The Paradox of Deontology

Andreas Bruns

Submitted in accordance with the requirements for the degree of Doctor of Philosophy

The University of Leeds

School of Philosophy, Religion and History of Science

September 2021

The candidate confirms that the work submitted is his own and that appropriate credit has been given where reference has been made to the work of others.

For my parents, my grandfather,
and my brother

Acknowledgements

Many people have contributed, directly or indirectly, to the completion of this thesis. I would first like to thank my supervisors, Gerald Lang and Pekka Väyrynen. You have somehow managed to offer only encouragement at all stages of this project while still making me feel that I can do better. I owe much of this work to our extensive discussions and I am immensely grateful for your continuous and patient support.

I would like to thank the University of Leeds for funding my research and the Royal Institute of Philosophy for awarding me a Jacobsen scholarship for my final year.

Others had a direct impact on this thesis. Thanks to Susanne Burri, Daniel Elstein, David Faraci, Shaun Nichols, and Benedict Rumbold for providing helpful comments in the context of conferences, workshops, and at other occasions. I would also like to thank Helen Frowe and Richard Rowland for providing many helpful comments in the context of examination. I would like to extend my thanks to my fellow postgraduates, especially to Miriam Bowen, David Heering, and Olof Leffler. Thanks to Sören Brandes, Kevin Jones and Alex Stamatiadis-Bréhier for reading parts of this thesis prior to submission.

Doing a PhD was one of the most important experiences of my life and I am immensely grateful to the friends who accompanied me during this period. Thank you, Becca, Jacopo, Jamie, Janset, Joe, Judith, Kamilla, Kevin, Konstantin, Madda, Roberta, Sarah, Simon, Vic, the fellowship of Casa Morandi, and all the others who helped making my time in the UK very special. My warmest thanks to Leonie, Katja, and Annika for being there during the final months of my PhD and ever since. Finally, I am deeply grateful to Sören and Raphael for accompanying me, not only during the time of my PhD, but throughout the most part of my life.

This thesis is dedicated to my family: to my parents and my grandfather who have always supported my choices, and to my brother, Philipp, who has been a source of inspiration for me in many ways. Moreover, I am grateful for my other family in Leeds. Thank you, Francesca, for being one of the best friends I will ever have. And thank you, Alice, for everything. To grow is one of the most significant things two people can do together.

Thesis Abstract

This thesis develops a deeper understanding of and provides an answer to the paradox of deontology. Traditional deontological views include deontic constraints that prohibit us from harming innocent people even to prevent greater harms of the same type. Although constraints correspond to widely shared moral intuitions, they seem to make traditional deontology unavoidably paradoxical: for how can it ever be morally wrong to *minimise* morally objectionable harm?

The thesis argues that previous attempts to solve this paradox have been insufficient because they have failed to distinguish clearly between two distinct puzzles that together constitute the paradox. The first puzzle—the rationality paradox—says that if we think that we should not harm others in a certain way, it is rational by default to think that we should minimise the occurrences of that kind of harm overall. Thus, to answer the rationality paradox the deontologist must justify constraints by reference to some value that *cannot* be furthered by minimising the occurrences of harm. However, this will make her vulnerable to a second puzzle—the value paradox—which says that in the face of the severe harm that awaits the greater number of individuals, it seems morally inappropriate to be concerned with anything other than the minimisation of the occurrences of that kind of harm overall.

The thesis develops a comprehensive approach that can address both these paradoxes. The hyperinviolability account developed in this thesis shows that traditional deontology ceases to appear paradoxical once we understand it as an agent-neutral moral theory that gives priority to our moral standing over the moral significance of what might happen to us.

Table of Contents

General Introduction

It is an essential part of common moral thought that we should make the world a better place. We should do something—if not our best—to help the less fortunate, to increase the well-being of others or reduce their suffering, to make our technologies more efficient and sustainable, to strive for more social justice, and to reduce prejudices and discrimination. When we act it seems appropriate to be oriented towards the aim that the world should, to the humblest degree, be in a better state than it would have been without our doing.

It is another essential part of common moral thought that there are things that we just should not do. Murder, torture, or enslavement are good examples. These kinds of acts might be wrong even if they have good consequences. For instance, you should do something to help the less fortunate. But you should not murder your rich uncle to inherit his millions and give the money to charity. Similarly, you should save lives. But you should not harvest the organs of a healthy patient and use them to save a handful of other, dying patients. Sometimes, it doesn't seem to matter how much good you could do because of the acts which would produce these goods. Sometimes, that is, common-sense morality places non-negotiable *constraints* on our action.

This thesis is about such constraints. More precisely, it is about one feature of constraints that many philosophers have found utterly puzzling: constraints prohibit certain kinds of acts *even when an act of one such kind*

would minimise instances of the same kind of act overall. For instance, you should not murder your rich uncle even if James will otherwise murder his two half-as-rich uncles to obtain the same amount of money for charity. And you should not harvest the healthy patient's organs even if your colleague will otherwise harvest the organs of two slightly less healthy patients to obtain the organs needed. There might be a way of arguing that you do not actually make the world a better place by murdering your uncle or by harvesting the organs of a healthy patient, even if by doing so you could prevent many more deaths. Murder is among the worst things you could do, morally speaking, and as such makes the world worse even if you save more lives than you take. But what is the rationale for saying that you should not minimise the number of murders when it is certain that *some murders* will be committed in any case?

Impersonal morality, it seems, must judge a world that contains the murder of two uncles or two patients to be worse than a world that contains the murder of only one uncle or one patient. It seems appropriate, then, to say that there is a sense in which constraints require us to make the world *a worse place*—one that contains many more murders, tortures, or enslavements. This seems a rather irrational request for morality to make. The problem of how to make sense of this puzzling feature of constraints has come to be known as *the paradox of deontology*.

Of course, it is not obvious that morality must have this puzzling feature. Common-sense could simply be in the grip of an intractable misconception of what morality really entails. This is why ethical theory may diverge from common-sense and has a range of possible views to offer.

Constraintism, as I will call it, is the view that there are constraints of the kind that prohibit you from minimising the number of uncles murdered for the cause of charity. *Eliminativism about constraints* is the name I give to the view that denies the existence of constraints. On an eliminativist view, there may still be restrictions on murder and other acts of serious harm such

that you should *not* murder your uncle even to help the less fortunate. But these restrictions only hold unless breaching them is the only way to minimise the number of murders overall. Once your murder prevents James' two murders, your act of murder cannot count as wrong.

A first limitation of this thesis is that it will not give a conclusive answer to the question which of these two views—constraintism or eliminativism—*is true*. I will give reasons why we should accept a constraintist position. But I will not present a complete argument with the intention to prove the truth of constraintism. Whether we should choose to accept constraintism or eliminativism depends—as with so many questions in normative ethics—on what kind of considerations we think should be given priority in ethical theory. Rather, my main concern lies with the puzzle about constraints and hence, with *the internal coherence* of constraintism. As such, I do not intend to take on the (rather futile) task of convincing, say, a radical utilitarian—who believes that we must always, without exception, increase well-being for the greatest number—that there are constraints. However, what I have to say should be relevant even to radical opponents of constraintism as the thought that constraintism lacks internal coherence is usually a key element in their rejection of this kind of view.

In a word, this thesis gives an answer to the paradox of deontology. Constraints cease to appear paradoxical, it argues, once we understand constraintism as a theory that gives priority to the moral significance of persons over the significance of what happens to them. Constraints give expression to our elevated moral worth and, as such, are grounded in *the shared moral ideal of the inviolability of persons*.

This answer is not a new one—it is a kind of answer that has been given before. Most notably, Frances Kamm has made great efforts to show that constraints on action could be justified based on considerations about the moral status of persons as inviolable beings. I shall refer to this idea as *the status rationale* for constraints or *the inviolability account*. I take much

inspiration from Kamm's work and this thesis may be understood as an expansion of Kamm's inviolability account. Yet it should be understood as offering more than just a defence of Kamm's view.

Most importantly, I develop a novel understanding of the paradox of deontology as a two-staged, constraints-sceptical argument. The argument begins with the thought that whatever feature it is that makes murder wrong must seem to make fewer murders morally preferable to more murders. So, how can it be rational to think that murder is wrong even when it would prevent more equally wrongful murders, i.e., where it would prevent more acts which have the very feature in virtue of which murder is wrong? I call this *the rationality paradox*.

Any proposed solution to the paradox of deontology has focused solely on the rationality paradox. For instance, *the agent-centred approach*, as I call it, aims to justify constraints by reference to the idea that morality gives each of us a special concern with what *we* do. According to this approach, morality asks us, first and foremost, to ensure that we ourselves do not act in certain ways even if others do. For instance, morality requires that you do not murder your own uncle, even if that means that James will murder two of his. Morality does not endorse James' murders. And it does not hold that there are no circumstances in which you should prevent murders committed by others. It just asks you, so to speak, to have your own moral house in order before attending to any other morally relevant business.

The agent-centred approach might be able to avoid the rationality paradox. If what you ought to do morally is to avoid murdering anyone *yourself*, then we have an explanation as to how it can be rational that you do not murder your uncle even where this would lead to more murders overall. However, this is only the first part of the paradox of deontology—the first stage of the constraints-sceptical argument. I believe that one reason why previous attempts to solve the paradox have been found unsatisfactory is that they

have failed to clearly address its second part, which I refer to as *the value paradox*.

The value paradox holds that where the harm the agent could do and the harm she could prevent are of the same type, it seems inappropriate to focus on anything other than the minimisation of that type of harm overall. For instance, we could just assume that there are agent-centred values which make it preferable from the agent's perspective that *she* does not commit murders, even if this would lead to many more murders overall. But how can it be appropriate—*morally* appropriate, that is—to be so concerned with such agent-centred values in the face of the greater harm one could prevent, especially where the greater harm is of the same type and thus directly comparable to the lesser harm? The agent-centred approach, it seems, simply fails to account for a central feature of moral values as something that is valuable *beyond the agent's limited perspective*.

This is only one face of the value paradox. As we will see, the value paradox has many faces. This is so because its shape depends on how we choose to answer the rationality paradox. For instance, the inviolability account also faces a variant of the value paradox. Suppose that constraints give expression to a certain moral status—inviolability—which is valuable in itself and which we lack if there are no constraints. According to the inviolability account, permission to murder someone in order to minimise the number of murders in total would not further the relevant value but would rather deny it. If plausible, this move avoids the rationality paradox. But again, the constraints-sceptic can ask: how can it be appropriate to be so concerned about our inviolability status, where this would mean that many more of us *will* actually be violated?

Whatever answer the constraintist might provide to avoid the rationality paradox, this will make her vulnerable to the value paradox which, in turn, will draw her back into the grip of the paradox of deontology. One central insight of this thesis is that in order to escape the paradox of deontology,

we must give a comprehensive answer that can address *both* the rationality and the value paradox. My own account—*the hyperinviolability account*—aims to break out of the constraints-sceptical cycle by providing such a comprehensive answer.

As such, this thesis takes certain ideas for granted and it might be useful to make two of these ideas explicit here. First, I take the idea for granted that advocating a paradoxical view is an issue—at least in philosophy—and that someone who is accused of advocating a paradoxical view has two general options: either to find a plausible explanation that de-bunks the charge of paradox; or—especially if the first option seems impossible to achieve—to give up on their original view in favour of a coherent alternative position. This thesis represents a choice of the first option. However, it should be mentioned that there is nothing about the claim that constraints appear paradoxical that could force the constraintist to choose between these options.

This brings me to the second idea this thesis takes for granted. A thesis focused on providing an answer to a single problem in analytical philosophy is bound to express the view that the problem itself is one worth addressing. However, some constraintists might think that the paradox of deontology is *not* a problem worth addressing. In particular, they might think that constraints only appear paradoxical from the perspective of a certain ethical tradition—like classical utilitarianism—and that it cannot be the task of non-utilitarians to make sense of parts of non-utilitarian ethical theory from the perspective of utilitarianism. Of course, there might not be a clear line between what counts as rejecting a problem altogether and what counts as proposing a solution to it. But this thesis takes for granted that the paradox of deontology *is* a problem, and that it is not just a problem if viewed through the lens of an alien ethical tradition. Instead, it aims to understand the paradox, in its strongest form, as *an internal problem* of constraintist views.

The thesis will be structured as follows. *Chapter One* analyses the historical and systematic origins of the paradox of deontology as well as the

general options of how to approach it. I illustrate the importance of the distinction between the rationality and the value paradox and argue that the agent-centred approach is insufficient because it lacks conceptual means to address the value paradox.

Chapter Two lays the foundation for an alternative to the agent-centred approach. My impression is that any convincing solution to the value paradox will have to depart from the idea that constraints give any special significance to what the agent herself does. (I will argue for this impression later.) Thus, I propose that an answer to the paradox of deontology should take what I call *the agent-neutral approach*.

The agent-neutral approach aims to justify constraints without reference to the idea that morality asks us to give priority to our own actions. Instead, it aims to justify constraints as part of a moral view that gives shared moral aims to all agents. However, it is most common to think that we *must* refer to agent-centred values in order to make sense of the peculiar normative force of constraints. The chapter rejects this *standard view*, argues that we can make sense of constraints in solely agent-neutral terms, and that constraintists have good reasons for favouring an agent-neutral account of constraints.

Chapter Three introduces the central ideas of the moral status rationale for constraints. I reconstruct Kamm's inviolability account as a systematic answer to the rationality paradox and address various issues with Kamm's own account.

Chapter Four develops Kamm's view further into the hyperinviolability account and provides an answer to the value paradox. The major objection against Kamm's account is that it cannot fully justify constraints because there are other dimensions to our moral worth than our inviolability. Most importantly, constraints express the view that we are *more inviolable*, but they also express the view that we are *less saveable* because if there are constraints, then morality does not require that we are saved under certain

circumstances. I will refer to this objection as *the saveability challenge* and show how the saveability challenge can be interpreted as a variant of the value paradox. I then present a novel response to that challenge which aims to show that it fails as an internal criticism against the inviolability account.

Chapter Five finally examines the relationship between constraintism and consequentialism. Constraints have usually been understood as a distinguishing mark of non-consequentialist or deontological ethics. But since the advancement of the idea of *consequentialising*, some have argued that constraints can be given a consequentialist reinterpretation. The question whether there is a plausible consequentialist account of constraints is relevant for two reasons. For one thing, examining the possibilities to consequentialise constraintism will further our understanding of the current debate about the practice of consequentialising, its ramifications, and its limits. More importantly, for another thing, a plausible consequentialist account that would justify constraints might provide a powerful alternative to the hyperinviolability account. I will argue that there is no such plausible account because none of the different versions of consequentialised constraintism can successfully avoid the value paradox.

1 *Origins of the Paradox of Deontology*

1.1. Introduction

All moral theories say that there are times when it is wrong to kill, torture, or otherwise seriously harm innocent people. Only some moral theories, however, say that there are times when it is wrong to kill, torture, or otherwise seriously harm the innocent *even if* by doing so the agent could prevent comparably greater harm. That is, only some moral theories are versions of a normative ethical view what I shall call *constraintism*.[1]

Constraintism appeals to widely shared moral intuitions. Research into the psychology of moral decision making suggests that people find it hard to endorse certain kinds of acts, even in situations where outcome-based evaluations suggest their preferability (Cushman 2015). But despite its being a central feature of ordinary morality many philosophers have come to find constraintism to be a deeply puzzling view. For how can it be wrong, say, to prevent a greater number of killings by committing a single killing if *whatever it is* that makes killing the innocents morally objectionable seems *to make it*

[1] This version of moral theory is often simply called *deontology*. I prefer to introduce the new name constraintism because there are both deontological views that do not subscribe to constraintism as well as consequentialist views that aim to accommodate it (see Section 1.3).

worse if there are more rather than fewer killings in total? Constraintism systematically requires us to make the world a worse place—one that contains many more killings or tortures than it would contain if such acts were permissible under the relevant type of circumstances. How can it be rational for an ethical theory to make any such request?

For this reason, constraintism has been said to be surrounded by a "distinct air of paradox" (Scheffler 1985: 409). This chapter investigates the historical and systematic origins of this paradox. Section 1.2 develops an understanding of constraints, their source, the kinds, and the contents of constraints. Section 1.3 clarifies the relation between constraintism and the consequentialism/deontology distinction in normative theory. Section 1.4 analyses the original formulations of the paradox of deontology by Robert Nozick and Samuel Scheffler and argues that both are insufficient to show that there is any *deontological* paradox surrounding constraintism.

Section 1.5 develops an alternative understanding of the paradox in terms of a conflict between two kinds of reasons against rights violations. Section 1.6 argues that there is no easy way out of this conflict that would set off from the claim that reasons against rights violations are agent-centred or agent-relative in a substantive sense. Moreover, it identifies the two parts of the paradox—the rationality paradox and the value paradox—and lays out the requirements for an alternative, agent-neutral approach to the paradox. And finally, Section 1.7 aims to develop a deeper understanding of the paradox by identifying the constraints-sceptical dialectic that has been nurturing its persistence in moral philosophy.

1.2. Understanding Constraints

Constraintists do not believe that we should never prevent harm, nor that it cannot ever be morally right to cause harm to prevent comparably greater harms. Instead, constraintists believe that it is *sometimes* morally wrong to inflict harm *even to* prevent comparably greater harms. In other words, they

believe that morality sometimes places a deontic constraint on our action.[2] Recall the two examples from the General Introduction to this thesis:[3]

> **Inheritance.** You could murder your rich uncle to inherit his millions, give the money to charity, and save many others from starvation.
>
> **Transplant.** You are a surgeon giving a healthy patient a routine check-up. You have five other patients waiting for an organ transplant. Since they are waiting for different organs, it so happens that you could save each of them by harvesting the organs of your healthy patient.

To say that you ought not to kill in either case is to say that there is a deontic constraint on killing the one even to prevent the deaths of many. That is, constraintists believe that the duty not to kill exhibits a certain stringency. It is impermissible to kill *even when* this would prevent many more deaths than lives you would take.

Perhaps, a deontic constraint on killing in *Inheritance* and *Transplant* could be justified by reference to the thought that killings are worse than deaths. You should not kill because it is worse, morally speaking, when you take someone's life than when you fail to prevent even a greater number of deaths. (Though one might ask why, exactly, the killing of your uncle should be worse than the starving of the many.) But constraintists believe that the

---

[2] Nozick (1974) has introduced them as side constraints, whereas Scheffler (1985) calls them agent-centred restrictions. Others speak of agent-relative restrictions or constraints (Brink 2006, Moore 2008, Emet 2010, Lippert-Rasmussen 2009) or of deontic or deontological restrictions or constraints (Brand-Ballard 2004, Oberdiek 2008, Alm 2009, Chappell 2011, Otsuka 2011, Johnson 2019). As a matter of terminological choice, I shall call them deontic constraints and use the term *restrictions* to refer to restrictions on the agent's conduct more generally. Thus, deontic constraints are a proper subset of deontic restrictions (Otsuka 1997: 202 fn5). As a matter of *significant* terminological choice, I refrain from calling constraints agent-centred or agent-relative to avoid the presumption that they are *in fact* agent-centred or agent-relative in any substantive sense; more on this in Chapter 2.

[3] The *Inheritance* case is used by Kagan (1989): 4. The *Transplant* case goes back to a famous example introduced by Thomson (1985): 1396.

duty not to kill is even more stringent than that. They believe that you should not kill even to prevent *more killings*, like in the following two cases:

> Inheritance Paradox. You could murder your rich uncle and use the inheritance money to save many others from starvation. If you don't do it, it is certain that James will murder his two slightly less wealthy uncles to obtain the same amount of money to save the same amount of people from starvation.
>
> Transplant Paradox. You could harvest a healthy patient's organs and use them to save other five dying patients. If you don't do it, your colleague will harvest the organs of two other slightly less healthy patients to obtain the organs needed and use them to save the five.

If killings are worse than deaths, we should think that you must go to greater length to prevent them. Yet constraintists believe that you should not kill even in *Inheritance Paradox* and *Transplant Paradox*. Even the prospect that you could prevent more killings is insufficient to justify an act of killing.

From here on, a deontic constraint is therefore understood as any moral principle that takes the following form:

> Deontic Constraint. It is impermissible to φ even to prevent more further φ-ings.[4]

Note the significance of the phrase *even to*. A constraint on killing may prohibit you from killing in many cases such as *Inheritance* and *Transplant*. But to qualify as a constraint it must prohibit killing *even when* this would prevent

[4] The prime case of a constraint is one that prohibits us from harming *innocent*, non-threatening people. A small subset of constraintists, however, might believe in pacifist constraints which prohibit us from harming a threatening aggressor to prevent that he harms his victims. While such pacifist constraints clearly quality as constraints and should therefore not be excluded from the definition, it is worth noting that many constraintists will reject constraints for cases of self- or third party-defence.

more further killings, and even when these further killings are *wrongful* killings (Otsuka 1997: 202).

1.2.1. The Paradoxical Nature of Constraints

A constraint on killing an innocent person to prevent the deaths of many others might appear puzzling. But the case that reveals the killing-constraint to be most puzzling is the case in which it prohibits killing even when this would prevent further killings, that is, the case in which killing would *minimise the total number of killings*. For whatever it is that makes an act of killing morally objectionable must make more killings seem worse than fewer killings. How can it be wrong to kill to minimise killings overall?

