly evaluates possible worlds that differ from the actual world in more than what is up to us — we do not have the power to bring about the universal acceptance of a given principle. So, KE seems a prime candidate to face the DWO. Now, I will provide an example using Podgorski's language of utility mines, in order to show the structural issue.

[68] Podgorski (2018) p.290

Recall the original set-up for the utility mine case: we can imagine a world there is a rule, *R,* that says, 'do a jumping jack at noon every day' and you know that there exist two mines:

1. A good-mine that triggers when there is universal acceptance of *R*; and
2. A bad-mine that triggers when 1 or more people accept *R*, but there is not universal acceptance.

So long as the universal acceptance of *R* does not lead to contradiction, (KE1) will imply that acts which are permitted by *R* are morally permissible. So, the good-mine's structure does not need changing. But there are two changes that need to be made to this case for KE. First, *R* needs to be something that every rational being could arrive at a priori — we need a rule less arbitrary than 'do a jumping jack at noon every day'. Second, because the rule will be something that everyone can arrive at, we can assume that 1 or more people have already accepted the rule. Accordingly, we must change the trigger conditions of the bad-mine, so that it will only be triggered if *you* are moved to act by your assessment of distant worlds. So, we can restate the scenario:

We can imagine a world where there is a rule, *T*, that says 'tell the truth', and you know that there exist two mines:

1. A good-mine that triggers when there is universal acceptance of *T*; and
2. A bad-mine that triggers when *you* accept *T*, but not everyone does.

It is highly plausible that 'tell the truth' is a maxim which could be discovered a priori by every rational being. It also satisfies (KE1) — everyone can rationally will that 'tell the truth'

be universally accepted. And (KE1) will be entirely insensitive to the existence of the bad-mine. It does not matter that you will trigger a bad-mine by following *T* and telling the truth — (KE1) says that it is morally permissible to tell the truth. Since Kant believes that the universal acceptance of rules which permit lying leads to contradiction, then telling the truth is also the only act which is morally permissible. Therefore, (KE1) implies that you are morally obligated to trigger the bad-mine. This is an example where (i) and (iii) are violated — KE causes us to do something avoidably disastrous *because* we allow irrelevant facts about distant worlds to determine what we ought to do.

This utility mine example is very similar to Kant's Murderer at the Door case[69] — the case that motivates Korsgaard's attempted solution to the DWO. In each case, if *you* act in a way that can be rationally willed to be a universal law (tell the truth), the outcome will be disastrous. Contrarily, if you act in a way that would lead to a contradiction if willed to be a universal law (tell a lie), the outcome will be much better. The case is as follows.

> Case 11: Suppose you permit a friend to hide from a murderer in your house. Suppose further that the murderer comes to your door and asks if you know where the friend is. Imagine you are certain of the following consequences: if you tell the murderer the truth, he will kill your friend; if you lie to the murderer, your friend might survive. So, you have two options: act from *T* (tell the truth), leading to an avoidably disastrous result of your friend being murdered; or act from ¬*T* (lie), giving your friend a chance of survival.

[69] SRL 8:425 / 611

According to Kant, one must not lie, "*however great the disadvantage to him or to another that may result from it*."[70] Kant claims it is morally impermissible to lie to the Murderer at the Door. This is highly counterintuitive, and a clear example of the DWO. By deciding what you ought to do by considering whether your act could be rationally willed as universal law, you end up needlessly (and wrongly) helping someone murder your friend. We do something avoidably disastrous *because* we allow irrelevant facts about distant worlds to determine what we ought to do.

Case 11 is also a particularly interesting example of the DWO: it is different to the examples traditionally used to explain the objection, such as in (**3.1**). In those cases, an agent acts in accordance with a rule like *T*, which one could rationally will to be a universal law (or, in RC's terms, would make things go best if everyone accepted). But because not everyone accepts *T* in the real world, the agent's act has disastrous consequences. In Case 11, though, an agent acts in accordance with a rule, *T*, which one could rationally will to be a universal law (or, in RC's terms, would make things go best if everyone accepted), and *even in the world where everyone accepts T*, the agent's act has disastrous consequences. Such cases are often overlooked in the existing literature on the IWO. And they require attention — we know that *T* is permissible upon the FoUL. How does this happen?

The confusion arises because *T* is a singular rule among many. The FoUL implies that acting from *T* is morally permissible, and that acting ¬*T* is morally impermissible. But consider another rule, *K*, which says 'do not kill', and a contrasting rule ¬*K*, which says 'kill when it furthers your ends'. *K* would be morally permissible upon the FoUL — it would not lead to a contradiction. ¬*K* would be morally impermissible upon the FoUL — it would lead to a

[70] SRL 8:426 / 612

contradiction in the will. Your aim of murdering others to pursue your ends would be frustrated by the maxim's universal acceptance, as you would likely be murdered when it served someone else's ends. So, the FoUL implies that acting from T and/or K is morally permissible, but that acting from $\neg T$, $\neg K$, or from $\neg T$ and $\neg K$, is morally impermissible. The DWO applies in Case 11 because (despite there being universal acceptance of T) the murderer is not acting from K. The DWO arises for KE because we consider which *set of rules* we could rationally will, and this makes us insensitive to real-world facts about compliance to that set of rules.[71]

Now, let us look at how Korsgaard tries to resolve the Murderer at the Door case.

4.5 How Does Korsgaard Try to Avoid the DWO?

Korsgaard's solution to the Murderer at the Door case is not a direct response to Podgorski's DWO — Korsgaard's paper was published 32 years prior! So, it would be anachronistic to present Korsgaard as trying to solve the DWO. But she does frame her paper in terms of the IWO: she is attempting to resolve a problem with KE, that "it seems to imply that our moral obligations leave us powerless *in the face of evil*."[72] She writes that "Kant's theory sets a high ideal of conduct and tells us to live up to that ideal *regardless of what other persons are doing*. The results may be very bad."[73] This is essentially a statement of the IWO.

[71] Recall that RC also asks an agent to consider the set of rules, S, rather than a specific rule
[72] Korsgaard (1986) p.325
[73] Korsgaard (1986) p.325

So, if her attempt to deal with the IWO fails to resolve the DWO, because it misinterprets why Kantian Ethics faces the IWO in the first place, then we can say that Korsgaard has misunderstood the force of the IWO. This Section sets out her attempted escape.

Korsgaard makes 3 points in her argument to deal with the IWO:

1. It is morally permissible to lie to the Murderer at the Door upon the FoUL, but not upon the other formulations of the Categorical Imperative.
2. Therefore, Kant's formulations of the Categorical Imperative are not equivalent — the FoUL is more permissive than the other formulations.
3. This allows us to generate special Kantian principles for dealing with evil, which are sensitive to the real world.[74]

Since the second claim follows from the first, the rest of this Section sets out Korsgaard's reasoning behind the first and third claims.

4.5.1 The FoUL Permits Lying to Murderers, the FoH and KoE Do Not

To show this, I must first provide the other formulations of the Categorical Imperative: the Formula of Humanity (FoH) and the Kingdom of Ends (KoE).[75] Then I will show that lying to the Murderer at the Door is morally permissible upon the FoUL, but not upon the FoH or KoE. It will then follow that the FoUL is more permissive than the FoH and KoE. Recall the

[74] Korsgaard (1986) p.327
[75] Korsgaard does not discuss the Formula of Autonomy, presumably because Kant himself never explicitly provided it as a formulation of the Categorical Imperative. I also omit it for this reason. See Johnson & Cureton (2022) §7

FoUL: *an act is morally permissible only if such acts are permitted by some principle whose universal acceptance everyone could rationally will.*

Kant states the FoH as follows: "[act] so that you treat humanity, whether in your own person or in that of another, always as an end and never as a means only."[76] Korsgaard argues that we can restate this as: *an act is morally permissible if and only if those affected by the act could possibly assent to this mode of acting.*[77] This is different from treating someone in a way that they would never assent to. To draw out this difference, let us see why lying is morally impermissible upon the FoH. Lying is not morally impermissible because a person would never agree to be lied to. Instead, lying is morally impermissible because a person *could* never agree to being lied to — a lie is only successful if the other person does not know that it is a lie.[78] If you agreed to being lied to, it would become impossible to be lied to, as being deceived is a necessary part of being lied to. So, lies are morally impermissible upon the FoH.