Transforming the *Inheritance* and *Transplant* cases into the *Inheritance Paradox* and *Transplant Paradox* cases has given the acts on both sides the same type thereby making it hard to see how your killing could be more significant than a greater number of other killings. This strategy[5] can be used to reveal the particularity of moral duties that take the form of deontic constraints as well as their puzzling nature. Therefore, I shall refer to cases like *Inheritance Paradox* and *Transplant Paradox* as *paradox cases*.

It is natural to have certain reservations about the construction of paradox cases. After all, it is not obvious what the connection is between your potential killing and the killings potentially committed by others. If your killing would prevent James' two killings, we would expect that there is some causal connection between the two sets of killings. But all that seems to connect them is that James has decided that he will kill two if you refuse to kill one. There is no reason to think that once you refuse to kill, *it is necessary* that two

[5] Brand-Ballard (2004) calls this strategy *equalisation*. As he notes, it is a strategy traditionally used by critics of constraintism in order to reveal its puzzling implications.

further killings occur because James, too, could simply do the right thing and refuse to kill.

It might therefore be useful to mention one other paradox case, in which the connection between your potential killing and the killings you could prevent is more obvious:

> **Footbridge Paradox.** A villain has tied five strangers to the tracks, started an electric trolley that is now heading towards them, and afterwards fled the scene. You could save each of the five by pushing a massive sixth stranger off a footbridge. He would die upon hitting the ground, but his body would block the tracks.[6]

In *Footbridge Paradox* there is an obvious causal connection between your killing of the massive man and the saving of the five. Since the electric trolley is already on its way to run them over and the villain who set it off has fled the scene, it is necessary that they die unless you decide to kill to save them.

Yet, first and foremost, who is in the wrong is the villain—just like James would have been in the wrong if he decided to murder his two uncles. In *Footbridge Paradox*, it might be more obvious why your killing would be necessary to prevent more killings. But it remains a case of partial compliance. If everyone were to comply with the duty not to kill, you would not have to consider whether to kill to prevent more killings. From this perspective, constraints often take the form of prohibitions against *re*-acting in a certain way to the wrongdoing of others (Kagan 1989: 47).

This thesis focuses on the kind of paradox cases described above. If we can clear the air of paradox surrounding a constraint on *φ*-ing even to prevent more further *φ*-ings, I believe we have addressed the case which reveals that constraint to be most puzzling, and thus have cleared the air of

---

[6] This case is a variation of the famous massive man case introduced by Thomson (1985): 1409.

paradox surrounding the concept of that constraint itself. However, it is worth noting once more that it would be a mistake to think that the feature that it prohibits φ-ing even to prevent more further φ-ings is what a constraint on φ-ing is all about. Properly understood, a deontic constraint is nothing more than an implication of how constraintists think of certain moral duties more generally. Although a constraint on φ-ing exists only where it is impermissible to φ even to prevent more further φ-ings, the existence of such a constraint implies that it is impermissible to φ under a large set of circumstances where φ-ing would have desirable consequences.[7]

1.2.2. Rights-as-Constraints

What is the source of a constraint on φ-ing? I believe the reason why we should think that there are deontic constraints is that individuals have rights.[8] In general, rights are entitlements to something hold by an individual person, i.e., the right holder (Wenar 2021). But as a source of constraints the relevant kind of rights will primarily be entitlements that others do not perform certain acts, and thus, *negative rights* or *rights against interference*. As we will see in the next section, however, it might be that also *positive rights*—entitlements that others perform certain acts—may act as constraints upon action.[9]

Persons have rights not to be killed, tortured, enslaved, and so on. As rights theorists, constraintists believe that sometimes these rights act as

---

[7] Note that we may also find cases of constraints that do not involve unequal numbers but unequal degrees of harm. Suppose you could lightly torture Joe to prevent a torturer from gravely torturing Jim. A torture-constraint might prohibit the light torture of Joe.

[8] Here, I follow the tradition of treating constraintism as a certain way of interpreting individual rights. However, it should be noted that constraintists are not, qua constraintists, committed to a rights-based theory. One can imagine constraintist views whose central category is not that of a right. For instance, virtue ethics might seem like a natural soil for constraintist views. Virtue-based constraintism would centre around the idea that the virtuous agent does not commit certain kinds of acts, even in a non-ideal world in which others commit a greater number of them.

[9] On the distinction between negative and positive rights see also Foot (1967), Thomson (1985), Kamm (1992), and Draper (2005).

constraints on the prevention of greater harms. More precisely, they believe that it is sometimes impermissible to violate a right even to prevent more extensive violations of the same right in others. I refer to such rights as *constraining rights*:

> **Constraining Right.** *R* is a constraining right iff it is impermissible to violate *R* even to prevent more extensive violations of *R* in others.

Where a violation of a right would prevent more further violations of the same right, I shall refer to it as *a minimising violation* of that right (Lippert-Rasmussen 1996). A right *R*, then, is a constraining right just in case that it is impermissible to commit minimising violations of *R*.[10]

Often, the question whether there is a constraint on *φ*-ing will depend on the question whether individuals have a constraining right against *φ*-ing. Whether it is wrong to prevent five killings in *Footbridge Paradox*, for instance, depends on the question whether the massive man has a constraining right against killing.

Some philosophers think that rights *just are* constraining rights. Those philosophers would insist that the impermissibility of minimising violations is built into the concept of a right. This conceptual limitation makes sense if we think about how the concept of rights is used in contemporary moral and political philosophy. When people discuss the status of human rights, for instance, what they might be interested in are "the rights of individuals not to be violated, sacrificed, or used in certain ways, even in the service of valuable

---

[10] I take the concept of a constraining right from Kamm (2001): ch. 10. Sometimes, a distinction is drawn between rights violations and rights infringements. An action that opposes someone's right constitutes an infringement of that right. For an action that opposes someone's right to constitute a rights violation, the action must moreover be wrong (Thomson 1990, Oberdiek 2004). In the context of this distinction, it does not make sense to ask, "Is it wrong to violate right *R*?" For the use of the term violation rather than the term infringement would already answer the question with Yes. For the sake of simplicity, I shall avoid the infringement/violation distinction and instead speak of a rights violation wherever an action infringes a *prima facie* right.

ends" (Nagel 2008: 102). And one such valuable end might be the minimisation of violations of the same human rights in everyone. I, too, find it hard to make sense of the idea that there could be something like a human right against torture which does *not* make it impermissible to torture us even to prevent more others from being tortured. However, I want to leave room—conceptual room at least—for alternative understandings of rights that do *not* give rise to constraints on action. In other words, I take it that there is at least conceptual room for something like *a non-constraining right*:

> **Non-Constraining Right.** *R* is a non-constraining right iff it is permissible to violate *R* to prevent more extensive violations of *R* in others.

Again, some might be sceptical about calling such properties rights in the first place. But one way to understand the task at hand is that we are trying to uncover how it can be rational to favour the interpretation of rights as constraining rights over their interpretation as non-constraining rights (Heuer 2011: 39–40).

1.2.3. Which Kinds of Actions Are Constrained?

Are all acts which are usually wrong subject to a deontic constraint under conditions of partial compliance? The examples most frequently discussed in the literature are constraints on killing, torture, and promise breaking. If these constraints exist then are there any constraints, for instance, on stealing and damaging someone's property, on humiliation and bullying, or on using racial slurs?

My impression is that it is unproblematic to think that if there are constraints on killing and torture, then constraints exist also for the other action types mentioned above. For instance, it seems reasonable to think that the prospect of being able to minimise the total number of racial slurs being used cannot justify that I use a racial slur myself; or that the prospect of being able

to minimise the total number of individuals being humiliated cannot justify that I humiliate someone. I shall not try to give a full account of all types of actions possibly subject to a deontic constraint. In the context of what I have to say, the question which acts are constrained will however be relevant in two respects.

First, the question is relevant to the intuitive appeal of constraintism. Some philosophers might think that it is generally right, say, to kill or torture someone where this would prevent more killings or tortures.[11] But I suspect that these philosophers might be less prepared to claim that it can ever be right to, say, rape or enslave another person to prevent more rapes or enslavements. Of course, this is a suspicion rather than an argument. But my impression is that moral philosophy is morally detached, so to speak, from the topics of killing or torture in a way that it is not detached from the topics of rape or enslavement. In any case, it should be clear that if constraintism is false, then not only would it be right to kill or torture someone in order to prevent more killings or tortures, but the same would have to be true for acts of rape and enslavement. And I suspect that even the harshest critics of constraintism would find themselves troubled by this implication. Thus, how strong the intuitive appeal of constraintism is seems to depend on the action types suggested to be constrained.

Second, the question which acts are constrained might also be relevant when talking about the paradoxical character of constraintism. Heuer (2011) argues that the paradox of deontology does not arise for any action type subject to a deontic constraint. Instead, the paradox would only concern certain puzzling aspects of the ethics of killing. If this is true, then it is inappropriate to say that *constraints* are paradoxical. I shall discuss Heuer's view in more detail in Section 1.6. Others seem to think that the paradox lies with the very idea of deontic constraints itself, notwithstanding their content (e.g.,

[11] However, as we will see in the next section, it is far from obvious that many critics of constraints would want to commit to such claims.

Dougherty 2013: 534). My impression is that whilst the paradox of deontology concerns the very concept of a deontic constraint, the content of that constraint will make a difference once constraintists begin to respond to the charge of paradox. Here, constraintists will have different conceptual resources available depending on the content of a constraint. As we will see in Chapter 2, for instance, constraintists can refer to agent-relativity when justifying a constraint on promise-breaking but might not have the same conceptual resource available to justify a constraint on killing.

1.2.4. Types of Constraints

The nature and form of a constraint may depend on the kind of constraint in question. On a first level of distinction, it is essential to distinguish between two types of constraints which I shall call *general* and *special* constraints.

My main examples so far have been constraints on killing and torture. These constraints are *general* since they hold between everyone and anyone else and are not restricted on any further conditions. However, it is commonly held that there are also *special* constraints. Suppose, for instance, that Max has made a promise. Unfortunately, Max can keep her promise only if Chloe breaks two. It might seem reasonable to say that Max's priority should lie with keeping her own promise rather than seeing to it that the greater number of promises are kept. But such a constraint on promise breaking is *special* in the sense that it is conditional to the fact that Max has promised something thereby committed herself to it.

Thus, a constraint on breaking one's promise seems different from the general constraints on killing or torture since its existence is restricted on the further condition that the agent has made that promise. It is *special* in the sense that it arises from the special, as opposed to general, nature of the

agent's commitment.[12] The distinction between general and special constraints is congruent with H. L. A. Hart's distinction between special and general rights, whereby special rights, in contrast to general rights, arise "out of special transactions between individuals or out of some special relationship in which they stand to each other" (Hart 1955: 183).

On a second level of distinction, constraints may arise from either *negative* or *positive duties*. Roughly, a *negative* duty exists where the agent ought not to do something, a *positive* duty where she ought to do something (Singer 1965). It might be natural to associate constraints, first and foremost, with negative duties. This is so because the agent confronts a constraint as her negative duty and thus in the shape of what she ought *not* to do.

Sometimes, however, a constraint might arise from a positive duty. By way of illustration, consider the following case:

> Little Brother. Sean could save his little brother Daniel from the influence of a dangerous cult, but only if he does not help Jacob save his two little brothers from the influence of the same cult.

It might seem reasonable to say that Sean's priority should lie with saving *his own* little brother rather than seeing to it that the greater number of little brothers are saved. Thus, Sean might be subject to a (special) constraint that prohibits him from helping Jacob to save his brothers. But that constraint arises not from a negative duty, but from Sean's positive duty to care for Daniel.

Thus, whereas constraints are, by definition, a proper subset of negative duties, also a positive duty may give rise to a constraint on someone

[12] Heuer (2011) and Setiya (2018) also recognise the importance of the distinction between general and special constraints in this context. Portmore (2013a) proposes to separate agent-centred restrictions (what I call general constraints) from special obligations (what I call special constraints). I shall say more on the distinction in Chapter 2.

else's action.[13] On a rights-based view, constraints are negative or positive by virtue of whether it is a negative or a positive right that acts as a constraint on the agent's conduct. For instance, Daniel's positive right to be cared for acts as a positive constraint on Sean's action.

Thus, we can distinguish between four kinds of constraints based on the kind of moral commitment they arise from:[14]

> General Constraint. Arises from a general duty (not) to treat others in a certain way.
>
> Special Constraint. Arises from a special duty (not) to treat someone in particular in a certain way.
>
> Negative Constraint. Arises from a negative duty not to harm others in a certain way.
>
> Positive Constraint. Arises from a positive duty to aid others in a certain way.

Note that these kinds of constraints distribute over two levels of distinction. On a first level, each constraint is either general or special. On a second level, each constraint is either negative or positive. A constraint according to which you ought not to kill in *Footbridge Paradox*, for instance, is a *general-negative*

---

[13] Note that it might not be obvious how constraints from positive duties are to be formulated. I suggest: It is impermissible for Sean to fail to save one little brother, even to prevent that someone else fails to save two little brothers. As we will see in Chapter 2, a reference to the fact that the first little brother is Sean's brother must feature in an explanation as to why Sean's conduct is constrained in this way. The literature on constraints has largely overlooked cases in which constraints may arise from positive rights; an exception is Lippert-Rasmussen (1996): 346.

[14] In the next Chapter, we will learn about two further kinds of constraints, i.e., agent-relative and agent-neutral constraints. I choose not to make an ad hoc distinction here because according to the standard view in moral philosophy *all* constraints are agent-relative. (Hence, the definition of constraints as agent-centred restrictions.) In Chapter 2, I will defend the minority view that there are agent-neutral constraints, but this view will have to be motivated more carefully than it could be done in the context of the present chapter.

constraint. A constraint that prohibits Sean from saving Jacob's brothers rather than Daniel is a *special-positive* constraint. A constraint on twisting your child's arm to prevent five other children from getting their arms twisted would be *special-negative*. Such a prohibition might have special normative force even compared to a general constraint on twisting anyone's arm to prevent more extensive arm-twisting.

It is less obvious that there are *general-positive* constraints. If there are, their existence might depend on the temporal unfolding of events. To see why, consider the following case:

> Overlooked Castaways. While out on your boat, you see a castaway pedalling awkwardly in the water. He cannot swim well and is about to go under. You steer your boat towards him but, nearly there, you notice two other castaways in the other direction. Both seem equally unable to swim and about to drown. You could save them, but only if you turn around immediately and leave the first castaway to die.

There was a moment in time—before you set off to save the first castaway—when you might have been required to save the other two. But given that you already set off to save the first, it seems plausible to say that there is a general-positive constraint that prohibits you from turning around to save the others. Henceforth, whenever I say constraint, I mean a general-negative constraint. Where I have something to say about special or positive constraints, I will specify further which kind of constraint I am referring to.

Until now, I have used examples of paradox cases where the agent could φ to prevent φ-ings committed by *someone else*. While this is the most obvious kind of paradox case, it is important to note that there is another kind. It is commonly held that constraintists are committed to the idea that

the agent ought not to φ even to prevent more φ-ings *committed by herself*. Consider the following case:

> **Bomb Paradox.** You have set up a bomb in a busy mall. Five strangers are within its burst radius when you are struck with remorse. You want to defuse the bomb. But at this point, the only way to achieve this would be to push a sixth innocent stranger onto it.[15]

In *Bomb Paradox*, the question is whether it can be right for you to kill to prevent *yourself* from committing a greater number of killings. It is commonly held that a deontic constraint on killing would also prohibit the agent from minimising her own acts of killing. On a third level of distinction, we can therefore distinguish between *interpersonal* and *intrapersonal* constraints:[16]

> **Interpersonal Constraint.** It is impermissible to φ even to prevent other agents from committing more further φ-ings.
>
> **Intrapersonal Constraint.** It is impermissible to φ even to prevent oneself from committing more further φ-ings.

Finally, there might be constraints on doing harm even to prevent greater harm *of other types*. Suppose that you could prevent five tortures by committing a single act of killing, or that you could prevent five killings by committing a single act of torture. If there are two separate constraints—a

[15] I take this case from Kamm (1989): 225.

[16] The distinction is recognised by Johnson (2019) as well, although not as a distinction between kinds of constraints but between an interpersonal and an intrapersonal paradox of deontology.

constraint on killing and a constraint on torture—it seems plausible to think that there could be a constraint of the following form:

> **Cross-Type Constraint.** It is impermissible to φ even to prevent more further ψ-ings.

The case of cross-type constraints raises delicate questions about the comparative weigh of different kinds of harms. If it is wrong to kill even to prevent two further killings, is it still wrong to kill even to prevent five people from being gravely tortured for a year? More importantly, such cross-type cases seem to reveal that some constraints are much less stringent than others. Compare a constraint on torture with one on promise-breaking. A constraint on torture might imply that it is wrong to torture someone even to prevent two killings. But it is less obvious that a constraint on promise-breaking implies that it is wrong to break a promise even to prevent two killings. In fact, even on an absolute view it might be justified to break a promise in case that doing so would prevent two killings.

What does this mean? Why do some constraints apply across action types while others do not? As I argue in more detail in Section 2.5.2, this has to do with the different ways to justify a constraint. A special constraint on promise-breaking may be justified by reference to the agent-relative nature of the commitment to keep one's promise and is easily defeated by the prospect that one could protect something of great agent-neutral value—for instance, by preventing killings or tortures.

1.3. Constraintism and Ethical Theory

Note that, according to my understanding of constraints, anyone who thinks that it is wrong in a certain case to φ even to prevent more further φ-ings endorses a constraint on φ-ing at least for that specific case. In turn, someone might deny the existence of constraints altogether and adopt a position

which I shall refer to as *eliminativism about constraints* or *eliminativism*, for short:

> **Eliminativism.** There are no deontic constraints.

First and foremost, eliminativism is a possible position in the logical space of ethical theory. Most of those opposing constraintist views or questioning their rationality would be more accurately described as critics of constraintism or constraints-sceptics.[17] They do not explicitly endorse eliminativism. And in the end, some of them might be more sympathetic to constraintism than to eliminativism. What constraints-sceptics are sceptical about is less the claim that there are times when it is wrong to φ even to prevent more further φ-ings but more the possibility that there could be a convincing rationale for a deontic constraint as a general moral principle.

1.3.1. Absolute and Moderate Constraintism

Moreover, those endorsing a constraint on φ-ing may do so in either an *absolute* or a *moderate* fashion:

> **Absolutism.** It is always impermissible to φ even to prevent more further φ-ings.

> **Moderatism.** It is sometimes impermissible to φ even to prevent more further φ-ings.

---

[17] The list of constraints-sceptics includes, among others, Parfit (1984), Scheffler (1985), (1994), Kagan (1989), (1991), Cummiskey (1990), Bennett (1998): ch. 10, Lippert-Rasmussen (1996), (2009), Pettit (2000), and Otsuka (2011). Among these authors, Kagan's view might be the most fitting example for eliminativism. In *The Limits of Morality*, Kagan aims to deliver "a sustained attack on two of the most fundamental features of ordinary morality"—one of these are deontic constraints—which, in turn, he hopes will provide indirect support for consequentialism, conceived of as a view that denies both these features of ordinary morality (Kagan 1989: xii).

What I call moderatism here has sometimes been described as a kind of threshold theory holding that there are constraints but, at the same time, that there is a numerical threshold of preventable *φ*-ings above which those constraints give way.[18] Accordingly, absolutism would be defined as the view that there is no such threshold to deontic constraints—that they make it impermissible to, say, kill an innocent, no matter how many further killings the agent could thereby prevent.

However, it should be noted that absolute and moderate views do not have to be concerned solely with numbers. A moderate constraintist might hold that you should not lightly torture Joe even to prevent a torturer from torturing Jim but that once Jim's torture reaches a certain degree of gravity this might tip the scale and you should torture Joe. That means, moderatists are sensitive to the degree of harm you could prevent whereas absolutists believe that preventing no degree of torture could justify your torture of Joe.[19]

Moreover, some might argue that moderatism should not be described as a constraintist view at all. Instead, moderatists would believe that whereas doing harm is worse than allowing it, the degree of harm one would allow may sometimes justify doing lesser harm of the same type. They would not deal in constraints but in considerations about the gravity of harm. However, according to my understanding of constraints, in any case where the moderatist thinks that it is wrong to *φ* even to prevent more further *φ*-ings

---

[18] Some speak of *threshold deontology* in this context; see e.g., Moore (2019) and (1997): ch. 17.

[19] This might make absolutism seem like a rather implausible position. Note, however, that an absolutist could argue that what you are doing to Joe needs to reach a certain degree of gravity first before it is even appropriate to say that you would torture Joe. An absolutist might claim that you may lightly twist Joe's arm to prevent the torture of Jim because lightly twisting Joe's arm doesn't constitute an act of torture. But you may not pull Joe's fingernails, say, to prevent the torture of Jim, no matter how gravely Jim would be tortured. That is, absolute constraintists are not insensitive to degrees of harm. They only believe that preventing any degree of harm cannot justify doing any lesser degree of *the same type of harm*.

she, by definition, endorses a constraint on *φ*-ing. Whether this makes her a *constraintist* or not may be a matter open to discussion. In what follows, when I speak of a constraintist I mean someone who thinks that it is *generally* wrong to *φ* even to prevent more further *φ*-ings—for instance, to kill an innocent even to prevent ten further killings—even though she might endorse cases in which killing is justifiable—for instance, to prevent a hundred further killings. This would make her a *moderate* constraintist. An absolute constraintist, by contrast, believes that killing would be wrong even in the latter case.