Kant states the KoE: "act in accordance with the maxims of a member giving universal laws for a merely possible kingdom of ends."[79] Johnson and Cureton suggest formulating this as: *an act is morally permissible if and only if it accords with a principle that could be accepted by a community of fully rational, morally legislating individuals.*[80] In other words, act always as if you were living in a Kingdom of Ends.[81] Here, lying is morally impermissible because a community of fully rational, morally legislating individuals would not accept a principle in

[76] G 429 / 41; Korsgaard (1986) p.330
[77] Korsgaard (1986) p.331
[78] Korsgaard (1986) p.333. In the case where one knows that they are being lied to, we can say that the lie still fails – its aim is to deceive. By failing to be deceived, you have not been lied to.
[79] G 4:439 / 50, as found in Johnson & Cureton (2022) §8
[80] Johnson & Cureton (2022) §8
[81] Korsgaard (1986) p.344

which lying is acceptable. Lying to someone, in effect, is an attempt to disqualify them as a member of the Kingdom of Ends.[82] This is because deception violates a person's ability to act as a fully rational moral legislator — it uses their reason as a tool for some other end.[83] So, lies are morally impermissible upon the KoE.

Kant believes that these formulations are equivalent. But various Kantian scholars have noted that they are not equivalent.[84] There are various ways to interpret this fact. For example, Wood suggests that the FoUL and FoH, when combined, lead to the KoE.[85] Contrarily, Korsgaard suggests that the non-equivalency provides room for KE to include special moral principles to deal with evil.[86] She believes that such principles would make Kantian Ethics sensitive to the real world, and thus avoid the IWO. Let us show that they are not equivalent, by seeing how the FoUL permits lying to the murderer, whilst the FoH and KoE do not.

It seems that lying would be morally impermissible upon the FoUL because it could not be the universal method of achieving its intended purpose.[87] It is an act which only works because one makes an exception of oneself. If the maxim 'lie when it benefits you' were universally accepted, that would lead to a contradiction in conception — if the maxim were universally accepted, the entire practice of truth-telling would break down, to the extent that telling a lie is not an act that is possible to perform.

[82] Korsgaard (1986) p.336
[83] Korsgaard (1986) p.334
[84] For example, Korsgaard (1986); Wood (2009); Johnson & Cureton (2022)
[85] Wood (2009) p.233
[86] Korsgaard (1986) p.346
[87] G 424 / 42; Korsgaard (1986) p.328

However, Korsgaard claims that one can universalise the maxim ‘lie to deceivers’ without causing a contradiction.[88] When the Murderer asks you where your friend is, Korsgaard assumes that the Murderer assumes that you do not know their intentions; they assume you do not know you are talking to a murderer.[89] Suppose he pretends to have a benign motive in trying to find your friend. Thus, the Murderer is trying to deceive you. So, even though it is universally accepted that one is morally permitted to lie to deceivers, the Murderer will not think that you are lying to them, because they will think that you do not know that they are deceiving you. And this will be the case in every situation in which one has the chance to act upon the maxim ‘lie to deceivers’ — the deceiver will presume that you do not know that they are trying to deceive you, and so the efficacy of your lie is not compromised. Therefore, we can universalise acceptance of the maxim ‘lie to deceivers’ without a contradiction. So, it is morally permissible to lie to the Murderer at the Door upon the FoUL.

One might contest this argument from Korsgaard. Truth-telling was a perfect duty for Kant — violation of a perfect duty is morally forbidden, and invites moral blame.[90] Further, there is at least a concern whether the maxim ‘lie to deceivers’ could be universally arrived at a priori: it is unclear how specific one’s maxims can be.[91] Nevertheless, this is not the objection that this thesis will press. My aim is not to discern whether Korsgaard has correctly interpreted the non-equivalency of the formulations of the Categorical Imperative. Instead, it is to argue that, even if Korsgaard’s interpretation of KE is on the right track, it still fails to escape the DWO. Accordingly, let us see how the maxim ‘lie to deceivers’ is morally impermissible according to the FoH and KoE.

[88] Korsgaard (1986) p.329-30
[89] Korsgaard (1986) p.330
[90] Wood (2009) p.229
[91] See Parfit (2011) p.289

It is not possible to assent to being lied to — a lie is only successful when the other does not know that it is a lie.[92] The efficacy of lying to the Murderer at the Door is determined by the Murderer not knowing that you are lying to them. Recall that the FoH states that *an act is morally permissible if and only if those affected by the act could possibly assent to this mode of acting.*[93] It is not possible for the recipient of the lie to assent to being lied to. The Murderer cannot possibly assent to your lie. Thus, lying to the Murderer is morally impermissible according to the FoH.

Similarly, lying to the Murderer is morally impermissible according to the KoE. Recall that the KoE states that *an act is morally permissible if and only if it accords with a principle that could be accepted by a community of fully rational, morally legislating individuals.*[94] Presumably, the maxim 'lie to others' would not be accepted by a community of fully rational, morally legislating individuals. Again, it disqualifies the person being lied to from the Kingdom of Ends. So, the maxim 'lie to deceivers' would not be accepted either, as it would not make sense: there would be no deceivers to lie to. No principle which permits lying to the Murderer would be accepted in the Kingdom of Ends, as the KoE presumes the ideal behaviour of each of its members. Thus, lying to the Murderer is morally impermissible according to the KoE.

[92] Korsgaard (1986) p.333
[93] Korsgaard (1986) p.331
[94] Johnson & Cureton (2022) §8

So, lying to the Murderer at the Door is morally permissible upon the FoUL, and morally impermissible upon the FoH and KoE formulations of the Categorical Imperative. The FoUL is more permissive than the FoH and KoE. Now, we can look at Korsgaard's second claim: that this means we ought to provide special principles for dealing with evil.

4.5.2 Special Kantian Principles for Dealing with Evil

Korsgaard views the discrepancy between the FoUL, and the FoH and KoE, as an opportunity to construct a version of KE which is more sensitive to the real world. She designs a double-level Kantian Ethic, taking inspiration from Rawls' distinction between ideal and non-ideal theories of justice.[95] This Section sets out her double-level theory.

The FoH and KoE are stricter than the FoUL. Thus, Korsgaard labels them as ideal moral theory. She suggests that they are not designed for dealing with evil.[96] So, the FoH and KoE become an ideal which governs our daily conduct. However, in non-ideal cases, Korsgaard suggests that it is morally permissible to act in a way that violates the FoH and KoE. For example, by lying to a murderer. But consistently, the more permissive FoUL serves as a baseline for our behaviour — it is never morally permissible to violate the FoUL. So, we can state the rules of Korsgaard's double-level moral theory as follows:

> *An act is morally permissible if and only if such acts are permitted by some principle whose universal acceptance everyone could rationally will. And, only if conditions are ideal, an act is permissible when those affected by the act could possibly assent to*

[95] Korsgaard (1986) p.341-343
[96] Korsgaard (1986) p.346

those modes of acting. And, only if conditions are ideal, an act is permissible when it accords with principles that could be accepted by a community of fully rational, morally legislating individuals.

So, it seems as though Korsgaard has provided a way to make Kantian Ethics more sensitive to nonideal circumstances. But I will now show that this cannot solve the DWO.

4.6 Why Does This Fail?

This solution employs the FoUL as the baseline for morality — one is morally permitted only to act in ways that satisfy the FoUL. So, if the FoUL, as our baseline, makes permissible acts that are clearly morally impermissible, then it is clear there is a flaw. And, by thinking about our previous examples, it is clear the FoUL will cause one to:

(i) Do something avoidably disastrous; or

(ii) Avoid doing something wonderful without good reason; or

(iii) Allow irrelevant facts about distant worlds to determine what we ought to do.[97]

Recall (**4.4**) and our utility mine case. We can imagine a world there is a rule, *T,* that says, 'tell the truth' and you know that there exists one mine:

1. A good-mine that triggers when there is universal acceptance of *T*; and
2. A bad-mine that triggers when *you* accept *T*, but not everyone does.

[97] Podgorski (2018) p.290

There, I argued that (KE1), which is a restatement of the FoUL, is entirely insensitive to the bad-mine, and will morally require that you accept *T*. Nothing has changed — it remains true that you are morally required to accept *T.* So, this solution clearly fails — there are many cases where the FoUL will cause you to do something avoidably disastrous, or avoid doing something wonderful, *because* you allow irrelevant facts about distant worlds to determine what you ought to do.

Korsgaard noted that the FoH and KoE are too restrictive and idealistic. Her solution aimed to resolve this by making them a target to aim towards, rather than criteria for permissibility. But this does not change the fact that the FoUL suffers from the IWO too — by determining what we ought to do based upon an assessment of a distant world in which everyone accepts the maxim from which you act, the FoUL is insensitive to important moral facts. So, Korsgaard's double-level moral theory cannot avoid the DWO. And so, Korsgaard fails to solve the IWO, *because* she has misunderstood the force of the objection.

Conclusion

In this Chapter, I have done two things. First, I showed that the IWO's scope is wider than just RC — it applies to Kantian Ethics. Second, I showed that Korsgaard's attempt to salvage Kantian Ethics from its idealism fails to resolve the DWO. Thus, Korsgaard has misunderstood the force of the IWO. In the next Chapter, I look at Moral Contractualism. In doing so, I challenge one of Podgorski's claims about the scope of the DWO and, as a result, argue that the IWO has a wider scope than even he believes.

5. Extending the DWO to Rawls' Theory of Justice

Podgorski's argument is that the DWO will affect *any view which determines what we ought to do by evaluating possible worlds that differ from the actual world in more than what is up to us*.[98] Thus, the scope of the IWO includes Rule Consequentialism (Chapter 3), various other Consequentialist moral theories,[99] and Kantian Ethics (Chapter 4). He also notes that various forms of Contractualism would be affected by the DWO in the following passage:

> "Notably, while [the DWO] affects contractualist views across both the Hobbesian (Gauthier 1998) and Kantian (Scanlon 1982) spectrum, all of which have the above feature, it does not apply to Rawls (1971), who purports to be providing a view about which political or social arrangements are just rather than how individuals ought to act in the actual world. This reveals one way to involve the assessment of distant worlds in a moral theory without falling prey to the objection from utility mines: one may evaluate those worlds in order to morally assess something broader than an individual action—an institution, practice, or collective pattern of behavior, while denying that a positive assessment of these immediately transmits a permission to individuals to act accordingly."[100]

It is interesting that Podgorski believes that a wide array of Contractualist moral theories will suffer from the DWO, but that Rawls' Theory of Justice *will not*. This Chapter, then, looks to explain Podgorski's claims about the DWO in relation to Contractualism. It also challenges

[98] Podgorski (2018) p.279
[99] Such as R.M. Adams' (1976), C. Johnson's (1991), and J. Driver's (2007), as referenced by Podgorski (2018) p.291
[100] Podgorski (2018) p.292

the assertion that the DWO would not affect Rawls' Theory of Justice. In doing so, I argue that it is likely that the scope of the IWO is wider than even Podgorski believes.

To do this, I first define Contractualism, and explain the distinction Podgorski makes between Hobbesian and Kantian Contractualist Views (**5.1**), before providing three reasons why one might accept Contractualism (**5.2**). Then I show how the DWO affects Contractualist Views (**5.3**). Having done that, I explain why Podgorski thinks the DWO would not affect Rawls' Theory of Justice (**5.4**). I then argue that he is mistaken, and that — at least — a strong argument can be made to suggest that the DWO will affect Rawls' Theory of Justice in some way (**5.5**). In doing so, I show that the IWO has a wider scope than even Podgorski believes, and that a satisfying account of Rawls' Theory of Justice requires a response to the DWO.

5.1 What is Contractualism?

Contractualism is a broad term for a set of moral theories in which the moral status of acts is determined by some form of social contract or agreement.[101] The different types of Contractualist moral theory differ in at least the following areas.[102] First, who is making the agreement? Contractualist moral theories differ on how rational persons entering a social contract are, what their interests are, how many persons are involved in the contract, and so forth. Second, what does their choice consist of? Contractualist moral theories differ on the moral force of the principles that persons choose, to whom the chosen principles apply, for how long the principles apply, and so forth. Third, about what are they choosing principles?

[101] E. Ashford & T. Mulgan (2018) §0
[102] As set out by Schafer-Landau (2013) p.556

Contractualist moral theories differ on whether persons are choosing principles to cover all areas of their moral lives or just some aspect, on whether the principles bind persons for their entire lives or just some period of time, and so forth. Fourth, whether the contract is hypothetical or actual — some suppose that persons have actually entered into an agreement about moral principles, whilst some treat the contract as a hypothetical.[103] Different answers to these questions can lead to different Contractualist moral theories deeming permissible very different sets of acts.

Although there is considerable scope for disagreement between different versions of Contractualism, all plausible versions agree that the contract produced provides some form of constraint upon the acts of individuals. Thus, we can formulate a general version of Contractualism as follows:

> *(CT1): An act is morally permissible if and only if it accords with the set of acts or principles which have been agreed upon by the relevant agents.*

For Contractualism to provide a justified list of rules or principles, the Contractualist must do two things.[104] First, they must carefully describe the hypothetical situation in which the agreement upon which acts are morally permissible is made. This allows us to understand the conditions in which each person enters the social contract. Second, the Contractualist must take us through the reasoning by which that list of acts or principles is produced. Having done this, we know the list of morally permissible acts or principles. The only morally permissible acts are those which are permitted by the social contract into which those agents

[103] Gauthier (1991) 576-77
[104] Buchanan (1980) p.18

enter. Many Contractualists, such as Rawls, state that the contract must also include priority rules, so that we can understand how to adjudicate in cases where there are two or more conflicting principles or acts permitted by the social contract.[105]

Podgorski's passage mentions two different strands of Contractualism — the Hobbesian (characterised by Gauthier) and the Kantian (characterised by Scanlon). Whilst both would broadly agree with (CT1), the two have important differences, and produce different action guidance in certain cases. Let us look at them in turn.