The main advantage of moderate views is thus that they can accommodate intuitions about moral catastrophes and dire consequences. As we will see in Chapters 4 and 5, however, constraintists might also have good reasons to embrace an absolute view if they want to justify constraints.

1.3.2. Deontology and Consequentialism

It might come naturally to think of the distinction between constraintism and eliminativism in terms of the distinction between deontology and consequentialism. However, this would be a mistake.

Consequentialism, as I use the term, is a family of ethical theories that concern the good. Roughly, (act-)consequentialists argue that we should promote good outcomes. However, the focus on outcomes or consequences rather than acts themselves does not play an important role in contemporary consequentialism. As many consequentialists have pointed out, we may understand outcomes so broadly that a description of an act's outcome may include all kinds of features that are not typically considered features of an act's outcome or consequences.[20] Instead, what matters is the focus on

[20] Some consequentialists go as far as to claim that whatever features of an act one might hold relevant when determining the rightness or wrongness of acts can be conceived of as a feature of the act's outcome. This might mean that all moral theories could be represented as a version of consequentialism; more on this in Chapter 5.

goodness. Consequentialism conceives of moral rightness as *a function of moral goodness*. It is the view that the good is fundamentally prior to the right in that we first need to uncover the value of its outcome before we can determine whether an act is right, wrong, permissible, impermissible, and so on.

Consequentialism, then, is only constrained by what John Broome calls "the *structure of good*" (Broome 1991: 11). To be able to say that *φ*-ing is right, the consequentialist must be able to say that *φ*-ing is sufficiently good. To consequentialists, the right and the good are two sides of the same coin. I have more to say about consequentialism and its connection to constraintism in Chapter 5.

Non-consequentialism, or deontology,[21] is the view that moral rightness is *not* merely a function of moral goodness. Generally, deontologists give some room to considerations about which acts produce better outcomes. But deontologists are not committed to saying that the rightness of acts is solely determined by the relative goodness of their outcomes. Unlike on consequentialist views, on deontological views the right and the good *may come apart*.

It might come naturally to think of constraintism as a deontological kind of view. As Broome says, the constraintist argument against minimising violations may go "directly to what […] I ought not to do, without first estimating the goodness of the alternatives" (Broome 1991: 9). In turn, it is far from obvious what a consequentialist argument against minimising violations would look like. In order to make sense of constraintism in terms of the good, we would have to make sense of the claim that, other things being equal, it is sometimes *better* if there are more killings or tortures in total. Thus, on the

[21] I will follow the convention here to equate non-consequentialism with deontology, although this is, strictly speaking, inappropriate. Virtue ethics, for instance, is a non-consequentialist theory since it says that the manifestation of moral virtues in the agent's character is fundamentally prior to considerations about the outcomes of her acts. But deontology, if the terms is to be in the right place, tells us what we ought *to do* rather than what we ought *to be* (e.g., Alexander and Moore 2020).

face of it, constraintism is a central feature of deontological ethics and a counterexample to consequentialism.

This is a systematic point. But it also explains why—on a historical note—the classical constraintists come from a non-consequentialist, critics of constraintism from a consequentialist tradition. On the one hand, Kant thought of moral value as valuing actions for their own sake rather than their contingent consequences, and thus described morality itself as a system of deontic constraints (Heath 2008: 5). Taurek (1977) famously argued that it is a mistake to think of harms to separate persons as *greater* harm, thereby providing the grounds for a rejection of even the initial description of paradox cases as cases in which killings or tortures would *prevent something worse*.

On the other hand, it is not so obvious that morality places constraints on our action if, as Mill said, acts are right solely "in proportion as they tend to promote happiness, wrong as they tend to produce the reverse of happiness" (Mill 2002: 7). As it usually results in the coming about of much more harm overall, it is hard to see how a refusal to violate a deontic constraint could be said to promote overall happiness, defined in terms of the presence of pleasure and absence of pain.

For these reasons, many philosophers might identify constraintism with deontology or non-consequentialism. I think this is a mistake—at the very least, it is an inappropriate way of handling the consequentialism/deontology distinction. Contemporary moral theory knows both deontological views that are eliminativist about constraints as well as consequentialist accounts of constraintism. For instance, David McNaughton and Piers Rawling (2006) hold the case against constraints to be convincing and thus defend a Rossian account of deontology consisting of mostly *prima facie* duties. On their view, there are general prohibitions against killing and other acts of serious harm. But there are no constraints that would prohibit the agent from doing harm even to prevent more harm of the same type. And as we will see

in Chapter 5, consequentialists have explored several paths in order to try to accommodate constraintism.

That said, this thesis aims to justify constraints as a central feature of deontological or non-consequentialist ethics. Thus, where I do not specify otherwise, whenever I speak of constraintism I am referring to *deontological* constraintism.

1.4. A Puzzle About What?

I now turn to the question why constraintism might appear paradoxical. I will have more to say later about the status of the paradox of deontology, whether it is in fact a paradox, and about different versions of the problem. In this section, I ask in what way a commitment to the existence of deontic constraints might look problematic.

What the terminological considerations of the previous sections suggest is this. The puzzle about constraintism cannot simply be that deontic constraints do not require us to promote the best available outcomes. *Some* might puzzle over this. But it should be a puzzle only to consequentialists, not to deontologists. If there is a *deontological* paradox here, this paradox cannot be owed to the fact that there seems to be a contradiction between the following two statements:

(1) It is always better to promote outcomes that contain fewer killings.

(2) It is sometimes wrong to promote outcomes that contain fewer killings.

For unlike consequentialists, deontologists are not committed to the following statement about the nature of moral rightness:

(3) It is always right to promote better outcomes.

A consequentialist account of constraintism would seem paradoxical precisely because its *consequentialist* claim about moral rightness appears to contradict its *constraintist* claim about the wrongness of killing. Such an account would tell us that it is always right to promote better outcomes and, at the same time, tell us that it is sometimes wrong to promote outcomes which are presumably better than their available alternatives. As we will see in Chapter 5, the consequentialist's best bet might be to deny (1), arguing that outcomes that contain fewer killings are in fact outranked by their alternatives in paradox cases.

Deontologists, by definition, are not committed to (3). They may argue that the right and the good may come apart such that it is sometimes *not* right to do what is best. Hence, there is no deontological paradox if constraintism appears puzzling from the perspective of the good. Even critics of deontology must admit that the paradox *of deontology* cannot be owed to the fact that constraintism is incompatible with consequentialism's central claim about moral rightness.[22]

What then *is* puzzling about constraintism? As I shall argue, the best interpretation of the paradox of deontology—as a genuinely *deontological* paradox—identifies a conflict between two types of moral reasons against rights violations. Before turning to this interpretation in Section 1.5, I shall now discuss the original formulations of the paradox by Robert Nozick and Samuel Scheffler.

---

[22] Sophie-Grace Chappell has made a similar point. She argues that if "we can derive a contradiction from the premises 'It's good to keep deontological constraints' and 'The role of agency is to bring about goodness,' that does not, in and of itself, show that it's not good to keep deontological constraints. What it shows is that either it's good to keep deontological constraints, or it isn't the role of agency to bring about goodness" (Chappell 2011: 272). Her solution to the problem is to deny the second premise and to claim that constraints are instances of those times when the role of agency is *expressive* rather than *productive*. My—more charitable—interpretation of the paradox of deontology is that the problem she addresses here cannot be all that critics of constraintism mean when they talk about a *deontological* paradox.

1.4.1. The Concern-Focus Interpretation

The paradox of deontology in its familiar form is owed to a passage from Nozick's *Anarchy, State, and Utopia* (1974). In this passage, Nozick distinguishes between two ways to express a concern for individual rights. A concern for the non-violation of rights, he says, can either be conceived of as an end state to be achieved; or it can be placed as a constraint upon action. Nozick calls the first view *utilitarianism of rights*, the second *side-constraint view* (Nozick 1974: 28–30).

According to Nozick, the side-constraint view—I shall continue to say constraintism—stands in contrast to goal-directed views. Utilitarianism of rights is a goal-directed view as it tells us that if we really care about rights, we should see to it that no or as few rights violations as possible occur. In other words, if we care about rights, we should think of them as *non-constraining* rights—rights that it is permissible to violate even to minimise violations of the same rights overall. Constraintism tells us that we should think of rights as constraining rights. It tells us not to violate rights, even where this would help to achieve an end state that contains fewer rights violations overall.

Is constraintism a rational way to express a concern for rights? Nozick suspects that it might not be:

> Isn't it *irrational* to accept a side constraint *C*, rather than a view that directs minimizing the violations of *C*? […]. How can a concern for the nonviolation of *C* lead to refusal to violate *C* even when this would prevent other more extensive violations of *C*? What is the rationale for placing the nonviolation of rights as a side constraint upon action instead of including it solely as a goal of one's actions? (Nozick 1974: 30)

Nozick later reveals that, contrary to first appearances, it is *not* irrational to endorse constraintism. However, it is hard to tell from Nozick's questions what he thinks it is that makes constraintism even only *look* paradoxical in the first place.

Nozick's central idea seems to be this. If we care about the non-violation of rights, it seems that we should prefer that as few rights violations as possible occur. For this reason, constraintism looks paradoxical. I shall call this the *concern-focus interpretation* of the paradox of deontology:

> Concern-Focus Interpretation. Constraintism seems irrational because it endorses a certain kind of moral concern (about the non-violation of rights) but then requires the agent to act in a way that contradicts that concern.

The concern-focus interpretation is popular in the literature on constraints.[23] However, it rests upon the thought that a concern for the non-violation of rights should normally lead us to accept a goal-directed view that tells us to promote end states that contain fewer rights violations. But why should that be the case?

Lippert-Rasmussen (1999) has plausibly argued that this central idea presupposes a certain *kind* of concern, the kind of concern only a utilitarian of rights, not a constraintist, would have. As it were, a utilitarian of rights is concerned that no rights violations should *occur*. Perhaps, the rational way to express this concern in paradox cases is to see to it that as few violations as possible occur.[24] But the constraintist does not share this concern. As Lippert-Rasmussen rightly points out, someone who is concerned that no violations

---

[23] For instance, Frances Kamm asks: "Why is it not permissible, indeed obligatory, for us to [minimise rights violations] as an expression of respect or concern for rights?" (Kamm 1989: 252). And Heuer (2011) points out that favouring constraintism over utilitarianism of rights as a way of showing concern for rights is, without further argument, unjustified.

[24] We will see shortly that this is not as innocent an assumption as it might seem.

of rights occur, "no matter who performs them, has nonviolation of rights as an end state" (Lippert-Rasmussen 1999: 52). But Nozick defined the constraintist as someone who does *not* have the non-violation of rights as an end state. Instead, constraintists are concerned with the non-violation of rights in *some other sense* (further to be specified).

The answer to Nozick's (1974) key question—"How can a concern for the nonviolation of *C* lead to refusal to violate *C* even when this would prevent other more extensive violations of *C*?"—is then simply that the concern referred to in this passage is not the type of concern a constraintist would have. Nozick seems to assume that any concern for the non-violation of rights can be translated into a concern for their non-*occurrence*. But constraintists need not qua defender of constraints be concerned with the non-occurrence of rights violations at all. All that is clear at this stage is that constraintists think that no one should violate rights. The concern-focus interpretation of the paradox just falls short of showing anything more than that constraintism looks paradoxical *from the perspective of a goal-directed view*. But as I have argued earlier, this cannot be the grounds for there to be a *deontological* paradox surrounding constraintism.

1.4.2. The Goal-Focus Interpretation

Yet Samuel Scheffler has also felt the grip of Nozick's puzzle. In particular, Scheffler tries to explain why constraintists should care about the *non-occurrence* of rights violations:

> [Constraintists] need the idea that violations of [rights] are morally objectionable or undesirable, in the sense that there is a moral point of view from which it is preferable that no violations should occur than that any should. (Scheffler 1985: 415)

They need this idea, according to Scheffler, because they could not otherwise explain why people ought not to violate rights, for instance, in the pursuit of their self-interested ends. In general, constraintists want to say that I should not violate some right *R* even where this would prevent you from violating *R* more extensively. But if it is not in some sense undesirable that *R* is violated (by anyone), how can it be impermissible that I do so?[25]

Scheffler's second achievement in getting closer to the core of the puzzle is that he identifies the specific *form* of rationality behind the charge of irrationality against constraintist views. He calls this form of rationality *maximising rationality*:

> **Principle of Maximising Rationality (PMR).** [I]f one accepts the desirability of a certain goal being achieved, and if one has a choice between two options, one of which is certain to accomplish the goal better than the other, then it is, *ceteris paribus*, rational to choose the former over the latter. (Scheffler 1985: 414)

The *PMR* should be familiar from various contexts of action. If I want to gain the highest financial benefit from the items which I don't use anymore it is rational that I sell them to the highest bidder. If I need to get to work quickly it is rational to take the fastest route. And so on. Note that the *PMR* is not limited to cases where an agent must choose between two options. For every set of *n* options, it seems rational to choose the one option that is certain to accomplish the assumed goal better than any other option.

Under the assumption that *moral* rationality follows the same principle, it is rational that one does whatever helps to achieve one's assumed

---

[25] Strictly speaking, the question whether it is undesirable that *R* is violated seems to lead the constraintist into a dilemma (Lippert-Rasmussen 2009: 166). Either it is not undesirable that *R* is violated, in which case it is unclear why it would be impermissible for anyone to violate *R*; or it is undesirable that *R* is violated, in which case it is unclear why I shouldn't minimise violations of *R* since it should be more undesirable if *R* is violated more extensively.

moral goals better than any alternative. So construed, there is an apparent conflict between constraintism and the *PMR*:

> [Constraintist views] identify certain kinds of actions as morally objectionable or undesirable, in the sense that it is morally preferable that no such actions should occur than that any should, but then tell us that there are situations in which we must act in such a way that a greater rather than a lesser number of these actions are actually performed. (Scheffler 1985: 415)[26]

One reading of what Scheffler argues here is this. Scheffler thinks, contra Nozick, that constraintist views are goal-directed views. They *do* assign certain goals to agents. For instance, they assign agents with the goal that no rights violations occur. Call this goal *G*.

According to the *PMR* it might be rational—again, pending further investigation of this assumption—that an agent who has *G* as a goal chooses to perform a minimising violation as this would seem to accomplish *G* better than if she lets many more rights violations occur. Since constraintism prohibits this option, however, it makes an irrational request to that agent. I shall call this the *goal-focus interpretation* of the paradox:

---

[26] In this passage, Scheffler shifts the focus from the violation of rights (or constraints) to the acts that violate these rights. Some will find it counter-intuitive however to say that by violating the rights of five individuals an agent would perform five individual actions. In *Footbridge Paradox*, for instance, sending the trolley on its way to overrun five innocents seems to be one single act, even if it violates the rights of five individuals. Accordingly, some might find it counter-intuitive to say that you would minimise the total number of *objectionable actions* by killing the massive man. For this reason, paradox cases are sometimes constructed as cases in which one agent could prevent each of *n* other agents from committing one act of killing each (e.g., Hurley 2013a). It is only consequent that Scheffler himself seems to think of paradox cases in this way (Scheffler 1994: 84).

> Goal-Focus Interpretation. Constraintism seems irrational because it assigns a certain goal to the agent (that no rights violations occur) but then prohibits her from achieving this goal.

This is the first reading of Scheffler's puzzle. Lippert-Rasmussen, who reads Scheffler in this way, explains why the goal-focus interpretation does not bring us much closer to finding out what is wrong with constraintism: it seems that *G* is achieved *neither* by an agent who refuses to perform a minimising violation, *nor* by an agent who does so. If you violate someone's right, you obviously fail to ensure that no rights violations occur. If you refuse to violate that right, you fail to prevent a greater number of other rights violations and thus, obviously fail to ensure that no rights violations occur. It is impossible that *G* is achieved in paradox cases because such cases are defined *in terms of the impossibility* of ensuring that no rights violations occur. Thus, constraintists could simply "deny that in not performing a minimizing violation [the agent] achieves [*G*] less well, since whatever he does he does not achieve [*G*] at all" (Lippert-Rasmussen 1999: 53).

All that the agent *could* achieve in paradox cases is a quite different goal—the goal that as few rights violations as possible occur. Call this goal *G**. For constraintism to be irrational, then, two conditions must hold. First, constraintists must accept *G* and second, *G* must be better achieved in paradox cases if the agent achieves *G**. However, it is not obvious that constraintists are committed to either of these ideas.

Let's begin with the second idea: it is not obvious that if you have non-occurrence of *φ*-ings as a goal, you accomplish that goal better if there are fewer *φ*-ings. By way of illustration, suppose that you are playing the classic children's game *Operation*. You try to remove the Adam's Apple for £100. Your aim is to do so without touching the metal edge of the opening because if you do you lose the game. Evidently, once you've touched the metal edge and hear the buzzing sound it doesn't matter how many times you've touched it. Even if you touch it as few times as possible, you have failed to

accomplish your goal which was to remove the Adam's Apple without triggering the buzzer at all.

By analogy, *even if* constraints would think of the non-violation of rights as a goal to be achieved, they could insist that you fail to achieve this goal once you commit a minimising violation. The goal was that there are no rights violations, not that there are as few rights violations as possible. Thus, even if it constraintists would prefer that *no* rights violations occur than that any should, this would not seem to commit them to the claim that it is preferable that *fewer* violations occur *where this, in turn, would require the performance of a rights violation*. It is not irrational according to the *PMR* to choose not to perform a minimising violation simply because doing so would not accomplish the relevant goal (that no rights violations occur) any better than would a refusal to perform such violation.

One might think that there could be a way to resurrect the claim that once you have the non-occurrence of rights violations as a goal, you are also committed to make it your goal that fewer violations occur. However, the other, more fundamental problem with the goal-focus interpretation is that it is not even clear why constraintists would be committed the first kind of goal. Constraintism is not, as Nozick said, a goal-directed view. All the constraintist is committed to qua defender of constraints is the claim that it is wrong to violate rights even where this would minimise rights violations overall. She thinks *de facto* that we are usually required to achieve end states in which no rights violations occur. But she does not think so because she would think, *de jure*, that we ought to promote outcomes that contain no rightsviolations. On the best interpretation of constraintism, then, it seems that constraintists *do not assign any goals to agents* (Lippert-Rasmussen 1999: 54).

1.4.3. The Preference-Focus Interpretation

There is a second possible, though perhaps unpopular, reading of Scheffler's puzzle. For, strictly speaking, Scheffler's *PMR* applies not only to someone

who aims to achieve a certain goal but also to someone who merely *accepts the desirability* of that state of affairs obtaining.[27] Constraintism could seem irrational from the perspective of the *PMR*, even if constraintism does *not* assign any goals to agents, so long as it endorses the preference that no or as few rights violations as possible occur. Let me give an example to illustrate the difference.

Imagine that Chris wants to be a good father. He tries to show his children every day that he loves them. He tries to support them in any way possible, to educate them, and to put them under no unnecessary pressure. He really does his best. Chris also has a certain mental state about the future when his kids have grown up. He hopes that—as adults—his children will think that *he has been a good father*. At the very least, he would much prefer this to a future state in which they think that he has been a terrible father. Thus, Chris accepts the desirability of some future state obtaining—a state in which his children think of him as a good father. And he knows very well that his actions make it more or less likely that this state obtains.

Yet Chris finds it important *not* to assume an end state in which his children retrospectively approve of his parenting as a direct goal of his action. A good father, he believes, does what is best for his children *independently* of whether he will gain praise or recognition from them in the future. The point is that Chris does not have the retrospective approval of his children as a goal but *merely* accepts the *desirability* of that state obtaining. As it stands, however, this is enough to make Chris an appropriate addressee for Scheffler's *PMR*: he has the relevant preference. At any time, when Chris must choose

[27] To my knowledge, Scheffler is usually taken to be saying something about rational choice for an agent who acts upon a certain goal (the desirability of which to be achieved she of course accepts), not for an agent who accepts the desirability of a certain end state *without* making its achievement a direct goal of her action. To me, however, the latter agent is clearly within the range of addressees of the *PMR* as formulated by Scheffler.

between two options one of which is certain to achieve the relevant state of affairs better, the *PMR* holds that it is rational for him to choose that option.

By analogy, even if constraintists would *not* say that the goal that no rights violations occur should be a direct goal of our action, they too might be appropriate addressees for Scheffler's principle, so long as they accept that agent should prefer that no or as few rights violations as possible occur. Constraintists merely seem to think, as Scheffler says, that it is "preferable that no such [violations] should occur than that any should" (Scheffler 1985: 418). They could embrace the idea that it is desirable if no violations occur, without making this in any sense a goal of our action.