5.1.1 Hobbesian Contractarianism

The 'Hobbesian' strand of Contractualist moral theories referenced by Podgorski is often referred to as 'Hobbesian Contractarianism' (HC) in contemporary literature.[106] This is done to distinguish it from the other strand of social contract thought, Kantian Contractualism, (KC) that will be discussed in (5.2.2).[107]

HC refers both to a theory that legitimises government authority and a moral theory that grounds the moral permissibility of acts.[108] Both are based upon Thomas Hobbes' idea of the social contract,[109] a contract into which rational individuals willingly enter — they agree to limits upon their individual liberty for their mutual self-interest. We can describe a proponent of the former as a Contractarian about political theory, and the latter as a moral Contractarian. We are interested in moral Contractarians. The moral version of HC is based upon the notion

[105] Buchanan (1980) p.20
[106] Cudd & Eftekhari (2021) §0
[107] Cudd & Eftekhari (2021) §0
[108] Cudd & Eftikhari (2021) §0
[109] Lloyd, S. A. & Sreedhar S. (2022) §2

that *morality consists in those forms of cooperative behaviour that it is mutually advantageous for self-interested agents to engage in*.[110]

The moral Hobbesian Contractarian holds that an individual willingly agrees upon a set of moral principles that limit which acts are permitted for them and others, in order to make their life go better. Each rational agent appreciates that it would be mutually advantageous to enter into this contract, even though it limits the set of acts or principles which are permitted for that agent. It is advantageous for a given agent, *X*, that the set of acts or principles available to other agents, *Y*, *Z*, …, *etc.*, is morally limited. And it is advantageous for agents *Y*, *Z*, *..., etc.*, that the set of acts or principles available to agent *X* is morally limited. So, each willingly enters into the agreement to place limits on the set of morally permissible acts or principles available to an agent. Let us use an example:

> Case 12: Suppose that Agent *X* helped others only when she expected to benefit from doing so. There would be many cases where Agent *X* would not help others, even when it would not cost her much, and give a great benefit to the person(s) she helped. It would make things go much better if Agent *X* did help others in such cases. Suppose further that Agent *X* realises that it would be mutually advantageous if she and Agents *Y*, *Z*, …, *etc.*, agreed to help each other when it would not cost them much, and would give a great benefit to the person(s) being helped. Each would realise that those belonging to a group whose members adhere to such a practice of mutual assistance enjoy benefits in interaction that are denied to others. We may then represent such a practice as rationally acceptable to everyone.[111]

[110] Ashford & Mulgan (2018) §2
[111] Example adapted from Gauthier (1991) p.575-76

Not only is it mutually advantageous to enter into such agreements, but the Hobbesian Contractarian believes it is also mutually advantageous to stick to them. Whereas some Act Consequentialists might think it would be better to enter into an agreement and benefit from others following the rules, whilst secretly breaking the rules for one's own benefit, the Hobbesian Contractarian disagrees. Gauthier argues that this would limit the level of mutual advantage one can enjoy, and thus would make things go worse for you.[112] For example, if Agent *X* only keeps promises when it is advantageous to do so, this will eventually become public knowledge. Thus, Agents *Y*, *Z, ..., etc.*, will exclude Agent *X* from agreements which lead to great mutual advantage, but require rigorous compliance. Thus, Agent *X*'s inauthenticity limits the mutual advantage they can enjoy by entering into agreements. So, it is mutually advantageous to authentically enter into the social contract.

If we presume that it is mutually advantageous for us to make things go best (for each of us), we can formulate HC as the following:

> (HC1): *An act is morally permissible if and only if it accords with the set of rules or principles that are agreed to make things go best for each of us.*

Given the above discussion about how it is mutually advantageous to enter into a social agreement to limit one's own behaviour, one might think that HC and Act Consequentialism (AC) are extensionally equivalent — if it is mutually advantageous to enter into a contract and agree to a set of moral principles, then that would be what makes things go best. So, surely AC would require that one follow HC. But this is mistaken — there will be cases

[112] Gauthier (1991) p.576

where what is mutually advantageous for Agents *X*, *Y*, *Z*, …, *etc.*, is not what makes things go best. Thus, HC and AC will produce different action guidance. For example:

> Case 13: Suppose that you live in a small, isolated town in the desert. Call it Oasis. 100 people live in Oasis, and there is just enough food and water to sustain Oasis' population. Out of the desert arrive 5 refugees. Splitting Oasis' supplies between 5 extra persons is not enough to cause anybody to become malnourished, but it will non-trivially inconvenience Oasis' current citizens. If you refuse to help the refugees, they will certainly die. Suppose further that 5 persons dying makes things go worse than 100 people being non-trivially inconvenienced.

If, when entering into their social contract, the citizens of Oasis considered such a case, it seems that there are two acts available to them. First, they could welcome in the 5 extra persons, saving their lives and non-trivially inconveniencing the 100 citizens. Second, they could refuse to help the refugees, certainly causing their deaths, but not inconveniencing the citizens. If it is true that 5 persons dying makes things go worse than 100 people being non-trivially inconvenienced, then AC will require that Oasis helps the 5 refugees — there is no act which makes things go better. But HC will not require it. When entering into their agreement, the citizens of Oasis need not include the 5 refugees — if the agreement is actual, rather than hypothetical, this could concretely be the case. And Oasis' citizens will realise that it is mutually advantageous for them to refuse to help the refugees. Thus, it will *at least* be morally permissible upon (HC1) for the citizens to refuse to help the refugees. It is plausible that it will be morally required, because it would be mutually disadvantageous for the citizens to help the refugees. Thus, AC and HC provide different action-guidance.

Let us now look at how HC differs from Kantian Contractualism.

5.1.2 Kantian Contractualism

In HC, morality consisted in those forms of cooperative behaviour that it is mutually advantageous for self-interested agents to engage in.[113] Contrarily, Kantian Contractualism (KC) grounds the social contract in the equal moral status that all rational agents enjoy.[114] The view is based upon the Kantian theory discussed in (**4.1**). Given the equal status of all rational agents, KC looks to generate a set of rules or principles that a group of rational persons could agree to, *hypothetically*. Scanlon provides a prominent account of KC, which can be formulated as follows:

> (KC1) *"[A]n act is wrong if its performance under the circumstances would be disallowed by any system of rules for the general regulation of behavior which no one could reasonably reject as a basis for informed, unforced general agreement*".[115]

Thus, it bears a passing resemblance to Kant's Kingdom of Ends — *an act is morally permissible if and only if it accords with a principle that could be accepted by a community of fully rational, morally legislating individuals.*[116] There are some, such as Parfit, who suggest that KC coincides with the best version of Kantian Ethics.[117]

[113] Ashford & Mulgan (2018) §2
[114] Ashford & Mulgan (2018) §2
[115] Scanlon (1982) p.110; as found in Podgorski (2018) p.292
[116] Johnson & Cureton (2022) §8
[117] Parfit (2011) p.407. Whether Parfit is correct is not important to my overall thesis — the DWO will apply, nonetheless.

And KC is different to HC. Ashford & Mulgan differentiate between the two in a helpful way: "Under [HC], I seek to maximise my own interests in a bargain with others. Under [KC], I seek to pursue my interests in a way that I can justify to others who have their own interests to pursue."[118] We can use Case 13 (the town Oasis) again to show that KC makes permissible a different set of acts and principles to HC:

Upon KC, the only morally permissible acts are those which would not be disallowed by any set of principles for the general regulation of behaviour that no one could reasonably reject as a basis for informed, unforced, general agreement. The scope of KC is greater, then, than the community of citizens in the town of Oasis. Instead, it is supposed to apply to all rational beings. And these 5 refugees are rational beings. So, the morally permissible acts are those which any rational beings — not just those who live in Oasis — could not reasonably reject. Presumably, a rational being who is not a citizen of Oasis would argue that it disrespects the equal moral status of the refugees to prioritise the non-trivial but not serious inconvenience to 100 persons over the death of 5 persons. So, they could reasonably reject the principle which permits citizens of Oasis to refuse to help the refugees. Thus, whilst it was morally permissible upon HC to refuse to help the refugees, it is morally impermissible upon KC to refuse to help the refugees — hence, KC and HC provide different action guidance.

I have set out what Contractualism is, and explained the various strands mentioned by Podgorski. Before looking at how the DWO affects these theories, let us consider why someone might accept either version of Contractualism.

[118] Ashford & Mulgan (2018) §2

5.2 Why Accept Contractualism?

Like Act Consequentialism, Rule Consequentialism, and Kantian Ethics before it, Contractualism is a *prima facie* plausible moral theory with various strengths. Below are three that are shared by both HC and KC.[119] These strengths will also apply to Rawls' Theory of Justice.

By grounding morality in the notion of a social contract or agreement — actual or hypothetical — we can view the moral evaluation of acts as the outcome of a rational collective choice. The reason that an act is morally permissible or impermissible is because of an agreement between oneself and others. This provides strong explanatory power in cases where it seems that morality requires one to act seemingly against their own self-interest. That morality is based on agreed-upon principles tells us both *which* acts will be morally permissible, and *why* they are — it acts as a justification as well as determining which acts are morally permissible or impermissible.[120]

Contractualism also has the notion of obligation built into it — since it requires that persons willingly and genuinely enter into an agreement as to which acts are morally permissible, each individual can be thought of as making a basic commitment to the principles they choose. That commitment could be concrete upon a version of Contractualism in which the agreement actually occurs, or implicit upon a hypothetical version of Contractualism. Nevertheless, a sense of obligation is built into the acceptance of any plausible version of Contractualism. This, in turn, justifies the enforcement of those principles. Whereas the proponent of a moral theory like AC needs to explain why it is wrong to act in a way that

[119] As provided by Buchanan (1980) p.18
[120] Rawls (1971) in Shafer-Landau (2013) p.583

does not make things go impartially best, the Contractualist can easily explain why an act which violates its rules is judged as morally wrong — it violates the principles agreed upon in the social contract.

Third, because the principles are chosen voluntarily — either because of mutual advantage or because of equal moral status — the principles will generally turn out to be equitable, fair, and intuitive (at least for those who are party to the contract). This is because they require at least the people within a given community to agree to them, including those who are particularly poorly off. So, it is likely that Contractualism will produce highly intuitive principles.

These reasons suffice to show that the Contractualist project — in general — is well motivated. Now, we can look at how both HC and KC are affected by the DWO. Then, we can look at Rawls' Theory of Justice.

5.3 How Does the DWO Affect Hobbesian and Kantian Contractualism?

Both HC and KC evaluate worlds which differ from our own in more than what is individually up to us. They produce a set of principles or rules by imagining a world in which a given group of people agrees to live by those rules, and determining how well things would go in that world.[121] It should be clear that such moral theories will face the DWO. Podgorski writes that Contractualist theories also face the DWO because they involve the evaluation of a world in which a set of rules is generally agreed upon.

[121] As noted by Podgorski (2018) p.292

This Section provides a utility mine example for each of HC and KC, to show how each will cause us to either:

(i) Do something avoidably disastrous; or

(ii) Avoid doing something wonderful without good reason; or

(iii) Allow irrelevant facts about distant worlds to determine what we ought to do.[122]

It also eliminates some potential responses from proponents of HC and KC.

5.3.1 Hobbesian Contractarianism

Recall that we formulated (HC1) as follows: *An act is morally permissible if and only if it accords with the set of rules or principles that are agreed to make things go best for each of us.* Now imagine the following utility mine case:

Suppose that there exists a rule, *R*, that says 'do a jumping jack at noon every day', and you know that two utility mines exist:

1. A good-mine that triggers if and only if the rule, *R*, is agreed to by everyone in a community; and
2. A bad-mine that triggers if and only if 1 or more people agree to the rule, *R*, but not everyone agrees with this rule.

[122] Podgorski (2018) p.290

It is mutually advantageous to trigger a good-mine and release an unimaginable amount of wellbeing. So, if it is known that universal (or, group-wide) agreement with *R* will lead to the triggering of a good-mine, then it is mutually advantageous for us to agree to *R*. So, this rule would be accepted under (HC1). Acts that accord with *R* would be morally permissible. If we assume that the Hobbesian Contractarian wants to maximise mutual advantage — that they want to make only the agreements that are most advantageous to them — we can also assume that no conflicting rule will be agreed to (at least for this specific case). Thus, the only morally permissible act available to you upon (HC1) is to go outside and do a jumping jack, as it is the only action available to you that adheres to *R*.

Realistically, though, there would not be universal acceptance of *R*, and so when you go outside to do a jumping jack, you do so knowing that your action will most likely trigger the bad-mine and release unimaginable suffering. Thus, in this Case, we do something avoidably disastrous *because* we allow irrelevant facts about distant worlds to determine what we ought to do.

The Hobbesian Contractarian might try to resist this argument in two ways. They might first argue, as Gauthier does, that our disposition to violate moral rules is transparent.[123] This means that it is clear when we intend to stick to our agreement, and when we intend not to. So, in our imaginary scenario where individuals come together to decide whether to accept *R* as a rule, we can see clearly that not everybody intends to (or will not be able to) follow *R*. Accordingly, we will not accept it as a rule. Second, they might try, as Dworkin does,[124] to introduce some sense of insurance against wrongdoing into their agreement. Those making

[123] As noted by Schafer-Landau (2013) p.556
[124] Dworkin (1981) as set out in Wolff (2007) p.127-129

the contract, in the knowledge that sometimes people act wrongly, might want to safeguard against that, by not entering into agreements that can be ruined by a single wrongdoer — thus, even though universal agreement might trigger a good-mine, they might deem it too risky to enter into such an agreement.

Such responses, though, are flawed in the same way as the similar responses from proponents of RC in (**3.3.1**) — they misunderstand the force of the problem. The DWO does not affect HC because it requires universal agreement. The problem is that it is insensitive to important moral facts because it evaluates a world that is different from our own in more than what is individually up to us. Consider how HC might be formulated by the Contractarian who employs Gauthier's or Dworkin's response:

> *(HC2): An act, A, is morally permissible if and only if it accords with the set of rules or principles that are agreed to make things go best for us if followed by 90% of people.*

(HC2) considers that some people are wrongdoers, or that we will see that some people are not genuine in their agreement. Upon (HC2) we will only enter mutual agreements which take this into account. This view avoids the utility mines example — if 90% of people agreed to go outside to do a jumping jack at midday, a bad-mine would be triggered. Thus, according to (HC2), *R* is not mutually advantageous, and we ought not to agree to it. But (HC2) fails to avoid the DWO in the same way as Fixed-Rate Rule Consequentialism. For we can simply change the trigger conditions for our utility mine. Consider again the rule, *R*, that says 'do a jumping jack at noon every day', and two new utility mines:

1. A good-mine that triggers when there is 90% agreement with *R*; and
2. A bad-mine that triggers when 1 or more people agree to *R*, but there is not 90% adherence.

Suppose you know that fewer than 90% of people will do a jumping jack at noon today. Should you nevertheless follow *R*? Of course not, for that would lead to disaster. But (HC2) implies that you should! This is because it would be mutually advantageous to agree to *R* in a world where 90% of people agree with this rule, as the good mine would be triggered. In this Case, (HC2) implies that you ought to follow *R*, and that the only morally permissible action available to you is to go outside and do a jumping jack. Upon (HC2), irrelevant facts about other possible worlds continue to determine what is morally permissible in our world. So, these responses fail.

Let us now look at Kantian Contractualism.