Thus, a second reading of Scheffler—in fact, one that is closer to Scheffler's own wording—suggests the following interpretation of the paradox:

> Preference-Focus Interpretation. Constraintism seems irrational because it endorses a certain kind of preference (that no rights violations occur) but then require agents to act in a way that contradicts that preference.

The preference-focus interpretation brings us closer to an appropriate understanding of constraintism since it does not allege that constraintism would assign any goals to agents. But it does not bring us closer to getting a grip on the puzzle about constraintism.

To see why, note that the case of an agent who accepts the desirability of a certain state of affairs obtaining without making its achievement a direct goal of her action might pose a serious challenge to Scheffler's *PMR*. Suppose that Chris is considering which snack to get for his son. He can choose between chocolate ice cream and rat poison. The *PMR* says that if Chris prefers that a certain state of affairs obtains, then ceteris paribus he should choose the option that is certain to achieve this state better than its alternatives. Chris accepts the desirability of an end state to be achieved in

which his son thinks that he has been a good father. Giving him chocolate ice cream instead of rat poison is certain to achieve this state better, as giving him rat poison will probably kill him. Thus, Chris should get him chocolate ice cream.

So far so good. But the *PMR* gets wrong *why* Chris should give his son ice cream instead of rat poison. For the reason is that Chris wants to be a good father—and good fathers don't poison their children—*not* that poisoning him will prevent his son from appreciating Chris's parenting in the future.

To put the point more generally: if the agent merely thinks that it would be preferable if a certain state of affairs obtains but does not, at the same time, make the achievement of that state a direct goal of her action, then the agent's choice of action is independent of that preference. Chris might make all kinds of choices in his efforts to be a good father that *might* contribute to the coming about of an end state in which his children eventually think of him as a good father. But he will have done *every single one* of these actions for some reason *other than* that he desires this end state.

This suggests that the *PMR*, in the form proposed by Scheffler, is simply false. If an agent has a certain goal *G*, it seems rational that she chooses the option which is certain to accomplish *G* better. But if she *merely prefers that G obtains*, this is not enough to make her choice of the option that achieves *G* better rational. For what explains why it is rational to choose one option over the other is the agent's reason *for favouring this option*. For instance, what makes it rational for Chris to protect his children from harm is not that he accepts the desirability of an end state in which they approve of his parenting, but that he believes a good father protects his children from harm.

It is not irrational, then, if constraintism requires the agent to act in a way that contradicts a certain moral preference simply because the preference that no violations occur is not what should guide her action and thus cannot be what makes her choice of action rational or irrational. Scheffler

seems to be aware of the limitations of the *PMR* when he says that it is "not obvious that maximizing rationality constitutes the whole of rationality" (Scheffler 1985: 418). Yet for the *PMR* to be plausible in the first place it would have to be reformulated to apply only to agents who *have* a certain goal (rather than merely accepting the desirability of its achievement). In this narrower version, however, the *PMR* does not apply to constraintism at all because constraintism is not a goal-directed view. Constraintism does not assign the goal to agents that fewer rights violations occur, and not even the goal that no violations occur.

1.4.4. The Value-Focus Interpretation

Let me mention one last interpretation of the paradox of deontology before turning to the one which I favour. That is, Thomas Nagel phrases the paradox in terms of constraintism's relation to agent-neutral values:

> The logical peculiarity of [constraining] rights can be described by saying that they cannot be given an interpretation in terms of agent-neutral values—not even in terms of the agent-neutral value of what they protect. (Nagel 2008: 106)

By *agent-neutral* values, Nagel means "values of certain occurrences or states of affairs, which give everyone a reason to promote or prevent them" (Nagel 2008: 105). I shall discuss the agent-relative/agent-neutral distinction in detail in the next chapter.

The story Nagel tells here is consistent with the stories by Nozick and Scheffler. Nozick's utilitarian of rights is governed in her moral judgements by agent-neutral values in Nagel's sense. She interprets the value of the non-violation of rights in term of states of affairs in which no or as few rights violations as possible occur. Because the occurrence of rights violations is bad, every agent has reason to promote these states of affairs, even if that

requires some of them to perform minimising violations. In contrast, constraintists place the non-violation of rights as a constraint upon (end-state-directed) action. They do *not* interpret the violation of rights in terms of their agent-neutral disvalue, which would then give everyone reason to promote outcomes that contain no or as few rights violations as possible occur.

In *The Rejection of Consequentialism*, Scheffler makes a similar point. In fact, he defines deontic constraints as:

> restrictions on action which have the effect of denying that there is any non-agent-relative principle for ranking overall states of affairs from best to worst such that it is always permissible to produce the best available state of affairs so characterized. (Scheffler 1994: 2–3)

Thus, focusing on agent-neutral values and their promotion, we get yet another interpretation of the paradox of deontology:

> Value-Focus Interpretation. Constraintism seems irrational because it endorses the idea that constraints protect something of agent-neutral value but then require the agent to act in a way that diminishes rather than furthers that value.

As with Nozick's and Scheffler's interpretations of the paradox, however, it is not obvious that there is any *deontological* paradox here. For it is not clear to what extent constraintist are committed to the idea that constraints protect something of agent-neutral value, especially if agent-neutral values are defined, as Nagel does it, as the values *of occurrences or states of affairs* which give everyone a reason to promote them. As we will see in Chapter 2, this is a rather narrow way of understanding agent-neutral values, and one that precludes any moral theory which gives priority to something other than the promotion of outcomes from the realm of agent-neutral moral theories.

1.5. A Deontological Paradox

It is not obvious from any of the formulations of the paradox discussed so far what precisely it is that makes constraintist views look paradoxical. To be clear, I am not yet arguing that constraintism *is not paradoxical*. Rather, my point here is that it is not even clear in what way constraintism *even only looks paradoxical*.

Taking stock of what we have learned so far: at the heart of constraintism is the claim that it is impermissible to commit minimising violations of rights and it seems as though this puts constraintism into conflict with a certain powerful form of rationality, i.e., *maximising rationality*. Roughly speaking, maximising rationality is the idea that if something is valuable, then it is rational by default to maximise the presence of that value and minimise interference with it.[28] Thus, if constraintism looks *paradoxical*, then this must be because it looks *irrational* according to the standard of maximising rationality.

But what exactly is the conflict between constraintism and maximising rationality? As we have seen, Scheffler's *PMR* will not help to fully analyse this conflict. Either the principle is meant to apply only to an agent who has a certain goal, in which case it does not apply to constraintist views that do not assign the relevant goals to agents; or it is implausible as a principle of rationality because it does not capture the connection between an agent's choice of action and the reasons which make that choice rational. Therefore, in order to understand the conflict between constraints and maximising rationality, the first thing we will need is an alternative to Scheffler's *PMR*.

---

[28] A similar formulation of the idea of maximising rationality is given by Kamm (1992): 359fn. Note that in this passage Kamm does not talks about constraints but about agent-centred permissions not to do harm; I will say more about agent-centred permissions in Chapter 2.

1.5.1. Maximising Rationality, Revisited

Let me begin with a simple example. Ron owns a cabin in the forest. For many years the cabin has served as his personal retreat, a place where he could spend a few days by himself. Now that Ron has children, however, he needs the money more than he needs the cabin. He decides to sell.

The demand is limited but after an appropriate waiting period, Ron has two credible offers by serious buyers, one amounting to £42,000 and another one amounting to £35,000. Ron cares about the financial benefit and thus has a reason to accept the highest offer. However, Ron also cares about other things. Most importantly, Ron would like to see the cabin in good hands. He would prefer to sell to someone who can appreciate the secluded location and who would continue to use the cabin, as Ron did, as a personal retreat. After speaking to both buyers, Ron learns that the first buyer who offers £42,000 is interested merely in the land. He's going to tear down the cabin to build a small luxury spa resort for couples. The second buyer who cannot offer more than £35,000 loves the cabin as it is and is looking for a place to spend a few quiet days now and then. What should Ron do now? Should he settle for the second offer, even though he could get more money if he accepted the first?

One natural kind of answer is this: Ron should do whatever he has *most reason* to do. Ron has a reason to sell for the highest financial benefit, and he has a reason to sell to someone who shares his ideology. What he *ought* to do is what, balancing these reasons, he has *most reason* to do.[29] Thus, it is natural to think that agents often have different reasons for and

[29] This is a very common idea; see e.g., Dancy (2004), Alvarez (2010): ch. 1, Alvarez (2018): §2, Bader (2016), Lord and Maguire (2016): 3–23, Portmore (2014), (2019): ch. 6. It should be noted, however, that some are sceptical whether there is a clear sense to talking about what one *has most reason to do* and about *balances of reason* (e.g., Gert 2016, Kearns 2016). I take it that for my purposes, this debate can confidently be left aside. Moreover, there are open questions about how to factor in reasons *against* an option and how they relate to reasons *for* options. I will ignore these questions as well; see e.g., Snedegar (2017): ch. 1.

against acting in certain ways, which each contribute to what they have overall or most reason to do. They are *contributory* reasons; to talk about what one has most reason to do is just to "talk about where the contributory reasons come down" (Dancy 2004: 15–16).[30]

For instance, to say that Ron has most reason to accept this or that offer is to say that his contributory reasons—the balance of his reasons—come down on this or that side. Perhaps, his financial situation is precarious enough to urge him to sell for an extra £7,000, even to someone who does not share Ron's ideological convictions. Or maybe it is so important to him that the cabin is in good hands—and his financial situation allows it—that he rather sells for £7,000 less than seeing the cabin being teared down to make room for a spa resort. Either way, what Ron ought to do is a function of a balance of his (contributory) reasons.

Taking these considerations into account, I think all we need to make the conflict between constraintism and maximising rationality explicit is the following kind of idea:

> **Principle of Balanced Reasons (PBR).** If an agent has reason to φ and reason not to φ, then, other things being equal, she ought to do what balancing these reasons she has most reason to do.

Like Scheffler's *PMR*, the *PBR* is not limited to cases where the agent must choose between two courses of action. Ron could have a wider range of available offers to choose from—more reasons to balance. But the idea remains the same: what he ought to do is what, balancing these reasons, he has most reason to do.

However, unlike the *PMR*, the *PBR* captures the connection between an agent's choice of action and the reasons which contribute to it. Whereas

---

[30] Note that there are two notions of *reason* here, one as a mass noun (reason to do something) and another one as a count noun (*a* reason to do something). I will place no further importance on this distinction.

the *PMR* allows for the agent's choice of action to be rational even where the agent chooses not to act upon the fact that she desires the achievement of some goal, the *PBR* inquires the normative reasons the agent has for favouring this or that course of action.[31]

1.5.2. The Reason-Focus Interpretation

Now back to constraintism. For the sake of argument, suppose that I have reasons not to violate some right *R* in anyone and that I also have reasons to prevent violations of *R*. Suppose further that I can only either refuse to violate *R* or violate *R* to prevent *more* further violations of *R*.

On the one hand, constraintists claim that in this case I ought not to violate *R*. On the other hand, the *PBR* says that I ought to do whatever I have most reason to do. For constraintism to be rational it must therefore be true that I have most reason not to violate *R*. But how can this be? If I have both reasons not to violate *R and* reasons to prevent violations of *R*, shouldn't I have *more* reason to prevent the greater number of violations of *R*? How can I have most reason not to commit a minimising violation of *R*, if both kinds of reasons contribute to the balance of my reasons?

I think that we should understand the conflict between constraintism and maximising rationality in this way, i.e., in terms of a conflict between *two*

---

[31] Without going into debates about reasons and their kinds which could not be accurately addressed in this thesis, I generally assume that the *PBR* generalises over reasons that everyone—or almost everyone—accepts as *normative*, i.e., reasons for someone to favour doing this or that. Since my primary concern is with finding an acceptable formulation of the paradox of deontology, the *PBR* may in the context of this thesis be read as being restricted to *moral* reasons. Thus, what one ought to do morally is what one has most moral reason to do. This leaves room for the idea that what one ought to do morally may not be what one ought to do *all-things-considered*. There may well be additional—for instance, legal or prudential—considerations such that I have most moral reason to *φ*; and yet I ought to not to *φ*, *all-things-considered*, because *φ*-ing is illegal or imprudent (and these latter considerations are weightier than or take priority over the moral considerations).

kinds of reasons against rights violations.[32] Constraintism looks irrational because it prohibits the agent from doing what she seems to have most reason to do in paradox cases. I call this the *reason-focus interpretation* of the paradox:

> Reason-Focus Interpretation. Constraintism seems irrational because it endorses the idea that the agent has both reasons not to commit rights violations and reasons to prevent them but then prohibits her from doing what she seems to have most reason to do.

I think the reason-focus interpretation brings us closer to getting a grip on the puzzle, although it cannot be the final word. It depends on the truth of the statement that constraintism endorses both reasons not to commit rights violations and reasons to prevent them.

Constraintists certainly accept that we have reasons not to violate rights. But it is not obvious that they must accept that we also have reasons to prevent rights violations, let alone *pro tanto* reasons—standing reasons with genuine weight[33]—to prevent violations *where this would require the agent to violate rights herself*. In the next section, I will investigate the question further to what extent constraintists are committed to thinking that we have pro tanto reasons to prevent rights violations.

That said, I believe that resting the conflict between constraintism and maximising rationality on the *PBR* rather than Scheffler's *PMR* is a significant step forward. Whereas it proves difficult to ascribe to constraintists beliefs

---

[32] It might be tempting to phrase this conflict of reasons as a conflict between *agent-relative* reasons not to commit rights violations on the one, and *agent-neutral* reasons to prevent rights violations on the other hand. I think it is important not to give in to this temptation; one of the main objectives of this thesis is to argue that reasons of the first kind (reasons not to commit rights violations) can be conceptualised as robustly agent-neutral on constraintist views.

[33] In contrast to a *prima facie* reason which *appears to be* a reason for something but might turn out to have no weight or be no reason at all, a *pro tanto* reason stands and may only be outweighed or trumped by other reasons.

about the existence or relevance of certain moral goals (at least if goals are understood as states of affairs to be achieved), it should be much easier to do this with beliefs about the existence of reasons. The *PBR* does not depend on a teleological, goal-directed approach to ethics, and can be used to analyse *any* moral theory that accepts reasons as a normative category. Generally, any moral theory must make requests that are consistent with the thought that we should do whatever we have most reason to do morally. And what we have most reason to do morally depends on what is fundamentally prior in ethics—the promotion of good outcomes, the treatment of persons as ends-in-themselves, the principles rational self-interested agents would agree upon, and so on.

Thus, on the reason-focus interpretation, the paradox of deontology exists because constraintists seem committed to accepting the truth of each of the following statements:

(1) I have reasons not to violate *R.*
(2) I have reasons to prevent violations of *R.*
(3) I can only either refuse to violate *R* or prevent a greater number of violations of *R*. (Introduction of paradox cases)
(4) I ought to do what I have most reason to do. (Introduction of the *PBR*)
(5) Balancing my reasons against rights violations, I have most reason to prevent the greater number of violations of *R*.
(6) Thus, I ought to prevent the greater number of violations of *R.*
(7) I ought not to prevent the greater number of violations of *R*. (Introduction of a deontic constraint)

The difficulty of stating a problem like this as a paradox is, of course, that paradoxes exist because of certain *appearances*—appearances that call for an explanation rather than for being left untouched (Pleitz 2018: 12–15). Many constraintists will have immediate notions about why one or another of these statements are false, or why she is not in fact committed to them.

The primary difficulty of addressing an informal paradox properly—one that is not a strictly logical paradox—is not actually how to solve it but to allow for it to be stated in some form such that, subsequently, one can think about its solution. I therefore ask the reader to be patient with me when I explain why I think this is an appropriate way to state the paradox of deontology.

To begin with, what makes the above complex of statements a paradox is that there appears to be a contradiction between the statements (6) and (7). The assumption that I ought to prevent the greater number of rights violations directly contradicts the assumption that there is a deontic constraint on the minimising of rights violations overall.

The next question is: are constraintists committed to both (6) and (7)? It should be clear that they are committed to (7) since the existence of constraints is what constraintism is all about. Thus, the statement that constraintists will be sceptical about is (6). The truth of (6) depends on the statements (1)—(5), so what about the truths of those statements?

Statements (1), (3), and (4) should be unproblematic. Constraintists evidently think that we have reasons not to commit rights violations, so (1) should stand. Statement (3) merely states the possibility of paradox cases in which an agent could only prevent several rights violations by committing a single rights violation herself. I have stated examples of such cases earlier and thus, we can imagine such cases to exist. And statement (4) states the *PBR*, which I have motivated earlier.

Thus, constraintists are most likely to want to reject (6) via rejection of (5), i.e., the claim that in paradox cases I have most reason to prevent the greater number of violations of *R*. If constraintists aim to reject (5), then they will most likely want to reject (2). If I don't have reasons to prevent rights violations, only reasons *not to commit them*, then I could have most reason

not to commit a rights violation even where this would mean that I fail to prevent a greater number of other violations.

In the next section, I will therefore focus on the question whether constraintists are committed to the idea that there are pro tanto reasons to prevent rights violations. In general, this thesis is about a way in which constraintists might be able to resist (5), even if they accept the truth of (2). I will argue that even if we have reasons to prevent rights violations, we have most reason in paradox cases not to commit a minimising violation.

Before turning to the question about reasons to prevent rights violations, let me end this section with some remarks on the status of the problem stated above. Some authors have denied that the problem is a real paradox (e.g., Chappell 2011, Heuer 2011). It might not be a paradox in the strict logical sense of the term. That is, there is a difference between the paradox of deontology and, say, the logical paradox apparent in the statement: 'This sentence is false.'

However, I think the above formulation suggests that the problem *can* be given the form of a paradox, i.e., the form of an argument "that appears to be valid from premises that appear to be true to a conclusion that appears to be unacceptable" (Pleitz 2018: 12). There appears to be a contradiction between two statements, (6) and (7), that appear to be true because they appear to follow from premises that each appear to be true.[34] That said, the question whether the paradox of deontology deserves to be called a paradox is of minor importance. Even if it did, it would be a mistake to think of it as a trap from which the constraintist has no way out. Paradoxes call for an

[34] Pleitz (2018): 16 also emphasises the *democratic* element to charactering philosophical problems as paradoxes. Some philosophical problems that have long been treated as paradoxes—think of *Achilles and the Turtle*—are not treated as paradoxes anymore since the puzzling appearances that gave rise to them have plausibly been explained. Whether a paradox has in this way been resolved depends at least in part on how convincing the philosophical community takes the proposed explanations to be.

explanation of why at least some of the appearances creating the paradox are in fact illusory.

One way in which the paradox of deontology *is* clearly misnamed, however, is that it concerns constraintism rather than deontology. And so far, we have no reason to believe that the paradox *of deontology* is not a problem for non-deontological versions of constraintism, even though it seems that the problem should call for a different kind of solution on consequentialist accounts of deontic constraints. (More on this in Chapter 5.) But again, the main concern of this thesis lies with the question how deontology or non-consequentialism could be defended against the charge of paradox. Where I do not say otherwise (or speak of *consequentialist* or *consequentialised* constraintism), I use the term constraintism as shorthand for *deontological* constraintism.

1.6. No Obvious Way Out

The crucial question on which the existence of a deontological paradox of constraintism depends is to what extent constraintists are committed to the claim that there are reasons to prevent rights violations. That means, the easy way to avoid the problem would be to deny the existence of such reasons altogether.

Before exploring this path, I need to contextualise the way in which I intend to classify different ways to approach the paradox. I intend to distinguish the *agent-centred approach* from the *agent-neutral approach*. Roughly, the agent-centred approach argues that requirements not to violate rights are relative to the agent acting such that she has most reason not to violate some right *R herself*, even if her violation would prevent more extensive violations of *R* by others. In contrast, the agent-neutral approach aims to justify

constraints without reference to the idea that constraints are, in any substantive sense, agent-relative.

However, it is common to use a different two-way distinction to classify the various approaches to the problem: it is common to distinguish between *agent-based* and *patient-based* accounts.[35] Roughly, *agent-based* accounts aim to find the rationale for constraints on the side of the agent, in some feature or features of moral *agency*, whereas patient-based accounts aim to find it on the side of the victim of a minimising violation, in some feature of features of moral *patiency*.[36]

Examples for agent-based accounts are those which appeal to the notion of moral integrity, arguing that constraints allow agents to be good in a world that is not (Williams 1973, Fried 1978, Chappell 2007). Others aim to explain constraints by reference to the nature of moral evil as something that may not guide our actions, even as a means to some good end (Nagel 1986: 181–182), or by disconnecting moral wrongdoing from outcomes such that any agent who performs a minimising violation would in fact *maximise* wrongdoing (Brook 1991); I will say more about Brook's view later. All these accounts tell us that there is something about being a moral agent that justifies the existence of a deontic constraint on one's action.

Examples for patient-based accounts might be less diverse. In general, patient-based accounts identify the holders of constraining rights as being worth of certain forms of moral protection such that there are treatments which are precluded from the set of permissible ways of treating them. Thus, any patient-based account eventually aims to ground constraints in the *moral*

---

[35] Kagan (1989): 27–32 employs a three-way distinction between agent-, patient-, and relationship-based approaches. However, those approaches which Kagan calls *relationship-based* seem to me to always either ground constraints in some feature or features of persons qua agents or qua patients and are, in this sense, eventually either agent- or patient-based.