5.3.2 Kantian Contractualism

Recall that Scanlon formulates his Kantian Contractualism (KC1) as follows: "*an act is wrong if its performance under the circumstances would be disallowed by any system of rules for the general regulation of behavior which no one could reasonably reject as a basis for informed, unforced general agreemen*t".[125]

As Podgorski writes, "[… it] is hard to see how to interpret this in a way that doesn't somehow involve evaluating, from the individual agent's perspective at least, a world where

[125] Scanlon (1982) p.110; as found in Podgorski (2018) p.292

the rules are the basis for unforced general agreement. Utility landmines (tailored to the agent's tastes, perhaps) can have an effect on the choice-worthiness of that world.[126]

Due to its Kantian origins, we can use a similar utility mine case to undermine (KC1) as we did Kantian Ethics. The only acts that are morally permissible upon (KC1) are those which would not be disallowed by a system of rules for the general regulation of behaviour. Any act which one could rationally will to become universally accepted would surely fulfil this criterion. In other words, any act which is permitted by the Formula of Universal Law (FoUL) will be permitted by Scanlon's Contractualism — this is partly why Parfit believes that the most plausible version of Kantian Ethics coincides with Contractualism.[127] We have already seen a utility mine case in (**4.4**) that affects the FoUL. Accordingly, we have a case where an act would be permissible upon (KC1), and yet would cause us to do something avoidably disastrous because of irrelevant facts about a distant world. We need not rehearse the case again. Kantian Contractualism also suffers the DWO.

I have explained Podgorski's claims about Hobbesian and Kantian Contractualism suffering from the DWO. We can now look at Podgorski's claims about Rawls.

5.4 Rawls' Theory of Justice and the DWO

Before looking at whether the DWO affects Rawls' Theory of Justice, let us get clear on what Rawls' theory is.

[126] Podgorski (2018) p.292
[127] Parfit (2011) p.407 as found in Rosen (2009) p.78

5.4.1 What is Rawls' Theory of Justice?

Rawls' Theory of Justice aims to lay out a set of general principles of justice which underlie the various considered judgments we make in particular cases.[128] The general principles that Rawls arrives at are "the principles that free and rational persons concerned to further their own interests would accept in an initial position of equality [*the original position*]".[129] It can be described as a Kantian Contractualist theory of justice.[130] So, if our other Kantian Contractualist theories faced the DWO, we might expect Rawls to as well.

As noted in (**5.1**), there are three questions according to which we can differentiate between different Contractualist views. First, what are the people entering into the contract like? Second, what does their choice consist of? Third, about what are they choosing principles?[131] We need to answer these questions before we can properly formulate Rawls' theory. Rawls' description of the original position — in which individuals enter into the social contract — answers these questions. Thus, it helps us to characterise Rawls' view.

5.4.2 What is the Original Position?

There are four main elements of the original position.[132]

First, the rational motivation of each person in the original position is assumed to be the same — each individual is motivated to pursue their life plans (and thinks of themselves as a

[128] As identified in Buchanan (1980) p.6
[129] Rawls (1971) p.11, as found in Buchanan (1980) p.18
[130] Ashford & Mulgan (2018) §2
[131] Schafer-Landau (2013) p.556
[132] As identified in Buchanan (1980) p.19-20

rational agent with a worthwhile life plan).[133] It is further assumed that it would be rational for each person to gain as large a share of primary goods as possible, because these are goods that will be useful in the pursuit of any life plan.

Second, each individual in the original position is under a *veil of ignorance*.[134] Nobody knows anything about themselves, other than that they are a rational agent motivated to pursue their life plans and maximise their share of primary goods. No individual knows whether they are from the dominant social group, how old they are, their physical attributes or talents, and so forth. This ensures that the individuals do not try to affect the choice of principles to advantage themselves or disadvantage others, based on a knowledge of individuals' respective talents and attributes.[135]

Third, there are formal constraints on the types of principles that individuals in the original position are permitted to choose. The principles must be: *general*, they must cover all questions about social justice; *universal in application*, they must apply to all members of society; *universalizable*, they must be principles whose universal acceptance we can endorse; *publicizable*, if they are to guide and justify our acts and policies, they must be known and understandable by all; *adjudicative*, they must provide a way of ordering conflicting claims to settle disputes; and *final*, they must be ultimate principles which provide a final court of appeal for disputes about justice.[136]

[133] Buchanan (1980) p.19
[134] Buchanan (1980) p.19
[135] Buchanan (1980) p.19
[136] Rawls (1971) p.131-35 as found in Buchanan (1980) p.20

Fourth, there is a list of competing principles of justice from which those in the original position can choose. This list includes various versions of utilitarianism, as well as Rawls' principles of justice.[137]

So, why does Podgorski think that the DWO does not apply to Rawls' Theory of Justice?

5.4.3 Why Might the DWO Not Apply?

Based on our understanding of the original position, it looks like a Rawlsian moral theory would be threatened by the DWO. If the rules of Rawls' system are those that are chosen in the original position, one might formulate a Rawlsian Ethic as:

> *(R1): An act is morally permissible if and only if it accords with the set of rules which would be agreed upon by rational persons in the original position.*

This immediately looks to share the structural problems that cause RC, KC, and other forms of Contractualism to suffer the DWO. When we assess an act under (R1), we are considering a distant world in which that act has become a universal rule. Accordingly, we can use the same reasoning here as we did in (**5.3.2**) — if an act is morally permissible upon the FoUL, it would be agreed upon by rational persons in the original position. We have already seen that the FoUL suffers from the DWO. So, (R1) will also face the DWO. So why does Podgorski think this does not apply?

[137] Buchanan (1980) p.20

The reason Podgorski thinks that Rawls is safe from the DWO is because Rawls is not trying to provide a moral theory. Rawls himself notes that his theory is not a complete contract theory — it is only applicable to *justice*, not to an entire ethical system.[138] That is, the principles chosen in the original position are meant to regulate our basic social and political institutions, rather than directly regulating the actions of individual members of society.

Accordingly, a formulation like (R1) would be a misappropriation of Rawls' aim. Recall, Podgorski writes that "[the DWO] does not apply to Rawls (1971), who purports to be providing a view about which political or social arrangements are just rather than how individuals ought to act in the actual world. This reveals one way to involve the assessment of distant worlds in a moral theory without falling prey to the objection from utility mines: one may evaluate those worlds in order to morally assess something broader than an individual action—an institution, practice, or collective pattern of behavior, while denying that a positive assessment of these immediately transmits a permission to individuals to act accordingly."[139]

Rawls' Theory of Justice — according to Podgorski — *does* morally assess a distant world. But a positive moral assessment of that distant world does not immediately have any consequences for what individuals ought to do in the real world. On Rawls' theory, the positive assessment of a distant world only has implications for which principles should regulate our basic social and political institutions. So, the DWO cannot apply, as Rawls' Theory of Justice causes no individual to either:

[138] Rawls (1971) in Shafer-Landau (2013) p.584
[139] Podgorski (2018) p.292

(1) Do something avoidably disastrous;

(2) Avoid doing something wonderful for no good reason; or

(3) Allow irrelevant facts about distant worlds to determine what one ought to do.

I think that we can challenge this claim from Podgorski. In the final Section of this Chapter, I argue that Rawls' Theory of Justice faces a dilemma — either it must still be affected by the DWO, or it lacks a purpose. I then conclude that the DWO has a wider scope than even Podgorski believes.

5.5 Why the DWO Does Affect Rawls' Theory of Justice

In Rawls' Theory of Justice, a positive moral assessment of a distant world does not immediately have a consequence for what we ought to do in the real world. But, presumably, a positive moral assessment of a distant world must *at some point* have *some consequence* for what we ought to do in the real world. If not, then what purpose does Rawls' Theory of Justice have? It would be strange to have produced a system of just principles with no attached obligations or directions to try to realise those principles. Rawls' Theory of Justice is surely intended to be normative, rather than just evaluative.[140] If Rawls' Theory of Justice *is* purely evaluative, then Podgorski is correct — the DWO will not apply. But then it seems to me that the theory is inert — it serves no purpose.