[36] I understand moral agency and moral patiency as the two moral faces of personhood. Qua agents, persons *act* in morally relevant ways. Qua patients, they *are acted upon* in morally relevant ways.

standing of persons (e.g., Kamm 1989, Brook 1991, Nagel 2008). In Section 1.6.4, I shall say more about why I take my distinction between the agent-centred and the agent-neutral approach to be more useful than the conventional agent-based/patient-based distinction. For now, it should just be noted that these are two different ways of classifying the various ways to approach the paradox.

1.6.1. The Agent-Centred Approach

As already noted, the easy way to avoid the paradox of deontology would be to simply deny that we have reasons to prevent others from violating rights. In this regard, Ulrike Heuer has argued that there is "no reason for violating a person's right […] when doing so would prevent a greater number of rights violations" simply because there "is no reason to prevent rights violations *per se*. There is only a reason not to commit them" (Heuer 2011: 261).

According to Heuer, the rejection of reasons to prevent rights violations committed by others follows from the idea that the moral worth of an action depends on whether the agent acts *for the right kind of reasons*. Since it is not within the reach of *my* agency to make others act for the right kind of reasons, she argues, it cannot be the direct goal of my agency to make it happen that they do (Heuer 2011: 251). But if there are no reasons to prevent rights violations per se, then it seems that all I am left with are reasons not to *commit* them. I will thus have most reason not to commit a rights violation even if my violation would minimise the total number of comparable violations by everyone.

This seems to bypass the paradox. As we have seen, the problem arises just in case that I have *both* reasons not to violate *R* and reasons to prevent violations of *R*. Denying the existence of the second kind of reasons avoids the relevant conflict of reasons.

Heuer's view is a version of what I shall call the *agent-centred approach* to the paradox which argues that reasons not to violate rights are

relative to the agent acting such that she has most reason not to violate some right *R*, even if her violation would prevent more extensive violations of *R* by others. Roughly, an *agent-relative* reason is a reason only for the particular agent, as opposed to an *agent-neutral* reason which is a reason for everyone.[37] Others have favoured the agent-centred approach. In particular, McNaughton and Rawling (2006) argue that the idea that requirements not to violate rights are agent-relative in this sense is sufficient to justify constraints.

Note that the agent-centred approach is one that proposes an agent-centred or agent-relative *justification* of constraints. It is very common to think that constraints take an agent-relative *form*, as restrictions on the conduct of the particular agent. But according to the agent-centred approach, constraints are not merely agent-relative in form but agent-relative in a substantive sense. The agent-centred approach says that there is a deontic constraints on *φ*-ing *because* the agent's primary concern should lie with her own acts of *φ*-ing. Chapter 2 will provide a detailed analysis of the distinction between formal and substantive agent-relativity.

Is the agent-centred approach sufficient to avoid the paradox? The most popular objection against the agent-centred approach is that it could not accommodate *intrapersonal* constraints. That is, it cannot explain why I ought not to violate *R* even to prevent *myself* from violating *R* more extensively. What if I could—like in Kamm's *Bomb Paradox* case—prevent *myself* from killing five by killing a single sixth person? Even if I have no reason to prevent *others* from violating *R* because my reasons not to violate *R* are

[37] This is a very rough definition. But it shall be sufficient for the purposes of the present chapter. Chapter 2 will provide a detailed analysis of ways how we are to understand agent-relativity.

personal or agent-relative reasons, shouldn't I have reasons to prevent *myself* from doing so?

Heuer bites the bullet here and denies the existence of intrapersonal constraints altogether. In a case where I could kill one to prevent myself from killing five, she argues, I am essentially presented with the choice, "killing either one or five, and in that case we may be allowed to aggregate" (Heuer 2011: 261). On this view, minimising the number of my own rights violations is precisely what I should do in the relevant type of case.[38]

However, this is a minority position.[39] A more common response to the challenge of intrapersonal constraints frequently put forward by proponents of the agent-centred approach is this: constraints give each agent a special concern not only with her own choice of action, but with her own *present* choice of action. In addition to being agent-relative, the argument goes, constraints are *moment-relative* in that they require me to ensure that *I* do not perform certain kinds of actions *now*. If my reasons not to violate *R* are relative to the present moment, then I have most reason not to violate *R now*, even if I could thereby prevent myself from violating *R* more extensively *at other times* (e.g., Brook 1991, Broome 1991, Johnson 2019). Some have even argued that moment-relativity renders agent-relativity obsolete and that all we would need to make sense of constraints is the claim that the *present*

---

[38] It should be mentioned here that Heuer's account—if plausible—would still avoid a portion of the paradox as intrapersonal constraints make up only a proper subset of constraints. Heuer's account might still explain why there are *interpersonal* constraints.

[39] What is meant here is the view that we should kill to prevent ourselves from committing more killings in the future. Many believe that we should minimise the number of killings we commit if we must choose between, say, steering a trolley towards one innocent person or towards five. Heuer's view—and the minority position referred to here—is that this case is equivalent to the one in which we set up a bomb to kill five and could defuse it by using the body of a sixth person.

time-slice of the agent must not do certain acts even to prevent the outcomes of past or future acts (done by other agents or her past or future self).[40]

Is the agent-centred approach—extended with a moment-relative requirement to focus on one's actions *in the present moment*—sufficient to avoid the paradox? I do not want to argue that this is not a plausible extension to the agent-centred approach. Here, my quarrel with this kind of approach is a different one: the agent-centred approach—even if plausible—avoids only *one part* of the paradox of deontology which I shall call the *rationality paradox*. It cannot, I argue, avoid its other part which I call the *value paradox*.

1.6.2. The Rationality Paradox

Let me begin with raising a critical question about the agent-centred approach: Do the agent-relative reasons of others give *me* reasons to do anything? Heuer's answer is *No*. She argues that our reasons against rights violations are not just relative to the agent acting. They are *personal* reasons. A *personal* reason, according to Heuer, is an agent-relative reason whose "presence does not give any reasons to others" (Heuer 2011: 254). Thus, from the mere fact that *you* have a personal reason to *φ* does not follow anything *for me*. In particular, it does not follow that I have a reason to ensure your *φ*-ing.

This exposes Heuer's view as a radical variant of the agent-centred approach. McNaughton and Rawling, for instance, do *not* deny that the agent-relative reasons of others may give us reasons to act in certain ways.

[40] Arguments that go into this direction have been put forward, e.g., by Broome (1991): pp. 9–10 and Brook (1991). I will not discuss these views in any further detail here. But I take it that it is not obvious that agent- and moment-relativity can be detached from one another in the relevant way because a moment-relative constraint would require the present time-slice *of the particular agent* not to perform a minimising violation.

Many others allow for this feature of agent-relative reasons (e.g., Nagel 1970, Parfit 1979, Sen 1982). Why does Heuer take the more radical position?

Heuer argues, contra McNaughton and Rawling, that agent-relative reasons alone—if such reasons are understood as giving reasons to others—are insufficient for establishing deontic constraints (Heuer 2011: 254 fn36). This is so because if *your* agent-relative reasons not to violate rights give me reasons to ensure that you do not violate rights, then we are back in the grip of the paradox: it seems that where I could violate *R* once and prevent more extensive violations of *R* committed by you, I should have *more* reason to make you comply with your agent-relative reasons not to violate *R* than not to violate *R* myself. Heuer thinks that only once we deny that I have *any* reason to prevent you from violating *R per se*, we can plausibly justify the existence of a constraint on minimising the total number of violations of *R*.

I do not think that advocates of the agent-centred approach have to take this radical step—although, as we will see shortly, it might not be that radical after all. The idea that reasons against rights violations are agent-relative alone might not explain why I should not perform a minimising violation, if your agent-relative reasons, too, give me reasons to prevent you from violating rights. But proponents of the agent-centred approach could invoke the idea that constraintism tells us which kinds of normative considerations *should take priority* over which other kinds of considerations. They could claim that if reasons against rights violations are agent-relative, all this means is that my reasons not to violate rights myself take priority over my reasons to prevent violations by others (or myself at other times than right now), and even a greater number of such violations. Preventing someone else from violating rights might be an important moral business; but it is more important (in some sense further to be specified) that the agent should violate rights herself.

One might find these more or less appealing claims for a moral theory to make. But they reflect an idea central to common-sense morality, i.e., the

idea that, as Williams said, “each of us is specially responsible for what he does, rather than for what other people do” (Williams 1973: 99). Normally, I have a certain direct influence on what *I* do *right now*, which I lack with regard to the actions of others or my future or past self. Why should we not think that this is a morally relevant fact? Why should we not think that this direct influence makes me, in some sense, *more* responsible for what I do in the present moment?

This way, I think the agent-centred approach could avoid one part of the paradox of deontology which I call the *rationality paradox*. The rationality paradox says that if I have both reasons not to violate *R* and reasons to prevent violations of *R*, then it seems irrational by default to prohibit me from performing a minimising violation of *R* since *whatever it is* that makes my violation of *R* worth avoiding will seem to make the greater number of violations of *R* more worth avoiding. The agent-centred approach provides an answer to *this* part of the paradox: reasons against rights violations are relative to the agent acting. Thus, when avoiding rights violations, each agent should give priority to those violations *she* would commit in the present moment. Minimising the number of rights violations committed in total should just not be her immediate concern.

1.6.3. The Value Paradox

Thus, the agent-centred approach has the conceptual resources to explain why minimising the total number of rights violations should not be the agent’s primary moral concern. That is, the agent-centred approach can avoid the rationality paradox.

However, I think the agent-centred approach faces a particularly severe version of another problem which I shall call the *value paradox*: why should we think that the agent’s primary concern should lie with her own actions? Paradox cases involve the expectation of severe harm to individuals who matter morally in their own right. In the face of the harm that awaits the

many of them, how can it be morally appropriate to think that it is a significant fact who would cause that harm? How can *my* agency be so significant, if compared to the significance of what would happen to *your* victims? The agent-centred approach, it seems, simply fails to account for a central feature of moral values as something that is valuable *beyond the agent's limited perspective*.

Heuer too acknowledges that there is a problem here. For even if we do not have reasons to prevent rights violations *per se* (because such reasons are entirely personal), this does not mean that we do not have reasons *whatsoever* to prevent what would happen to the victims of rights violations. Given the significance of what would be done *to them*, I think we must acknowledge that there are moral reasons—and not particularly weak ones—to protect them. A reason to prevent a rights violation is never just a reason to prevent that rights violation *per se* but a reason to prevent the mistreatment of an individual.

However, once we acknowledge the existence of such reasons, the agent-centred approach is drawn back into the grip of the rationality paradox. For how can it be that I have most reason not to violate someone's right even to prevent more comparable violations, if I also have reasons to prevent each of these other violations?

In fact, even Heuer agrees that the fact that reasons to prevent rights violations *per se* do not exist would not give us a conclusive answer what we should do in paradox cases because we have other reasons for preventing what would happen to the victims of rights violations (Heuer 2011: 265).[41] But Heuer wants to vindicate the agent-centred approach by claiming that

[41] This move is a necessary one, of course, because Heuer's view needs to account for cases in which the agent could easily and *without* violating anyone's rights herself prevent a rights violation. Where she could, for instance, prevent Jones from throwing a child into a shallow pond—she's must stronger than Jones and could easily hold him back—the agent should have reason to do so. On Heuer's view, this reason cannot be a reason to prevent Jones' rights violation *per se* (those do not exist). Thus, there must be some other kind of reason to prevent Jones' rights violation.

this problem would lie not "with the very idea of [deontic constraints], but with puzzling aspects of the ethics of killing" (Heuer 2011: 238).

It is not obvious to me how Heuer intends to single out the case of killing. What goes for killing should, I think, also apply to cases of torture, enslavement, abuse, rape, and other kinds or serious harm.[42] Phrasing the reasons against rights violations as intimately personal such that there are no reasons to prevent rights violations by others *per se* does not fully avoid the paradox of deontology because constraintists seem committed to existence of pro tanto reasons to prevent rights violations, *more broadly understood*.[43] It is hard to see how constraintists could plausibly deny that we have *any* pro tanto reasons to prevent what would happen to the victims of rights violations. The commitment to the claim that such reasons exist seems itself to be a *moral* one.

The value paradox is not an entirely new problem. Others have raised similar worries about the agent-centred approach. For instance, Kagan argues that where "I have some control over whether others shall do harm, the [constraintist] still needs to explain why the harm-doing of others should not be of as much concern to me as the harm I may bring about myself" (Kagan 1989: 126). And Alexander and Moore wonder how a "secular, objective morality can allow each person's agency to be so uniquely crucial to that person" and call this a "moral" paradox (Alexander and Moore 2020: §4). However, I want to offer a new understanding of this problem, not as an objection against the agent-centred approach, but as an integral *part of* the paradox of deontology itself, a part which any response to the paradox must address. I shall say more about this understanding of the value paradox in the concluding Section 1.7.

---

[42] Heuer is right, however, in making a difference between the case of killing and the case of promise breaking. I will account for this difference in Chapter 2 when discussing the agent-neutral or agent-relative character of general and special constraints.

[43] Kagan (1989): 47–50 argues that constraintists are committed to the existence of pro tanto reasons *to promote the good*. I think for the paradox of deontology to arise all we need to say is that they should accept the existence of a subset of such reasons, i.e., pro tanto reasons to prevent the harms associated with rights violations.

1.6.4. The Agent-Neutral Approach

Any account that we could hope can address the above version of the value paradox is one that does *not* hold my own agency to be so significant that I should have a special concern with my own actions. In other words, I believe a convincing answer to the paradox of deontology—one that can address both the rationality paradox and the value paradox—must take what I call the *agent-neutral approach*.

To solve the paradox, this thesis sets off from Frances Kamm's *inviolability account* as a version of the agent-neutral approach. Kamm has argued that individuals who are protected by constraining rights are "more potent individuals than they would be otherwise" (Kamm 1989: 254). A moral system that endorses the existence of this type of rights gives expression to the person as a *more inviolable*, and hence more valuable type of being, if compared to a system that denies the existence of constraints and with it the relevant concept of the person. If there were no constraints, we might save more lives—but we would have to live our lives believing "in a less sublime and elevated conception of ourselves" (Kamm 1989: 254).

If it can be given a plausible explication, the inviolability account provides an answer to the rationality paradox. That is, it is not irrational to prohibit the agent from minimising the total number of rights violations once we understand that the value of having those rights lies in the moral status which they give expression to, a status that would be denied if it were permissible to perform minimising violations. Chapter 3 will develop this part of the solution in detail. As we will see, the inviolability account does not rest on any understanding of reasons against rights violations as substantively relative to

the agent acting.[44] Thus, it does not face the version of the value paradox that troubles the agent-centred approach.

Yet the inviolability account, too, faces *some* version of the value paradox. Why should we think that our inviolability is so significant that we should not prevent the harms we could prevent, especially if these threaten a greater number of people? How can it be morally appropriate be so concerned about our inviolability status, where this would mean that many more of us will *actually* be violated? Just like with the agent-centred approach, the value paradox draws the agent-neutral approach back into the grip of the rationality paradox: why should we not focus instead on actual violations and minimise the number of actual violations in everyone? In fact, I think the inviolability account faces two specific variants of the value paradox, one that takes the form of an *external* criticism and one that takes the form of an *internal* criticism.[45]

The *external* variant says that we cannot take for granted, without further argument, that we should care more about our moral status than about the things that may happen to us. Perhaps, a world without constraints is one in which we must believe, as Kamm says, "in a less sublime and elevated conception of ourselves" (Kamm 1989: 254). But, at the same time, it is a world in which fewer of us end up being violated (provided, of course, that the permissibility to save the greater number in paradox cases actually increases the number of cases in which agents do so). It doesn't seem unreasonable to prefer such a world to one with constraints on the grounds that

---

[44] As already noted, it is one thing to think of constraints as agent-relative *in form* and another thing to think that they give the agent a special concern with her own actions. A proponent of the inviolability account may understand constraints as agent-relative in form without thinking that we need to refer to agent-relativity to explain their normative force. More on this in Chapter 2.

[45] I understand an *external* criticism as one that criticises an argument based on one or more premises which are not themselves part of the original argument. In contrast, an *internal* criticism attacks an argument based on one or more premises which *are* part of the original argument, aiming to show that the argument is inconsistent or insufficient. More on this in Chapters 3 and 4.

the first world would contain fewer killings, tortures, enslavements, less abuse, deception, and promise breaking in total. In Chapter 3, I will aim to address this external variant of the value paradox.[46]

The *internal* variant of the value paradox says that we cannot take for granted, without further argument, that we should care more about our inviolability than about other dimensions of our moral worth. In more detail: if there are constraints, then there are more things that others may not do to us even in the service of good ends. At the same time, however, there are more things that others may *allow to be done* to us in the service of respecting the inviolability status of a few. Others may allow for our rights to be violated where only a minimising violation could save us. How, then, are we beings of higher worth *overall* if there are constraints? As far as the things that others may allow to be done to us are concerned, we seem to be *less* significant beings if there are deontic constraints. This is often referred to as the *saveability challenge* to the inviolability account.[47] Chapter 4 will present a novel response to this challenge.

Before concluding the considerations of this chapter, let me say why I think that distinguishing between the agent-centred and the agent-neutral approach is preferable to the common practice of separating agent-based from patient-based accounts.

According to my interpretation of the paradox of deontology, any account to justify constraints must aim to explain the agent's moral reasons against minimising violations. Perhaps, this is also what Stephan Darwall has

---

[46] There might be some expectation management in order here. I will not present a complete argument that shows that we should care more about our moral standing than about the morally significant things that might happen to us. I will merely argue that it is not irrational to think that there is a plausible perspective in ethics from which our moral standing is fundamentally prior to the things that might happen to us. Whether one chooses to accept a moral theory that gives priority to our moral worth or to the minimising of morally objectionable harm ultimately depends on one's convictions about what the right kind of moral theory should look like.

[47] The objection has first been raised by Kagan (1991). Lippert-Rasmussen (2009) and Otsuka (2011) have challenged the inviolability account on similar grounds.

in mind when he states that "no justification for agent-centred restrictions can be found so long as we begin by looking *outside* the moral agent" (Darwall 1986: 291). Admittedly, Darwall might just think so because he thinks of constraints as *agent-centred* restrictions. But in this case, his claim is trivial: how could a restriction that centres on the agent acting (in some substantive sense) be justified without reference to that agent? On a more charitable interpretation, Darwall points us to the fact that even a patient-based response to the paradox must eventually explain why the agent should have most reason not to commit a rights violation, even where she also has reasons to prevent other more extensive violations.

Moreover, the use of the agent-based/patient-based distinction might suggest that all agent-based accounts take the agent-centred approach. This would be a misunderstanding. For instance, Brook (1991) points out that it should come natural to constraintists to follow an ethical tradition that holds intentions (or something related to intentions) to be the bearers of right- and wrong-making features. This would disconnect moral wrongdoing from outcomes in a way that renders the very idea of *minimising* violations of rights unintelligible. If all prior attempts to violate people's rights constitute wrongs that cannot be undone by preventing the bad consequences from occurring, any so-called *minimising* violation of a *R* would just constitute an additional wrong and therefore, *maximise* violations of *R* overall.[48] The account Brook has in mind is agent-based insofar as it sets off from what it means for an agent to do wrong, but it is not based on any conception of the agent's moral reasons as personal or agent-relative. The interesting question, I think, is whether we approach the paradox from the perspective of an *agent-centred* or an *agent-neutral* conception of morality.

[48] Setiya (2018) puts forward a similar argument but argues in favour of an agent-neutral *consequentialist* account of constraintism. I will come back to Setiya's proposal in detail in Chapter 5.

1.7. Conclusion

The considerations of this chapter suggest that the paradox of deontology might be best understood not as a single problem, but as a two-staged constraints-sceptical argument. The argument begins with the rather informal thought that it seems *worse* if there are more rights violations overall and that it is unclear how a plausible moral theory can require us to make the world a worse place. Constraintists can try to address this thought by identifying some feature *F* of minimising violations that makes it wrong for the agent to commit them. But then, the constraints-sceptic can point out that since the act the agent would commit and the acts the agent could prevent are described as equally significant in paradox cases, *whatever F is taken to be*, the acts the agent could prevent should also manifest *F*. If *F* is the relevant feature that makes an act morally objectionable, how can it be prohibited to minimise manifestations of *F* in everyone's action? This is what I have called the *rationality paradox*.

To answer the rationality paradox, the constraintist must explain why minimising the manifestations of *F* in everyone's action cannot be—or shouldn't be—the agent's primary concern. As we have seen, the constraintist can refer to agent-relativity, for instance, and identify *F* as a feature of acts which the agent, as opposed to others, would perform in the present moment. Or she could identify *F* as the feature of acts which come up against a certain kind of moral status.

Either way, this makes the constraintist vulnerable to the value paradox. For the constraints-sceptic can now question the constraintist value-assumption: why should it be more important to avoid actions that are *F* than minimise manifestations of some other candidate wrong-making feature *G* in everyone's action—whereby *G* might, for instance, be defined in terms of the harms caused to the victims of rights violations?

This way, the constraints-sceptical dialectic aims at drawing the constraintist back into the grip of the paradox, whatever answer she may try to

give to the various sceptical questions.[49] The aim of this thesis is to break out of this dialectic. The rationality paradox and the value paradox together constitute the paradox of deontology. Only if we can explain both *that* the agent should be concerned with something other than the minimisation of rights violations overall and *how it can be plausible* to assume that she should have this other concern, we can hope to find a satisfactory solution to the paradox.

Note that the value paradox has many faces. It manifests in a variety of objections against ways to answer the rationality paradox and, as such, depends on the details of that answer. For instance, the value paradox troubles the agent-centred approach in a different shape than the one in which it troubles the agent-neutral approach. It may take the form of an internal or an external criticism against the inviolability account; and so on. It is still insightful to identify these objections and criticisms as instances of one and the same problem, or one and the same stage of the constraints-sceptical argument as doing so reveals something about the structure of the paradox of deontology, and the requirements for a convincing solution to it.

[49] Thus, it is important to note that the value paradox is *not* an entirely separate issue. It is simply another stage of the challenge posed by the paradox of deontology, a stage at which the constraints-sceptic might try to draw the constraintist's answer to the rationality paradox back into the grip of the paradox.

2 *The Significance of Agency*

2.1. Introduction

As we have seen, an approach to answer the paradox of deontology may take one of two forms. The agent-centred approach aims to justify constraints by reference to the thought that reasons against rights violations are—in some substantive sense—relative to the agent acting. According to this approach, the rationale for constraints is to be found in the idea that each agent's primary moral concern must lie with her own (present) actions. By contrast, the agent-neutral approach aims to justify constraints without reference to the idea that reasons against rights violations are agent-relative in any substantive sense.[50]

However, this characterisation is somewhat deceptive. It takes the agent-centred and the agent-neutral approach to be two options on a par with each other, two ways to respond to the charge of paradox, as if we could simply choose the one which would seem most promising to us. As a matter of fact, however, the agent-centred approach rests on moral philosophy's

[50] As we will see later, there is a sense in which any moral reason is relative to the owner of that reason. Thus, if the agent-relative/agent-neutral distinction is to be a meaningful one, there must be a more substantive sense in which some, but not all reasons are agent-relative reasons (see Sections 2.2.2 and 2.2.3).

default understanding of deontological constraintism. According to what deserves to be called the *standard view*, constraintists *must* refer to agent-relativity in order to explain the peculiar normative force of deontic constraints.[51] How else could they justify the claim that the agent *herself* ought not to kill or torture even where this would mean that others commit many more killings or tortures? According to the standard view, deontic constraints *just are* agent-centred restrictions.[52]

Many constraintists have embraced the standard view. I think this is a mistake. As I have argued in the last chapter, the agent-centred approach cannot avoid the value paradox, i.e., the problem that it seems morally unacceptable to think that my own agency matters so much more than the fate of those I could save. If there is a morally acceptable justification of the claim that I ought not to kill even to prevent more killings, then this justification must rest on something other than the fact that I *in particular* should not kill.

In this chapter I therefore aim to do two things. For one thing, I show that we can make sense of deontic constraints in solely agent-neutral terms. It is possible, at least, to understand constraints without reference to agent-relativity.

For another thing, I argue that an agent-neutral account of constraintism is preferable to an agent-relative one. Its preferability rests, as I see it, on three major advantages. First, an agent-neutral account, unlike an agent-relative one, allows constraintists to conceptualise morality as a shared endeavour describable in terms of moral aims that all agents share. Second, an agent-neutral account captures the difference between *general* and

---

[51] By the "peculiar force" of constraints (McNaughton and Rawling 1991: 169) or their "logical peculiarity" (Nagel 2008: 106) what is meant is simply the feature that constraints prohibit minimising violations of rights.

[52] Much work in defending the standard view has been done by David McNaughton and Piers Rawling (1991), (1992), (1995a), (1995b), (1998), and more recently by Matthew Hammerton (2017), (2019), (2020). The list of its supporters further includes Parfit (1984), Kagan (1989), Dreier (1993), Zong (2000), Hare (2013), and Setiya (2018). This is just to mention a few. As Michael Ridge notes, the idea that constraints are necessarily agent-relative "is close to being an orthodoxy" (Ridge 2017: §7).

special constraints, a difference which is untraceable on an agent-relative account. And third, an agent-neutral account helps deontological constraintists to resist the claim that consequentialism, *not* deontology, provides the best overall account of deontic constraints. An agent-relative account, by contrast, makes constraints an easy target for consequentialising.[53]

Section 2.2 begins with some remarks on the agent-relative/agent-neutral distinction when applied to the logical form, substantive content, and rational justification of constraints. Section 2.3 asks what motivates the standard view. In this context, I discuss Tom Dougherty's argument for agent-neutral constraints and argue that it is insufficient to show that constraintists do not need to refer to agent-relativity to make sense of deontic constraints. In a nutshell, the problem with Dougherty's view is that he rejects an agent-relative *justification* of constraints without offering any alternative justification. Section 2.4 illustrates how offering such an alternative, agent-neutral justification of constraints could redeem the promises of a deontological account of constraintism that generates the desired verdicts in all relevant types of cases and is robustly agent-neutral. Section 2.5 lays out the advantages of an agent-neutral account of constraintism as opposed to a standard agent-relative one. And finally, Section 2.6 draws some general lessons from this on the conceptual distinction between agent-relativity and agent-neutrality.

---

[53] This last point may not strike the reader as an immediate advantage of agent-neutral accounts. So long as the consequentialised version of constraintism secures the correct extension (renders the correct kinds of actions impermissible) and deontological constraintists care more about this extension than the fact that they are deontologists, consequentialised constraintism might even hold certain promises for those constraintists (see Chapter 5, Section 5.2). However, as I argue in Chapter 5, consequentialised constraintism cannot avoid the value paradox because it is based on an implausible conception of value.

2.2. The Agent-Relative/Agent-Neutral Distinction

It is a commonplace in ethics to think that we can distinguish between agent-relative and agent-neutral moral theories. The agent-relative/agent-neutral distinction is generally understood as a binary opposition applicable to various aspects of morality including values, aims, reasons, and rules.[54] Moral theories, I assume, can be agent-relative or agent-neutral only by virtue of the fact that they understand these other aspects of morality in agent-relative or agent-neutral terms.

The agent-relative/agent-neutral distinction is widely acknowledged as an important conceptual resource. For one thing, it seems to provide an adequate terminology for talking about the normative force of agent-centred restrictions, special obligations, and agent-centred permissions.[55] For another, it has come to play an important role in taxonomizing normative theories. Most notably, the agent-relative/agent-neutral distinction is commonly used to draw the central line between deontological and consequentialist moral theories. Almost two decades ago, Thomas Hurka called it one of the "greatest contributions of recent ethics" (Hurka 2003: 628)—and it has since not lost its prominent place in moral philosophy. Despite its popularity, however, it is difficult to get a firm grip on the distinction, and some have

---

[54] Nagel (1970), (1986), (2008) applies the agent-relative/agent-neutral distinction to moral reasons and to values. Pettit (1987), too, talks about reasons. Parfit (1984) applies the distinction to moral aims. Scheffler (1985), (1994) talks about agent-relative/agent-neutral principles and standpoints. Dancy (1993) draws the distinction in terms of points of view from which we can recognise reasons. And McNaughton and Rawling (1991), (1992), (1995a), (1995b), (1998) primarily use the distinction to analyse moral rules.

[55] In contrast to natural duties owed to all persons qua persons, *special obligations* are duties owed to particular persons and are usually associated with the existence of special relationships such as friendship (e.g., Jeske 1998). *Agent-centred permissions* (or *agent-centred options*) exist where morality permits us to act suboptimally to preserve something that we care about. For instance, it might be permissible for me to do a PhD in philosophy instead of dedicating my time solely to the task of improving the situation of children in Uganda. Other examples for agent-centred permissions are morally permissible self-defence and permissions to fail to maximise the good where this would require self-sacrifice (e.g., Steinhoff 2016, Quong 2016, Lazar 2019).

suspected that it might not be a meaningful or useful distinction in the first place (Korsgaard 1993, Rønnow-Rasmussen 2009, Schroeder 2011).

To a large extent, the present chapter can be read as a critical analysis of the agent-relative/agent-neutral distinction. However, my main motivation for providing such a critical analysis is not so much to give an account of the distinction itself than to understand the extent to which constraintists might, roughly speaking, be committed to agent-relativity. And my primary motivation for gaining such an understanding is to overcome the standard view, to pave the way for an agent-neutral approach to the paradox of deontology.

Accordingly, I shall not begin by trying to give a precise definition of agent-relativity and agent-neutrality. A deeper understanding of these terms is one of the aims of this chapter, not its starting point. But it might be useful to say what I think these terms mean in the roughest possible sense: agent-relativity, as I see it, is the thought that when determining the deontic status of an act the identity of the agent acting *matters*—agent-neutrality is the absence of agent-relativity.[56] In what follows, I will elaborate on what this could mean.

2.2.1. Agent-Referencing

Sometimes, it matters morally whether I or someone else ought to perform an act. The most obvious candidate of an agent-relative moral commitment is perhaps the duty to keep one's promises. Since no one else can keep my promises because they are *mine*, it seems to matter that *I* ought to keep

[56] One advantage of understanding agent-relativity in this way is that it allows us to understand related phenomena such as patient-relativity and moment-relativity. Patient-relativity is the thought that the identity of the patient acted upon matters when determining the deontic status of an act. Moment-relativity is the thought that the identity of the moment in which an act is performed matters when determining the deontic status of that act. This chapter focuses on the phenomenon of agent-relativity. I have little to say about patient- and moment-relativity here. For a more detailed analysis of these phenomena, see Hammerton (2016).

them. What must feature in a full explanation of my duty to keep some promise is a reference to the fact that the relevant promise is one that *I*, not anyone else, have made.[57]

Many philosophers since Nagel have therefore argued that agent-relativity is indicated by a form of *back-reference* to the agent acting (e.g., Nagel 1986, Pettit 1987, McNaughton and Rawling 1998, Rønnow-Rasmussen 2009). Nagel himself, who focused the discussion primarily on moral reasons, says that an agent-relative reason is one that cannot "be given a general form which does not include an essential reference to the person who has it" (Nagel 1986: 152–153). Accordingly, an agent-neutral reason is distinguished by the possibility to state that reason in a general form *without* any such agent-reference.

An agent-reference is essential, then, if and only if it *must* feature in a general statement of someone's reason for action—or a full specification of that reason (Pettit 1987: 75). If it would not matter that *I* ought to keep some promise, a general statement of my reason to keep that promise would not have to feature a reference to myself, the promisor. The reference to myself would not be essential but *eliminable*.[58] But since it matters that *I* ought to keep my promise, it seems that a general statement of my reason to keep it must feature a thus essential reference to me.

Although the definition of agent-relativity in terms of agent-referencing might seem sufficiently clear, a closer look at the question when agent-

---

[57] What if I promised the following: someone will come and pick you up at 10pm? It looks like *anyone* could come and pick you up at the right time, and my promise had been kept. So, how does it matter that *I* keep it? It matters that I keep my promise because the *responsibility for redeeming that promise*, we might say, still lies with me, not with anyone else. Suppose that I fail to make sure that someone picks you up but, by accident, someone else picks you up not even knowing that I have promised you anything. In this case, my impression is that I would have failed to keep my promise, even though the outcome of *what* I have promised is secured.

[58] Here, I follow the most common interpretation of an *essential* agent-reference as one that is *ineliminable* (e.g., McNaughton and Rawling 1998: 38); see Sections 2.2.2 and 2.2.3 for a more detailed discussion of eliminability.

references are essential and when they are eliminable will reveal that it is not so obvious after all what it means for an agent-reference to be *truly essential*. One difficulty in getting a firm grip on the agent-relative/agent-neutral distinction is, I think, that the occurrence of an agent-reference is a matter of the *logical form* of a moral principle, whereas the question whether this reference is essential goes *beyond* matters of mere form.[59] Another way to read this chapter is therefore as a proposal on how we are to mediate between questions concerning the form, content, and justification of deontic constraints when applying the agent-relative/agent-neutral distinction.

2.2.2. Formal Agent-Relativity

First and foremost, one might think that the agent-relative/agent-neutral distinction is a distinction between different forms that moral principles could take. They may take an *agent-relative* or an *agent-neutral* form. Consider the generic constraint:

> Deontic Constraint. It is impermissible to φ even to prevent more further φ-ings.

As it stands, it is just not obvious whether this is an agent-relative or an agent-neutral principle. The statement contains no agent-reference, but it is not even clear what a similar statement that *did* contain such a reference would look like.

David McNaughton and Piers Rawling have therefore suggested that when determining whether a principle takes an agent-relative or an agent-neutral form our first step must be to give that principle a form which makes its agent-relative or agent-neutral character *visible*. They argue that any

[59] By logical form I simply mean the abstract form (or structure) of a principle, as distinguished from its particular content or *what* it is that principle requires us to do.

moral principle, whatever its content, can be given the form of a rule that states: *Each agent should ensure that…*, followed by a statement of that principle's ethical content (McNaughton and Rawling 1992: 873). For the sake of clarity, they use a semi-formal notation containing agent-variables ("*x*", "*y*", etc.). Thus, on McNaughton and Rawling's account, any moral principles can be given the form:

(*x*) (*x* should ensure that […])

The content of what it is that each agent should ensure is to be inserted in the square brackets.

A rule is *formally agent-relative* if and only if there is an occurrence of "*x*" in the square brackets bound by the initial universal quantifier. Otherwise, it is *formally agent-neutral*. (Expressed in ordinary language terms, formally agent-relative rules contain an agent-reference "*she*" after the "ensure that…", which refers to the agent to whom the rule is addressed.)[60]

By way of illustration, we can give the constraint an agent-relative form on McNaughton and Rawling's account:

Agent-Relative Constraint (ARC). (*x*) (*x* should ensure that [*x* does not *φ* even to prevent more further *φ*-ings])[61]

ARC should be read as: for any agent, that agent should ensure that *she* does not *φ* even to prevent more further *φ*-ings. *ARC* is formally agent-relative

[60] McNaughton and Rawling lay out their semi-formal account in several joint contributions, but perhaps most concisely in McNaughton and Rawling (1995b). For a criticism of their account see Rønnow-Rasmussen (2009); for a recent defence of their account see Hammerton (2019).

[61] To my knowledge, McNaughton and Rawling do not make it explicit that this would be the correct translation of a constraint on their account. However, it is suggested by the way they mean to translate more general deontological rules like "Each agent should not lie" (McNaughton and Rawling 1991: 177) or "Each agent should not kill innocents" (McNaughton and Rawling 1995b: 34).

because there is an occurrence of "*x*" in the square brackets bound by the initial universal quantifier.

As we have seen, the question whether the agent-reference represented by the occurrence of "*x*" in the square brackets is an essential one, we need to know whether it is ineliminable—but ineliminable *in the light of what*? On merely formal grounds, there is nothing that keeps us from giving the constraint an alternative, *agent-neutral* reading. All we need to do is to introduce a second agent-variable "*y*" to replace the agent-variable "*x*" in the square brackets which only referred to the agent to whom the constraint is addressed. We get:

> Agent-Neutral Constraint (ANC). (*x*) (*x* should ensure that [(*y*) (*y* does not *φ* even to prevent more further *φ*-ings)])[62]

ANC should be read as: for any agent, that agent should ensure that *no one* *φ*-s even to prevent more further *φ*-ings.

ANC contains no agent-reference in the square brackets that would be bound by the initial universal quantifier; hence, *ANC* is agent-neutral in form. *ANC* of course *entails ARC*—and some might wonder whether it deserves to be called a constraint just because it entails that (agent-relative) constraint. I will address this worry in Section 2.4.3. If we grant for the moment that *ANC* can be called a constraint, then it is evident that *ANC* is an *agent-neutral* constraint. It does not tell the agent only to ensure that *she* does not *φ*. Instead, it tells the agent to ensure that she does not *φ and* to ensure that no one other than herself *φ*-s to prevent more *φ*-ings. This is so because the agent-variable "*y*" can refer either to the agent to whom *ANC* is addressed ($x = y$) or to any other person ($x \neq y$). As with any other agent-neutral principle, *ANC* of course may require the *particular agent* to act in a

[62] This is in line with how McNaughton and Rawling handle the translation of other agent-relative rules into agent-neutral ones. A similar transition from a formally agent-relative constraint to an agent-neutral one is proposed by Dougherty (2013): 531.

certain way. Agent-neutral rules require the agent to ensure "that something is true of *her*—but only insofar as she is *one amongst many*" (McNaughton and Rawling 1991: 179); more on this later.

When I say nothing keeps us from transforming *ARC* into *ANC*, what I mean is simply that it is possible to yield *ANC* from the structure of *ARC*. This has no bearing on the question whether either of *ARC* or *ANC* are *plausible* readings of a constraint on *φ*-ing. *ARC* and *ANC* are just *possible* readings of that constraint, and the difference between them is, first and foremost, a difference in their logical form or structure.

Note that, on merely formal grounds, nothing keeps us from eliminating an agent-reference from a moral principle, *nor* from inserting one. By way of illustration, consider:

> **Minimising Killings Principle.** Killings should be minimised.

It might seem natural to want to give this principle an agent-neutral form:

> **Agent-Neutral Minimising Killings Principle (ANM).** (*x*) (*x* should ensure that [(*y*) (*y* minimises killings)])[63]

But nothing seems to keep us from giving it an agent-relative form either:

> **Agent-Relative Minimising Killings Principle (ARM).** (*x*) (*x* should ensure that [*x* minimises killings])

On merely formal grounds, there is nothing that would distinguish the minimising killings principle as an *agent-neutral* principle from a constraint on killing. Thus, if *ARC* is supposed to be *truly* agent-relative, as the standard view

---

[63] Alternatively, *ANM* could be stated as: (*x*) (*x* should ensure that [killings are minimised]). I take these two statements to be logically equivalent. The ethical content that killings should be minimised will in any imaginable case mean that *some agent* should see to it that there are fewer killings.

suggests, then there must be some other sense in which the agent-reference it contains is *ineliminable*. For *any* agent-reference is eliminable on merely formal grounds. In other words, the sense in which the agent-relative/agent-neutral distinction concerns the mere form of moral principles cannot be the sense in which the standard view takes constraints to be truly agent-relative. Evidently, it is possible to give *any* moral principle both an agent-relative and an agent-neutral form. Therefore, if the standard view would simply hold that constraints are formally agent-relative, then it would be a trivial view: all moral principles can be given an agent-relative form.

I do not think that the standard view is trivial. It takes constraints to be agent-relative *in a more robust sense*. Constraints are *robustly agent-relative* if and only if they *cannot* be given an agent-neutral form, i.e., as Nagel said, if they cannot be given a general form that does not contain an essential agent-reference.[64] But if it seems that nothing keeps us from giving an agent-relative principle an agent-neutral form, and vice versa, then how can any principle be in this sense robustly agent-relative?

2.2.3. What Is Robust Agent-Relativity?

Before tracking down the sense in which the standard view takes constraints to be agent-relative, let me point out that the considerations of the last section pose a general challenge for advocates of the agent-relative/agent-neutral distinction.

The challenge might be most obvious for the case of agent-neutral reasons. As Rønnow-Rasmussen puts it, practical reasons are *personalisable*,

---

[64] McNaughton and Rawling (1995b): 34 speak of "genuine" agent-relativity. My impression is that the term *genuine agent-relativity* is ambiguous. Since on McNaughton and Rawling's account every moral rule can take an agent-relative form, one could think that agent-relativity is a feature of all moral rules and in this sense of the word *genuine*. I prefer to say *robust agent-relativity* because what makes a moral rule agent-relative in a non-trivial sense is, as Nagel pointed out, the impossibility to give it an agent-neutral form: the agent-relative form of some moral rules is *robust*, i.e., unchangeable in some sense to be specified further.

i.e., each practical reason can be thought of as a reason *for someone* to do something. This means that there is a sense in which any practical reason is relative to the owner of that reason:

> Since all reasons apparently are reasons for someone and a reason is only a reason for someone, if it somehow involves or refers to this someone, it follows that all reasons are by their very form reasons that refer to the person who has the reason to *φ*. This, in its turn, is just another way of saying that all reasons are on entirely formal grounds, agent-relative reasons. (Rønnow-Rasmussen 2009: 230–231)

But if all practical reasons are personalisable and in this sense agent-relative, how could there ever be such a thing as an *agent-neutral* reason? Rønnow-Rasmussen continues:

> It is always logically legitimate to ask what it is about some fact that makes it into someone's reason for action. And in the case of a so-called agent-neutral reason, the test is how does *P* (which expresses this fact), that allegedly contains no reference of any sort to *x*, express a reason *for x* to *φ*? (Rønnow-Rasmussen 2009: 234)

Rønnow-Rasmussen's point seems to be that we cannot even make sense of a reason that is *not* a reason *for someone*. If this is so, and if being a reason for someone means that the reason somehow refers to its owner, then what constitutes true agent-neutrality? If there are no reasons that are entirely agent-neutral, then a truly agent-neutral reason would be an illusionary concept.