Rawls does note that his work is mostly concerned with ideal theory.[141] In other words, his Theory of Justice is setting out the principles which best exemplify justice in ideal conditions,

[140] As Gilabert (2011) argues that they are, as found in Valentini (2012) p.658
[141] Valentini (2012) p.660

rather than in the real world. Yet his theory is *not* supposed to be a purely descriptive account of which principles of justice would work best in an ideal society. Invoking the notion of justice implies normative force —Valentini notes that principles of justice are *weighty*, they are often seen as giving rise to rights.[142] It is generally agreed that Rawls' principles of justice ought to yield concrete judgments about the justice of specific institutions and practices, and guide us in developing policies and laws to correct injustices in the basic structure.[143]

And if Rawls' thought experiment in the original position ought to yield concrete judgements about the justice of specific institutions and practices, and guide us in developing policies and laws to correct injustices in the basic structure, then there is scope for the DWO to apply to Rawls' Theory of Justice. If a positive or negative assessment of a distant world provides any form of transmission into an evaluation of what one ought to do, then there will surely be a utility mine case which can hijack this process of transmission.

So, it seems that Rawls's Theory of Justice will be caught in a dilemma. The first option: positive assessments of distant worlds transmit no corresponding obligations to individuals to try to bring about the just system. In this case, Rawls' Theory of Justice provides little, or no, practical help in contemporary political philosophy.[144] The second option: positive assessments of distant words do transmit corresponding obligations to individuals *in some way* to try to bring about the just system. In this case, Rawls' Theory of Justice is vulnerable to the DWO *in whichever way the corresponding obligations are transmitted*. For Rawls' theory to be consequential in any way, there must be at least one case where this normative response is the deciding factor between two courses of action. There will likely be many. And

[142] Valentini (2012) p.658
[143] As stated by Buchanan (1980) p.9
[144] See Valentini (2012) p.654-662 for a summary of accusations along this line

in this case, one is allowing irrelevant facts about faraway worlds to determine what one ought to do. Inevitably, in some of those cases, those facts will cause one to do something avoidably disastrous, or to avoid doing something wonderful for no good reason. Since there are various ways in which an obligation could be transmitted from a positive assessment of Rawls' distant world, I cannot provide an exact account of how this would work, but hopefully I can show that the process of transmission leaves Rawls' theory vulnerable.

Let me provide an example to do this. The example uses universal compliance to show the structural threat that the DWO places upon Rawls' Theory of Justice.

> Case 14: Suppose that everyone has a button in front of them. Suppose further that and if everyone presses their button, then our social and political institutions will henceforth be regulated by the principles selected in the original position. However, if some people press their buttons and others do not, the world will explode.

If you accept that the realisation of the principles of the original position would be overwhelmingly positive, we could characterise this case using Podgorskian language as follows. Suppose that there exist two utility mines:

1. A good-mine that triggers if and only if everyone presses their buttons; and
2. A bad-mine that triggers if and only if one or more people press their buttons, but not everyone presses their button.

What does Rawls' Theory of Justice imply in this case? Would deliberators in the original position agree that everyone would be required to press the button?

If yes, then we seem to get a requirement to press the button in the real world. Since not everyone would press the button in the real world, Rawls' Theory of Justice seems to be falling victim to the DWO — it is requiring, or making permissible, or incentivising, disposing, individuals to act in a way that, given the certainty of noncompliance, will make them seriously vulnerable, releasing a bad-mine. It is allowing irrelevant facts about faraway worlds to require people to do something disastrous.

If no, because the deliberators in the original position do not choose principles to govern the behaviour of individuals, then Rawls' theory is inert — whilst telling us how just social and political institutions would be governed, it fails to tell us anything about when and how we are required to bring about the formation of these institutions. Rawls' Theory of Justice would have no application.

If one is tempted to respond that deliberators would choose not to require that everyone press the button because such a policy is far too risky, then I would remind the reader that the *degree of ideality* is not the problem here (as was shown in (**3.3.1**) with Partial Rule Consequentialism, and (**5.3.1**) with Partial Hobbesian Consequentialism). To offer this response would be to misunderstand the force of the IWO. We can imagine the same scenario where the problem persists with slightly different utility mines:

1. A good-mine that triggers if and only if there is 50% or more compliance with a given rule, *R*; and
2. A bad-mine that triggers if and only if one or more people comply with *R*, but there is less than 50% compliance with *R*.

Here, the degree of ideality might more closely resembles a real-world case. Yet irrelevant facts about other possible worlds continue to determine what is morally permissible in our world. Rawls' Theory of Justice thus seems to suffer the DWO. The alternative is that Rawls' theory tells you nothing about what you ought to do in the face of an unjust situation.

Conclusion

In this Chapter I have set out how the DWO applies to Contractualism, and have explained Podgorski's claims about how the DWO relates to Rawls' Theory of Justice. I then argued against Podgorski's claims about Rawls. I hope to have made a case for the claim that Rawls' Theory of Justice could face the DWO. This is a significant claim about the scope of the IWO and runs against Podgorski's understanding. Rawls' theory seems to be stuck between two untenable positions: either it has no practical applications for political philosophy, since the assessments it makes have no influence on what individuals ought to do to bring about just institutions; or it faces the DWO, and hence causes one to either do something avoidably disastrous or to not do something wonderful for no good reason, because of irrelevant facts about distant worlds. I hope at least to have shown that a proponent of Rawls' Theory of Justice needs to find a path out of this dilemma.

So, the IWO has a very wide scope: it affects various versions of Rule Consequentialism, Kantian Ethics, and Moral Contractualism. I believe I have made a strong case that Rawls' Theory of Justice also faces the IWO. At least, a proponent of Rawls' Theory of Justice must show how the IWO does not affect it. I have also shown, using the DWO, that various attempts to resolve the IWO fail, because they misunderstand its force. The problem is not

the degree of ideality, but that these moral theories allow irrelevant facts about distant worlds determine what individuals ought to do.

Conclusions

The IWO is clearly a problem for various moral theories. Presumably, a moral theory ought not cause you to do something avoidably disastrous. It also ought not cause you to not do something wonderful without a good reason. And it obviously should not do these things because of irrelevant facts about distant possible worlds.

In this thesis, I have defined various moral theories and provided reasons why one might find each plausible. I have set out the IWO and how it affects each of these moral theories. Using Podgorski's DWO, I have argued that various attempts to solve the IWO have failed. In doing so, I have made the following claims about the scope and force of the IWO.

The IWO's scope is very wide. It is an objection that proponents of many moral theories must find a response to, including Rule Consequentialists, Kantian Ethicists, Contractualists and Moral Contractarians. I have argued that it also affects Rawls' Theory of Justice. I claim that either Rawls' theory has no practical applications for political philosophy, since the assessments it makes have no influence on what individuals ought to do; or it faces the DWO and will cause us to either do something avoidably disastrous or to not do something wonderful for no good reason, because of irrelevant facts about distant worlds. This shows that the IWO has a wider scope than even Podgorski believes. The implication of this claim is that the Rawlsian theorists, like those drawn to RC or other forms of Contractualism, must also find a route out of this dilemma.

And the IWO's force is widely misunderstood by the proponents of the moral theories that it affects. Using Podgorski's DWO, I have argued that various attempts to solve the IWO have

failed. For RC, this included changing RC's level of compliance (**3.3.1**), and using conditional rules like 'follow R, unless X' (**3.3.2**). For Kantian Ethics, this included Korsgaard's suggestion of a double-level Kantian moral theory (**4.5**). For Contractualists and Moral Contractarians, this included assuming that the disposition to violate moral rules is transparent, and introducing insurance against wrongdoers (**5.3.1**). In many of these cases, proponents of the moral theories tried to resolve the IWO by changing the *degree of ideality* of their moral theory. In each case, I have shown that a specific formulation of the IWO, the DWO, still applies. This shows that the degree of ideality is not the source of the IWO's force. Instead, these moral theories face the IWO because they allow irrelevant facts about distant worlds determine what an individual ought to do in the real world. This makes them insensitive to facts about the real world.