Advocates of the agent-relative/agent-neutral distinction may respond to the personalisability challenge by denying the relevance of this mere formal sense of agent-relativity. Parfit, for instance, explains that since "even

agent-neutral reasons can be, in this [formal] sense, agent-relative, this sense is irrelevant to the discussion" (Parfit 1984: 143). But this is simply begging the question. For the fact that there is a sense in which all practical reasons are agent-relative does not by itself show that this very sense is irrelevant to the discussion, *especially* if this supposedly trivial sense affects the intended applicability of the distinction. Instead, it could as well show that the distinction between agent-relativity and agent-neutrality itself is a trivial one, if we cannot identify some other sense of the distinction in which only some, but not all reasons are agent-relative.

McNaughton and Rawling offer the following response to the personalisability challenge. They explain that while any agent-neutral rule (reason, etc.) can be converted into an agent-relative one "there is no way of converting a genuinely [agent-relative] rule to [agent-neutral] form without a change in content" (McNaughton and Rawling 1993: 86). For instance, once we transform the agent-relative constraint *ARC* into *ANC*, we have given the constraint an agent-neutral form. But we have also changed its content. *ANC* requires the agent to ensure that *all* agents do not φ even to prevent more further φ-ings, whereas *ARC* requires the agent only to ensure that she *herself*—a very small subset of 'all agents'—does not φ even to prevent more further φ-ings. According to McNaughton and Rawling, this would indicate that *ARC* is *genuinely agent-relative*: it cannot be given an agent-neutral form without altering its original content.

However, I do not think that this resolves the issue. If the criterion for robustness is that a rule cannot change its form without also changing its content, then *all* agent-relative and *all* agent-neutral rules are robust.

To see why, recall the agent-relative version of the minimising killings principle, *ARM*. *ARM* cannot be given an agent-neutral form without a change in content. Once we translate *ARM* into *ANM*, the agent is no longer only required to minimise killings *herself*. She is now also required to ensure that others minimise killings. Thus, *ARM* is robustly agent-relative on McNaughton

and Rawling's account. In turn, once we translate *ANM* back into *ARM*, the agent is no longer required to ensure that *everyone* minimises killings. She is now only required to ensure that she *herself* does. Thus, *ANM* must be robustly agent-neutral. The same goes for the constraint. *ARC* and *ANC*, the two possible readings of a constraint on *φ*-ing, cannot be translated into one another without changing the content of what the constraint requires the agent to do. Thus, *ARC* comes out as *robustly* agent-relative on McNaughton and Rawling's account, whereas *ANC* comes out as *robustly* agent-neutral.

This should not come as a surprise. Since McNaughton and Rawling had defined agent-relativity in terms of agent-references that occur *within the content-part* of moral principles, it is not surprising that no moral principle can change its form from agent-relative to agent-neutral, or vice versa, without changing its content as well.

To be clear, the problem here is not just a conceptual one. It is not that on McNaughton and Rawling's account there is no such thing as *non-robust* agent-relativity or agent-neutrality. The problem is that what we needed—or rather what the defender of the standard view needed—was not proof that the *agent-relative version* of a constraint is truly agent-relative (this is all McNaughton and Rawling's account allows us to say) but that the *constraint itself* is truly agent-relative (this is what it does not allow us to say). Until now, no argument has been provided to the effect that *ARC* is the correct or preferable reading of a constraint on *φ*-ing, the one that really captures what that constraint is all about.

Thus, if what I have said so far is plausible, then the most straightforward argument against the standard view would simply run as follows:

(1) Constraintism postulates constraints of the generic form 'It is impermissible to *φ* even to prevent more further *φ*-ings.'
(2) We can interpret that constraint as *ANC*: '(*x*) (*x* should ensure that [(*y*) (*y* does not *φ* even to prevent more further *φ*-ings)]).'

(3) *ANC* is robustly agent-neutral because it cannot be given an agent-relative form without altering its content.

(4) Therefore, constraints are robustly agent-neutral principles.

Of course, the same kind of argument could be used to show that constraints are robustly *agent-relative*. All we would have to do is to replace premise 2) with:

(2') We can interpret that constraint as *ARC*: '(*x*) (*x* should ensure that [*x* does not *φ* even to prevent more further *φ*-ings]).

The argument would then take an alternative route to the conclusion that constraints are robustly agent-relative.

Where does this leave us? I think McNaughton and Rawling rightly point out that robust agent-relativity can only be a matter of the content, not the mere form of moral principles. But I disagree—for the reasons given above—that robust agent-relativity is indicated by the fact that a change of form would entail a change in content. Instead, I think that robust agent-relativity is a matter of what we think is the *appropriate content* for a moral principle. In particular, whether the generic constraint is robustly agent-relative or not depends on whether we think *ARC* actually captures the content which we think that constraint should have. Let me elaborate on this.

Consider an advocate of *ARC* who goes on to reject that *ANC* is a plausible reading of a constraint on *φ*-ing. She might explain this by something along the following lines:

'We are not responsible for what others do, at least not as much as we are responsible for what we ourselves do. It is as Williams says, "each of us is specially responsible for what he does, rather than for what other people do" (Williams 1973: 99). Therefore, when I say that no one should kill to prevent more killings, what I really mean is that each agent should ensure that *she* does not kill. Each agent's primary concern should lie with her own

actions, not with those of others. Thus, I reject the view that *ANC* is a plausible reading of a constraint on killing because *ANC* does not capture the fact that each agent's primary moral concern should lie with what *she* does.'

It is obvious to me that an advocate of *ARC* who argues in this way makes a claim about *what it means* for there to be a constraint on *φ*-ing. She rejects *ARC* on the basis that *ARC* does not capture what she takes the content of that constraint to be. On her view, the generic constraint cannot be given an agent-neutral form because in its agent-neutral form the constraint does not capture what a constraint on *φ*-ing is all about. In this sense, her view is that the agent-reference in *ARC* is ineliminable—it is ineliminable *in the light of what she believes the content of that constraint should be*. Nothing more, nothing less. In other words, such an advocate of *ARC* thinks of constraints as agent-relative in a robust sense. She is a true agent-relativist or, more precisely, a true *agent-relative constraintist*.

In turn, it should be obvious that we can think of constraints as agent-relative *in form*, without thinking that they give each agent a special concern with what *she* does. Logical form and substantive content of a constraint are two separate things. Whereas calling a moral principle agent-relative is a statement about its logical form, calling it *robustly* agent-relative is a statement about what its content should be. Robustly agent-relative constraints *cannot* be given an agent-relative form only in the sense that a certain, agent-relative interpretation of what they are all about resists that transition. But so far, we have heard no argument as to why we should think that the agent-relative interpretation is the *correct* interpretation of a constraint on *φ*-ing.

2.2.4. The Justification of Constraints

Similarly, form and justification of a constraint seem largely independent issues. In general, it seems that we can think of constraints as taking an agent-relative form without thinking that they require an agent-relative

justification, one that makes use of agent-relative considerations as to why we should not perform minimising violations.

Moreover, we can think of constraints as being robustly agent-relative without having provided any justification for them. Saying that a constraint gives each agent a special concern with her own actions does not entail any explanation as to why we should think that this is a plausible claim to make. We can still ask: what is the justification for a constraint that tells me to ensure that *I* do not kill even if others do?

That said, there is a strong dependency between the content-interpretation and the justification of constraints. For instance, if we think of constraints as robustly agent-relative in the sense that they require us to prioritise our own compliance with them, then they call for an agent-relative justification, one that explains why constraints *as robustly agent-relative restrictions* are in place. For why should we think that constraints are robustly agent-relative, if it turns out that their normative force can be explained—if they can be justified—conclusively without reference to the idea that each agent has a special concern for what *she* does? An agent-relative justification would only explain why constraints *as agent-relative restrictions* are in place. Once we reject the standard view that constraints must be understood as agent-relative restrictions, constraints call for an alternative, agent-neutral justification.

Summarising the previous sections, I have argued that the binary distinction between agent-relativity and agent-neutrality is orthogonal to a threefold distinction between the logical form, substantive content, and rational justification of moral principles. Robust agent-relativity is a matter of what we think the content of a constraint should be and, as such, goes beyond matters of mere logical form. Whether we need to provide an agent-

relative or an agent-neutral justification of constraints depends on whether we think of constraints as robustly agent-relative or agent-neutral.

In general, my view is that constraints are best understood as robustly agent-neutral by virtue of their substantive content and as such require an agent-neutral justification. Yet at the same time they are still best understood as taking an agent-relative form: the agent acting usually confronts a constraint as a restriction on what *she* may not do, even if the reason *why* she may not do it is in no substantive sense relative to her.

2.3. Constraintism and the Issue of Agency

Before turning to the question what could make us think that constraints are robustly agent-relative principles, let me summarise what we have learned about the standard view on constraintism.

First, the standard view is not a view about the logical form of constraints. It is a view about the substantive content of constraints as robustly agent-relative principles. Constraints are robustly agent-relative just in case they give each agent a primary concern with her own actions. Agent-relative constraints, as Nagel says, "govern each agent's conduct only with respect to the killings that she might commit" (Nagel 2008: 105). An agent-relative account of constraintism thus requires us to attend especially to our own actions rather than to "adopt an impartial perspective that treats our actions as […] on a par with those of others" (Johnson 2019: 283).

Second, although the standard view is often articulated as a view on the status of *deontology*, it is more precisely described as a view on the status of deontological *constraintism*. It focuses on deontology's commitment to the existence of deontic constraints. In a more general context, deontologists often think that there are other moral principles that might indicate an underlying agent-relative understanding of morality, such as special obligations and agent-centred permissions. This is not the place to discuss whether the belief in something like special obligations would commit deontologists to

agent-relativity or not—one might think that it does.[65] The question I am interested in is whether the standard view is correct to assume that constraints *just are* agent-relative principles. I will therefore continue to talk about constraintism rather than deontology.

And third, although it is usually formulated as a view on *deontology's* commitment to constraints, the standard view strictly speaking implies that *all* constraintists are agent-relativists. Everyone who believes in the existence of deontic constraints is committed to the belief that the relevant moral prohibitions, like the prohibition against killing or torture, are relative to the agent acting in the robust sense described above.[66]

In what follows, I will discuss whether constraintists are committed to accepting the standard view. Comparatively few have explicitly argued against it.[67] Most notably, Dougherty (2013) aims to directly reject the standard view. I share his initial hope that "with a little reflection some [constraintists] would reject the claim that they are concerned with agent-relativity" and instead adopt a position according to which a constraint on φ-ing

---

[65] While I indeed find it hard to make sense of special obligations without reference to agent-relativity, I am not convinced that so-called agent-centred permissions or options are agent-relative in any deeper sense. On the one hand, personal relationships are optional: May might be Chloe's best friend, but she may as well not be. Thus, whether Max has some special obligation towards Chloe as her best friend depends on some fact about *Max* (whether Max *is* Chloe's best friend or not). On the other hand, whether Max is permitted to undertake graduate study in philosophy instead of optimising the situation of children in Uganda does not seem to depend on any fact about Max. It does certainly not depend on the fact whether Max has *an interest* in undertaking graduate study in philosophy in the first place. It is an option Max has, notwithstanding Max's interest in making use of that particular option.

[66] The standard view leaves conceptual room for views that hold that we should not perform minimising violations but are, at the same time, not constraintist views. One such view is articulated by Setiya (2018) who argues that agent-neutral consequentialism can accommodate the idea that we should not kill or torture, even where this would minimise killings or tortures overall. Yet he does not think of his view as a constraintist view because he understands constraints as agent-centred restrictions (in the robust sense). More on this in Chapter 5.

[67] Among these are Frances Kamm and Eric Mack. I will discuss Kamm's view extensively in the following chapters. Mack (1998) does, strictly speaking, not endorse the idea that constraints are agent-neutral because he holds that reasons we have to obey constraints are neither agent-relative nor agent-neutral.

"does not give the agent any special concern with her own [*φ*-ings]" (Dougherty 2013: 530–531). However, I think Dougherty's vision of an agent-neutral account of constraints fails to deliver on its promises. I aim to redeem those promises in Section 2.4.

2.3.1. The Special Case

First, we should ask what motivates the standard view. Why should we think that in order to make sense of the normative force of deontic constraints, constraintists must refer to the idea that agents should be concerned in particular with their own actions?

I think Eric Mack offers a helpful explanation in this regard (Mack 1998: 63). He asks us compare two general cases:

> Normal Case. *X* should not kill *A.*
>
> Special Case. *X* should not kill *A* even to prevent the killing of *B* and *C.*

It seems that if it were the agent-neutral disvalue of killing that normally makes my killing *A* wrong, then in the *Special Case* in which I could prevent two further killings by killing *A*, I should kill *A*. Simply put, one rather than two impersonally bad killings would occur if I did. Therefore, if I ought *not* to kill *A* even in the *Special Case* where this would prevent the killing of *B* and *C*, then this must be because what makes my killing *A* wrong is that *I*, as opposed to others, would be doing the killing. For the only other difference between the killing of *A* and the killing of *B* and *C* is that the killing of *B* and *C* entails the killing of an extra person. Certainly, this second difference makes the killing of *B* and *C* seem worse than the killing of *A*. From an agent-neutral standpoint,

it seems as though it simply cannot be preferable that I spare *A*'s life where this would mean that *B* and *C* get killed.

Thus, it looks as though the only way to make sense of the claim that I ought not to kill *A* is by reference to the fact that my primary concern should lie with my own potential killings. First and foremost, I am to avoid killing anyone myself. As an uninvolved impartial observer, it seems that I should want one instead of two to get killed. Yet if I should refuse to become the agent of that one killing, then this must be because *I* in particular ought not to kill. The deontic status of the act of killing *A* (its being impermissible) must depend on the fact that I, as opposed to others, would perform that act. Because the only other difference between the killing of *A* and the killing of *B* and *C* strongly suggests that I *should* kill *A*.

In comparison, take a standard eliminativist account of morality, i.e., classical act-consequentialism. Classical act-consequentialism appeals only to what makes outcomes better or worse, whoever would produce them, and tells us to always maximise impersonal good. In particular, it tells me to kill *A* in the *Special Case* because an outcome that contains the killing of one is better than an outcome that contains two other killings. Classical act-consequentialism, then, is what Parfit calls an *agent-neutral* moral theory because it gives "to all agents common moral aims" (Parfit 1984: 27). It gives me the aim[68] that killings are minimised, and it gives you the aim that killings are minimised, and so on. Deontological constraintism, however, does not seem to assign shared moral aims to all agents. Instead, it gives me the aim to ensure that *I* do not kill, and it gives you the aim to ensure that *you* do not kill.

Thus, one way to set up the standard view is to point out that deontological constraintism, as opposed to classical agent-neutral consequentialism, makes it a significant fact that I, as opposed to others, would perform an

[68] I understand *moral aims* broadly here—in contrast to the narrower term 'goals'—as whatever it is that our actions should be directed at or be guided by, morally speaking. Expressing respect for the value of a commitment to keep a promise can thus be a moral aim just like producing an outcome in which one has kept a promise.

act. The issue of agency—the question concerning who performs an action—is relevant to a classical consequentialist understanding of morality "only to the extent that it affects the results" (Zong 2000: 678). It is relevant only to the extent that it is sometimes within in the reach of *my* agency to produce better outcomes. But what I ought to do can fully be described on such an account by making use of passive phrases such as "It should be brought about that *S* obtains", where "*S*" stands for some state of affairs that entails the achievement or partial achievement of our common moral aims. What I ought to do on a constraintist account of morality, however, cannot fully be described by making use of passive phrases like "It should be brought about that…" because what matters is that *I* ought to do it.

Constraintists could interject here that their intention was never to claim that *I* ought not to kill in the *Special Case* but rather that *everyone* ought not to kill even to prevent more killings. This, after all, is what makes constraintism a *moral* theory: that its principles quantify over the set of all moral agents. How then does constraintism not give the same moral aims to everyone, for instance, the aim not to kill *A* in the *Special Case*?

It may be useful here to make a further distinction between *formal* and *substantive* moral aims (Portmore 2013b: 162). Advocates of the standard view could argue that insofar as constraintism tells all agents to obey a deontic constraint on killing, it can be said to give all agents the same *formal* aim. But insofar as obeying that constraint appears to have different implications for different agents, it does not assign to everyone the same *substantive* aims. Instead, it gives me the substantive aim to ensure that *I* do not kill, and it gives you the substantive aim to ensure that *you* do not kill.

This is just another way of saying that universality does not guarantee agent-neutrality (e.g., Nagel 2008: 105). If the agent-relative/agent-neutral distinction is to be a non-trivial one, it cannot be simply because rules address the set of all moral agents that these rules come out as agent-neutral. For instance, a moral rule that tells everyone not to lie might apply universally to

all agents and yet might be understood as giving each agent the exclusive aim that *she* does not lie.

2.3.2. The Special Prevention Case

Still, it is not obvious that constraintists cannot assign the same substantive moral aims to all agents. If John and Herbert are each required not to kill *A* in the *Special Case*, perhaps they should also want Donald—and each other—to refrain from killing *A* in the *Special Case*. All the constraintist needs is the quite natural thought that we should want others not to do wrong—we should be against wrongdoing—and the aim of not killing *A* then begins to look a lot like a substantive aim shared by all agents.

This seems to be what Tom Dougherty has in mind when he puts pressure on the standard view. First, he asks us to imagine a bystander-version of the *Special Case* where a previously uninvolved bystander must decide whether to intervene and prevent you from killing *A* (to prevent the killing of *B* and *C*). I call this case the *Special Prevention Case*:

> Special Prevention Case. *X* should prevent that [*Y* kills *A* even to prevent the killing of *B* and *C*].

For the sake of argument, assume that preventing the killing of *B* and *C* is not under *X*'s direct control. All *X* can do is prevent *Y* from killing *A*. What can constraintists say the bystander should do? Dougherty comments:

> A [constraintist] is free to say that the bystander should be opposed to your killing the single person, even though she knows that this will lead to more deaths overall. So if the bystander were able to intervene to prevent your killing the person you could kill, then she ought to do so. Similarly, the [constraintist] can say that the bystander ought to prefer that you do not kill. Indeed, I suggest that these are rather

> attractive claims for the [constraintist] to make. (Dougherty 2013: 530–531)

Thus, Dougherty argues that constraintists should prefer an account of constraints that can accommodate the *Special Prevention Case*. Such an account would include constraints that capture the thought that we should prevent others from violating those constraints, thus providing for the *enforcement dimension* of constraints.[69] Once constraintism accommodates the *Special Prevention Case*, it seems, not killing *A* looks like a substantive moral aim shared by all agents. Each agent should not kill *A* and each agent should prevent others from killing *A*. No differences are made between the agent's potential killing of *A* and someone else's potential killing of *A*. Thus, the identity of the agent acting does not seem to make a difference to the deontic status of the act of killing *A*.[70]

Agent-relative constraintism cannot—at least not without further argument—accommodate the *Special Prevention Case*. Recall the agent-relative constraint:

> **Agent-Relative Constraint (ARC).** (*x*) (*x* should ensure that [*x* does not *φ* even to prevent more further *φ*-ings]).

The constraint *ARC* only tells the agent to ensure that *she* does not kill even to prevent more further killings. It does not tell her to ensure that anyone

---

[69] I will say more about the enforcement dimension of (constraining) rights when I address Mike Otsuka's version of the saveability challenge in Chapter 4.

[70] Note that my duty not to commit minimising violations might still be more stringent than my duty to prevent them. Possibly, I ought to bear a high cost (such as losing an arm) rather than killing *A* in the *Special Case* but I do not have to bear a comparable high cost to prevent someone else from killing *A* in the *Special Prevention Case*. However, this does not render the duty not to commit minimising violations agent-relative. Within an agent-neutral framework there is space for the idea that duties not to do harm are generally more stringent than duties to prevent harm.

else does not kill *A*.[71] What we need in order to be able to add the *Special Prevention Case* to the extension of deontic constraints is the constraint *ANC*:

> Agent-Neutral Constraint (ARC). (*x*) (*x* should ensure that [(*y*) (*y* does not *φ* even to prevent more further *φ*-ings)]).

Since the agent-variable "*y*" may refer either to the agent herself or to any other person, *ANC* holds that each agent should not kill in the *Special Case* and, moreover, should ensure that no one else kills in the *Special Prevention Case*. On a view that aims to accommodate both the *Special Case* and the *Special Prevention Case*, a deontic constraint on killing is robustly agent-neutral insofar as only *ANC* secures that the constraint has the correct extension. Once we think that constraintism should accommodate the *Special Prevention Case*, the agent-neutral principle *ANC* looks like a perfectly plausible interpretation of a constraint on killing.