In making this argument, I hope to have made clearer the scope and force of the IWO.

Bibliography

Bibliography

1. **Arneson, R. (2005).** "Sophisticated rule consequentialism: Some simple objections." *Philosophical Issues*, *15*, 235-251.
2. **Ashford, E., & Mulgan, T**. (2018). "Contractualism", *The Stanford Encyclopedia of Philosophy* (Summer 2018 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2018/entries/contractualism/>.
3. **Brandt, R. B. (1984).** "Utilitarianism and Moral Rights." in *Canadian Journal of Philosophy*, 14.
4. **Buchanan, A. (1980).** "A critical introduction to Rawls' theory of justice." In *John Rawls' Theory of Social Justice: An Introduction*. Ed. Blocker, G. & Smith, J. Ohio University Press pp.5-41.
5. **Cudd, A., & Eftekhari, S. (2021)** "Contractarianism", *The Stanford Encyclopedia of Philosophy* (Winter 2021 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2021/entries/contractarianism/>.
6. **DePaul, M. & Hicks, A. (2021).** *"*A Priorism in Moral Epistemology*." The Stanford Encyclopaedia of Philosophy* (Summer 2021 Edition). Zalta, E. N. (ed.), URL = https://plato.stanford.edu/archives/sum2021/entries/moral-epistemology-a-priori/.
7. **Ebells-Duggan, K. (2011).** *"*Kantian Ethics*"* in *The Continuum Companion to Ethics,* ed. Miller, C., Bloomsbury Publishing, Sept 2011, pp. 168-189
8. **Gauthier, D. (1991).** *"*Why Contractarianism*?"* in *Ethical Theory: An Anthology, Second Edition.* Shafer-Landau, R (ed.), 2013, John Wiley & Sons, Inc. Reprinted from Peter Vallentyne, P. (ed.), Contractarianism and Rational Choice (Cambridge

University Press, 1991), 15–30. Reprinted with permission of Cambridge University Press.

9. **Hobbes, T. (1651).** "Leviathan" in *Hobbes: Leviathan: Revised student edition* (1996). R. Tuck (Ed.), Cambridge: Cambridge University Press.
10. **Hooker, B. (2000).** "Ideal Code, Real World." Oxford University Press, 2000.
11. **Johnson, R. & Cureton, A. (2022).** "*Kant's Moral Philosophy*", The Stanford Encyclopedia of Philosophy (Fall 2022 Edition), Zalta, E. N. & Nodelman, U. (eds.), URL = https://plato.stanford.edu/archives/fall2022/entries/kant-moral/
12. **Kant, I. (1785).** *"*Groundwork of the Metaphysics of Moral." Trans. Gregor, M. & Timmermann, J. *Cambridge Texts in the History of Philosophy*. Cambridge: Cambridge University Press. Referenced throughout as **Groundwork.**
13. **Kant, I. (1797).** "On a supposed right to lie from philanthropy." In: Gregor MJ, ed. Practical Philosophy. *The Cambridge Edition of the Works of Immanuel Kant*. Cambridge: Cambridge University Press; 1996:605-616. doi:10.1017/CBO9780511813306.014. Referenced throughout as **SRL**.
14. **Korsgaard, C. (1986).** *"*The Right To Lie: Kant on Dealing with Evil*." Philosophy and Public Affairs,* Autumn 1986, Vol. 15, No. 4, pp.325-349
15. **Korsgaard, C. (2012a).** *"*Introduction*"* in *Groundwork of the Metaphysics of Morals.* In M. Gregor & J. Timmermann (Trans.), Kant: Groundwork of the Metaphysics of Morals, Cambridge Texts in the History of Philosophy. Cambridge: Cambridge University Press.
16. **Korsgaard, C. (2012b).** "Kant's Formula of Universal Law." in *Creating the Kingdom of Ends*. Cambridge: Cambridge University Press; 1996:77-105. doi:10.1017/CBO9781139174503.004

17. **Larry, A. & Moore, M. (2021).** "Deontological Ethics", *The Stanford Encyclopaedia of Philosophy* (Winter 2021 Edition), Zalta, E. N. (ed.), URL = https://plato.stanford.edu/archives/win2021/entries/ethics-deontological/.
18. **Lloyd, S., A., & Sreedhar, S. (2022).** "Hobbes's Moral and Political Philosophy", *The Stanford Encyclopedia of Philosophy* (Fall 2022 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = <https://plato.stanford.edu/archives/fall2022/entries/hobbes-moral/>.
19. **McCarty, R. (2009).** "Kant's Theory of Action", *Oxford online edn*, Oxford Academic, https://doi.org/10.1093/acprof:oso/9780199567720.003.0001
20. **Nagel, T. (2007).** "The value of inviolability." *Morality and Self-Interest* (2007; online edn, Oxford Academic, 3 Oct. 2011), https://doi.org/10.1093/acprof:oso/9780195305845.003.0006
21. **Parfit, Derek. (1986).** "Reasons and Persons." Oxford, 1986; online edn, Oxford Academic, 1 Nov. 2003, https://doi.org/10.1093/019824908X.001.0001
22. **Parfit, D. (2011).** "On what matters" Vol. 1. *Oxford University Press*, USA.
23. **Podgorski, A. (2018). "**Wouldn't it be Nice? Moral Rules and Distant Worlds". *Noüs*, 52:2, (2018), 279-294.
24. **Rawls, J. (1971).** "A Theory of Justice." in *Ethical Theory: An Anthology*, John Wiley & Sons inc., Blackwell Publishers Ltd., pp.581-592
25. **Ridge, M. (2006).** "Climb Every Mountain?" *Ratio* 22(1), pp. 59–77, 2009.
26. **Rumbold, B. (2024).** "Does the Patterned View Avoid the Ideal Worlds Objection?." *Utilitas*, 36 (2), 130-147.
27. **Scanlon, T. M. (1984).** "Rights, Goals, and Fairness." In *Theories of Rights*, ed. Waldron, J., Oxford

28. **Schafer-Landau, R. (2013).** "Introduction to Part X" in *Ethical Theory: An Anthology*, John Wiley & Sons inc., Blackwell Publishers Ltd., pp.555-557

29. **Thomson, J. J. (1996).** "Rights, Restitution, and Risk: Essays in Moral Theory*".* (Cambridge, MA: Harvard University Press), 50., n.2.

30. **Urmson, J. O. (1958).** "Saints and Heroes". *Essays in Moral Philosophy*, ed. Melden. A. I., Seattle, University of Washington Press.

31. **Valentini, L. (2012).** "Ideal vs. non-ideal theory: A conceptual map." *Philosophy compass*, *7*(9), 654-664.

32. **Varden, H. (2010). "**Kant and Lying to the Murderer at the Door…One More Time: Kant's Legal Philosophy and Lies to Murderers and Nazis*." Journal of Social Philosophy*, 41: 403-421. https://doi.org/10.1111/j.1467-9833.2010.01507.x

33. **Wolff, J. (2007).** "Equality: The Recent History of an Idea." *Journal of Moral Philosophy,* 4(1), 125-136. https://doi-org.libproxy.ucl.ac.uk/10.1177/1740468107077389

34. **Wood, A. (2009)** "*Duties to Oneself, Duties of Respect to Others*" in The Blackwell Guide to Kant's Ethics, Chapter 11 https://onlinelibrary.wiley.com/doi/book/10.1002/9781444308488#page=237.