Moreover, Dougherty wants to reassure us that no disadvantage will arise from trading an agent-relative constraint for an agent-neutral one. Dougherty suspects that one reason why constraintists have so readily accepted the standard view is the fear that *ANC*-form principles would be "too narrow to cover the full range of cases in which a [constraintist] will judge that an agent ought not kill" and that an agent-neutral theory would not be able to offer the desired deontic verdicts "if we dreamed up ever more

---

[71] While this is true, agent-relative constraintists might have other options for accommodating the *Special Prevention Case*. Usually, a constraintist view will not be exhausted with a set of deontic constraints but will include a range of further moral principles such as those prescribing aid or rescue and those regulating the requirements concerning the preventable consequences of wrongdoing. Thus, an agent-relative constraintist could argue that besides a constraint on killing, her theory also includes a principle that requires agents to prevent others from killing. A theory that posits *ANC* therefore is not necessarily broader in scope as one that posits *ARC*. My point here is merely that once constraintists want to accommodate the *Special Prevention Case*, it begins to look like *ANC* is a perfectly plausible interpretation of a constraint on killing.

complex cases in which some killings depend on others not occurring" (Dougherty 2013: 532). To dispel such doubts, Dougherty makes two claims.

First, Dougherty claims that for any agent-relative constraint of the *ARC*-form governing only the agent's own conduct, there is an agent-neutral constraint of the *ANC*-form that governs that agent's conduct with respect to *everyone's φ*-ings. Thus, there is no agent-relative constraint that would not be covered by any *ANC*-type principle.[72]

Second, Dougherty argues that a theory fashioned out of *ANC*-type principles would be "just as broad in scope as its agent-relative cousin" (Dougherty 2013: 532). What he means here, as I understand it,[73] is that an *ARC*-based account and an *ANC*-based account would support the same first-order normative view: both views should yield the same verdicts in all relevant types of cases where a constraintist view holds that it is impermissible to *φ* even to prevent more further *φ*-ings. According to Dougherty, agent-relative and agent-neutral constraintism are *extensionally equivalent*.

If Dougherty is right about the extensional equivalence between an agent-neutral and an agent-relative version of constraintism, then constraintists really have nothing to fear. They could maintain the same first-order view, just based on a conception of morality that gives all agents shared moral aims. (I will attend to the advantages of such an agent-neutral account in Section 2.5.)

However, while I think that Dougherty proposes a plausible way of approaching an agent-neutral account of deontic constraints, I think his vision

---

[72] Dougherty does not offer a further argument to support this claim. Its plausibility seems to rest solely on the insight that we can transport the content of any *ARC*-type constraint into an *ANC*-type constraint, minus the agent-reference contained in the content-part. This is so because moving from *ARC* to *ANC* all that has changed is that "*y*" has replaced "*x*", any other content remains unchanged. Dougherty does not distinguish between special and general constraints, and it is unclear whether he thinks this also applies to special constraints. I will say more about the relevance of this distinction in Section 2.5.2.

[73] Many thanks to Tom Dougherty for clarifying this point with me in a personal correspondence.

of agent-neutral constraintism does not deliver on its promises. It is true that—as Dougherty himself puts it—we can dream up ever more complex cases that draw constraintists back into the grip of the standard view because it looks as though only an agent-relative account generates the desired verdicts for those cases. After examining this problem closer in the next Section (2.3.3), in Section 2.4 I argue that this is not the end of agent-neutral constraintism. Dougherty's account fails because it essentially rejects an agent-relative *justification* of the impermissibility of minimising violations without offering any alternative. I show how offering an alternative, *agent-neutral* justification could redeem Dougherty's promise of an account of deontological constraintism that is both robustly agent-neutral and generates the correct verdicts in all relevant types of cases.

2.3.3. The Special Case, Revisited

To begin with, Dougherty's assumption of extensional equivalence between agent-neutral and agent-relative accounts is rushed. It seems that the most obvious way to reach extensional equivalence between agent-relative and agent-neutral accounts would be to assume that agent-relative and agent-neutral *constraints* are extensionally equivalent. However, it should be easy to see that this is not the case. As we have seen, *ANC*'s extension is broader than *ARC*'s extension since *ANC* includes the *Special Prevention Case*. *ARC*, on the other hand, provides no direct guidance on what the bystander should do if you are about to kill *A* to prevent the killing of *B* and *C*.[74] Thus, *ARC* and *ANC* do not have the same scope—if, as I assumed, scope means extension here.

Thus, on any charitable interpretation, Dougherty cannot mean that agent-relative and agent-neutral deontology are extensionally equivalent *by virtue of ARC* and *ANC* being extensionally equivalent. For agent-relative

[74] Again, this is true *just considering ARC and ANC in isolation*. An agent-relative constraintist can come up with additional principles that demand that we prevent others from violating *ARC*.

constraintism to be just as broad in scope as an agent-neutral account, *ARC* alone will not do. It would have to be combined with some additional moral principle that requires agents sometimes to prevent others from killing even to prevent more killings.

Of course, this looks like a strength rather than a weakness of agent-neutral constraintism. Dougherty's argument was precisely that constraintists should want to adopt an agent-neutral account because, unlike an agent-relative account, its constraints clearly capture the thought that we should prevent others from violating those constraints. And I agree that the main advantage of an agent-neutral, as opposed to an agent-relative account of constraints, is that the former provides for the enforcement dimension of deontic constraints.

However, the lack of extensional equivalence between *ARC* and *ANC* is evident in other types of cases, too. In contrast to the *Special Prevention Case*, in these cases *ANC* turns out to be narrower in scope than *ARC*. Imagine the following scenario:

> Zealous Sidekick. Joker is using a trolley to kill five innocents. Bruce is standing on a footbridge next to a massive man. He could use the man's weight to stop the trolley by pushing him onto the tracks. Bruce's zealous sidekick, Robin, is standing on a second footbridge closer to the position of the five and next to two slightly less massive men. Bruce does not intend to use the massive man to stop the trolley. But he knows for certain that if he does not do it, Robin will use the two others to stop it. (Their body weights combined will be sufficient to stop it.) Either way, it is certain that someone will be killed to prevent the killing of the five.[75]

[75] I take inspiration for this case from a similar case in Kamm (1989): 251. Note that Kamm is using her example simply to trace the feature of constraints that they prohibit minimising violations of the very same constraints.

What should Bruce do? He can only *either* ensure that he does not kill *or* that someone else does not kill to prevent more killings, but not both. Bruce cannot ensure that *no one* kills even to prevent more killings but this, it seems, is precisely what *ANC* requires him to do. According to an *ANC*-based view, Bruce's moral situation is a tragic dilemma: he must choose between two mutually exclusive options, both of which would however make it the case that he has failed to satisfy the relevant moral principle.

However, the problem here is not that Bruce cannot satisfy *ANC*. It would be unreasonable to think that it is always within the agent's power to satisfy a given moral principle. And in those situations where it is not, the agent may not be required to do so—*ought implies can*. Rather, the problem is that it is clear what constraintists should want to say about *Zealous Sidekick*. They should want to say that Bruce should *not* kill even to prevent that Robin kills two. They should want to say this because it cannot be wrong for Robin to kill even to prevent more killings but, at the same time, be right for Bruce to do so. A deontic constraint on killing, agent-relative or agent-neutral, implies that *no one* should kill even to prevent more killings.[76] But to get to this claim, it seems, *ANC* will not do. Instead, it seems that the constraintist must tell Bruce that:

> your priority lies first and foremost with yourself. You must ensure that you do not kill innocent people. Maybe you are also required to ensure that others do not kill innocent people. However, if you have to choose between ensuring that you do not do this and ensuring that

[76] Of course, there might be a scenario in which it is wrong for Robin to prevent more killings but not for Bruce. For instance, Robin should not kill 1,000 innocents to prevent the killing of 1,001 others. But a moderate constraintist might claim that Bruce should kill one innocent to prevent Robin from killing 1,000 (to prevent the killing of 1,001 others). However, taking the *Zealous Sidekick* case as it is, allowing Bruce to kill would indeed deny the existence of a deontic constraint on killing such that it could not be wrong for Robin to minimise killings.

> someone else does not do it you must (all else being equal) choose the former. (Hammerton 2017: 324)

This seems to draw the constraintist back into the grip of the agent-relative interpretation of constraints. If she wants to claim that Bruce must not kill to prevent Robin from killing two (to prevent five further killings), then she should postulate an agent-relative constraint that governs the agents conduct primarily with regard to her own killings.

What the *Zealous Sidekick* case illustrates is that Dougherty's agent-neutral account cannot accommodate constraintist first-order commitments as effortlessly as he seems to think. Instead, Dougherty's agent-neutral account is at risk of collapsing back into an agent-relative one. In order to explain why Bruce should not kill in *Zealous Sidekick*, it seems that constraintists must hold that Bruce's primary concern should lie with his own actions. He should make sure that *he* does not kill, even if others do.

Moreover, although the *Zealous Sidekick* case has a more complex structure than the *Special Case*, it does not seem to raise any novel issue—one that would not have been present already in the *Special Case*. The question has always been: how can we justify that the agent ought not to kill where there is a trade-off between her own killings and more extensive killings by others? And the answer the constraintist seems committed to is that the agent should, first and foremost, avoid killing anyone herself.

I think the problem here is that the best interpretation of the significance that my agency seems to have in Mack's *Special Case* has always been that of a priority relation which holds between my potential killings and those potentially performed by others. Killings that I commit have a certain feature (they are committed *by me*) that killings committed by others lacks. When avoiding killings, I am to give priority to those killings which have this feature, meaning that I must ensure, first and foremost, that I do not kill before I attend to any other morally important business. After I ensure my own compliance with the prohibition against killing, it might be that I should try to ensure

that others comply as well. But again, I ought not to kill to ensure this because, for me, *my* compliance should take priority over theirs; for you, *your* compliance should take priority over others'; and so on. Dougherty denies this priority claim to the effect that he denies the very criterion which allowed agent-relative constraintism to prohibit killings in all cases in which a constraintist account disallows killing.

On the face of it, then, agent-neutral constraintism does not seem to be just as broad in scope as its agent-relative cousin. In fact, a constraint like *ANC* cannot accommodate all cases in which constraintists disallow minimising violations. As it stands, Dougherty's vision of agent-neutral constraintism does not deliver on its promises of a robustly agent-neutral account that would generate the desired verdicts in all relevant types of cases. In effect, his rejection of the central claim of agent-relative constraintism amounts to a rejection of the very feature that gave such views the force to accommodate the relevant first-order commitments.

2.4. First-Order Agent-Neutral Constraintism

As mentioned earlier, I do not think this is the end of agent-neutral constraintism. All we need to do in order to show that constraintists do not have to fall back on agent-relative considerations is to flesh out in more detail what a first-order account of agent-neutral constraintism might look like. All we need is to show that we can provide an alternative, agent-neutral justification for the claim that we should not kill even to prevent more killings, which also explains why the agent should not kill in cases like *Zealous Sidekick*. I believe that providing such a justification will redeem the promises of a robustly agent-neutral account of constraintism.

To begin with, recall the explanation as to why Mack's *Special Case* seems to commit constraintists to agent-relativity in the first place. There are only two differences between two sets of killings. First, the killing of *A* is committed by me whereas the killings of *B* and *C* are not. And second, the killing

of *B* and *C* contains the killing of an extra person. Since the second difference can hardly bring us to think that not killing *A* is the preferable choice, we are left with the first difference. If constraintists want to say that it is wrong for me to kill *A*, then they must say that this has something to do with the fact that *I* would be the agent who does the killing.

I think constraintists could easily resist this description. To begin with, they should reject the idea that what must be explained from a constraintist perspective is why the killing of *A* is *different* from the killing of *B* and *C* or why it is particularly wrong. In order to do so, constraintists could simply say something along the following lines:

'You say I need to point to some relevant difference between the killing of *A* and the killing of *B* and *C* to be able to claim that the killing of *A* is wrong. But rather, my point is that there is *no* relevant difference between the two sets of killings. It is wrong to kill *B* and *C*, and it is also wrong to kill *A*. If anything, I need to explain why it makes *no* difference that the killing of *B* and *C* contains the killing of an extra person.'

This is the first step in resisting the standard view. It shifts the focus of what constraintists need to explain. Someone who claims that killing *A* is *right* needs to point to some difference between the two sets of killings that explains why their deontic status is different. Certainly, it makes sense here to point to the second difference: the killing of *B* and *C* contains the killing of an extra person and is therefore impersonally worse that the killing of *A*. But constraintists—who think that all three killings are wrong—need not explain why it matters that some killings are committed by the agent whilst others are not. They need to explain why it does not matter that some sets of killings are greater than others. And while it is hard to see how the first kind of explanation could ever be an agent-neutral one, there is no reason to think that the second kind of explanation must refer to agent-relativity at all.

2.4.1. The Inviolability Account

Consequently, the second step in resisting the standard view must be to explain why the number of killings in the *Special Case* does not make a difference, or at least not so much that the deontic status of killing *A* changes from impermissible to permissible or required. I think constraintists could now say something along the following lines:

A, *B*, and *C* are all morally significant beings. It matters not only what others actually do to them; it also mattes what others *may* or *may not* do to them. If you may not kill *A* even to save *B* and *C*, then *A* is morally more important than *A* would otherwise be. *A* is more *inviolable*. Moreover, inviolability is a moral status. If *A* is more inviolable, then by extension so are *B* and *C*, who would each also have to be spared even to save the lives of more others. The point of constraints is that they give expression to this elevated moral importance.

I shall refer to this way of justifying constraints as the *status rationale* or the *inviolability account*. Chapters 3 and 4 will flesh out the inviolability account in detail and respond to the various challenges it faces. It is not my intention to pre-empt this work here. Rather, my intention is to illustrate how the inviolability account—*if plausible*—could bring the rejection of the standard view to a close.

First, the inviolability account gives clear guidance on what we ought to do regarding minimising violations of rights, including in the *Zealous Sidekick* case. If we begin with the status of the person who would have to be killed to prevent more killings, we have an explanation as to why Bruce should not kill in *Zealous Sidekick*. *A* is inviolable, which means that *A* may not be killed even to prevent the killings of *B* and *C*. *B* and *C* too are inviolable, which means that Robin may not kill them even to prevent five further killings. If Robin goes ahead, however, this cannot mean that *A* loses that inviolability status. On the inviolability account, a constraint on killing yields the verdict

that none of the killings in *Zealous Sidekick* (including those performed by Joker) can be permissible.

Second, the inviolability account supports a robustly agent-neutral account of deontic constraints. The inviolability account does not rely on any understanding of the prohibition against killing as relative to the agent acting. Instead, it holds that what *constrains* the particular agent's conduct is the status of others as morally important beings who may not be sacrificed in certain ways.[77] The inviolability account holds that neither Bruce nor Robin has permissible means available to prevent more killings. My impression is that this is a very plausible claim to make. Think again of the *Transplant* case where you are a surgeon who can only save five patients by harvesting the organs of a healthy sixth patient. What is going on in this case is that there are no permissible means available to you to help your five patients.

The inviolability account simply holds that things lie in a similar way in *Zealous Sidekick*. Under different circumstances, Bruce should prevent Robin's killings. But given the circumstances here, Bruce cannot permissibly prevent Robin's killings because he would have to kill to achieve this. What matters is not the importance of *Bruce's* as opposed to others' killings, but his lack of permissible means to prevent others from killing.[78] The inviolability account does not emphasise a distinction between what the agent does and

---

[77] A critic of the inviolability account might interject that this simply makes it a patient-relative rather than an agent-relative account. What matters seems to be the status *of the particular person* who may not be sacrificed. However, I think this is false. The agent is not required to give priority to the status of *her* potential victim over the moral status of the potential victims of other agents. Rather, her victim's status constitutes a moral barrier that the agent must not cross even to enforce the interests of others who have the same status. To pre-empt an analogy which I use in the next chapter: that I cannot pass through a solid wall does not mean that my physical incapability to do so nor the wall itself have any special significance compared to the physical incapability of others to pass through other equally solid walls.

[78] Of course, whether an agent has permissible means available or not to prevent others from doing wrong depends on facts about that agent. Bruce is standing on the footbridge, where *his* only means to stop the trolley would be to sacrifice an innocent person. However, this makes Bruce's reason not to kill agent-relative in only the formal sense in which all practical reasons are personalisable and thus depend on facts about the particular position the agent occupies in the world.

what others do, but a "distinction between what it is *permissible to do to a person* (that is, his status) and what *happens* to persons" (Kamm 1992: 386). Inviolability is, as Kamm says, a victim-focused, agent-neutral, and non-consequential value.

2.4.2. The Nested Structure of Constraints

The inviolability account might explain why Bruce should not kill in *Zealous Sidekick*. It holds that a constraint on killing implies, as to its *content*, that the agent ought not to kill even to prevent someone else from killing more others, where this person would kill to prevent even more killings. This might also have implications for the *structure* of constraints.

In particular, I want to suggest that we should think about the logical form of constraints as sometimes having a *nested* structure. For instance, a constraint that prohibits Bruce from killing in *Zealous Sidekick* might take the form:

> Nested Constraint. Bruce should ensure that [*he* does not kill even to ensure that {Robin does not kill even to prevent more killings}].

Bruce confronts a constraint on killing as a constraint on what *he* may do. But the constraint can be given an agent-neutral form. Since no reference to agent-relativity is needed to explain why Bruce may not kill, we can give the nested constraint the form:

> Nested Agent-Neutral Constraint (NNC). (x) (x should ensure that [(y) (y does not φ even to ensure that {(z) (z does not φ even to prevent more further φ-ings)})]).

NNC should be read as: for any agent, that agent should ensure that no one φ-s even to ensure that no one (else) φ-s even to prevent more further φ-

ings. Still, *NNC* is nothing more than a combined statement of what an agent-neutral killing-constraint would imply for each particular agent involved.

NNC is agent-neutral since there is no occurrence of "*x*" in the square brackets bound by the initial universal quantifier. As such, *NNC*, like *ANC*, may be addressed to a bystander. Imagine, for instance, that Barbara would be up on the footbridge with Bruce. Barbara is much stronger than Bruce and it will take her nothing to hold Bruce back if he decides to go ahead and shove the massive man off the footbridge. *NNC* gives clear guidance on what Barbara should do: she should ensure that Bruce does not kill even to ensure that Robin does not kill even to prevent more killings.[79]

It is reasonable to think that constraints take such a nested structure, mapping the chain of killings that prevent more killings in some given case. By adding an agent-variable for each agent involved, we can map the feature of a constraint that it prohibits *each agent* from killing even to prevent more killings, which is precisely what a constraint on killing was all about. It might look as though each agent must give priority to not committing any killings herself. And *de facto* the existence of an agent-neutral constraint on killing requires each agent not to kill, even if others do. But the reason why each of them must not kill even if others do is not that each of them should have a special concern with their own actions. Rather, the reason is that the moral status of persons renders all killings impermissible that would be performed to prevent more killings. No reference to agent-relativity is needed to explain this. On a robustly agent-neutral account of constraintism, each agent's primary concern lies *de facto* with her own killings, but not *de jure*.

---

[79] Note that constraintists are free to say that if Barbara could prevent *either* Bruce *or* Robin from killing, she should prevent the greater number of killings. As said before, constraintists are not qua constraintists committed to number-scepticism; see also Section 3.2.3.

2.4.3. Two Worries

Before turning to the final part of this chapter, in which I shall discuss the advantages of agent-neutral constraintism, I want to address two worries about such an agent-neutral account as presented above.

A first, general concern is that an agent-neutral account might simply overstrain the concept of a constraint. Following Scheffler's definition of constraints as agent-centred restrictions, constraints are usually understood as "limitations on the conduct of the individual agent" (Scheffler 1994: 81). They are negative duties not to cause harm or principles that prohibit the performance of certain types of action. Similarly, Dougherty explains, "I will use the term 'constraint' […] to denote a deontic prohibition on certain action-types like killing" (Dougherty 2013: 528 fn.).

However, it seems that the agent-neutral principle *ANC* is not covered by this definition. *ANC* entails not only a negative duty not to *φ* but moreover a positive duty to make sure that others do not *φ* even to prevent more *φ*-ings. The worry here is that calling *ANC* a constraint pushes the boundaries of the definition of constraints too far. The part of *ANC* that corresponds to what Dougherty and others in the debate call a constraint is the negative duty not to *φ* even to prevent more *φ*-ings. It is the part of the principle that is equivalent to a statement of *ANC* in which "*x*" and "*y*" refer to the same agent. And some might find it hard to see how *this* duty, the negative duty not to perform certain types of actions, by itself could be expressed in agent-neutral terms. In a nutshell, a possible objection against my account of agent-neutral constraints is that what I have shown, at best, is that there is *some* agent-neutral principle that has a constraint as one of its parts, not that that constraint itself can be given an agent-neutral form.[80]

I do not think that this objection is particularly worrisome. As it has already been noted, a specific prohibition against *φ*-ing to prevent more

[80] Thanks to Benedict Rumbold for pointing me in the direction of this worry.

further *φ*-ings is nothing more than an implication of how constraintists think of certain moral duties more generally. It is an implication of a constraintist account of the duty not to *φ* when applied to a specific type of case (where more *φ*-ings can only be prevented by fewer *φ*-ings). For instance, the claim that one must not kill to prevent more killings is an implication of a constraintist account of the general duty not to kill. It states, *Do not kill anyone (not even to prevent more killings)*.

Thus, I take it that even if we can only show that there is some general moral principle like *ANC* that constraintists can subscribe to, this would be sufficient to show that constraintists can express the duty not to kill in agent-neutral terms. For *ANC* both entails a deontic prohibition against *φ*-ing to prevent more *φ*-ings and takes an agent-neutral form.

Insofar as a constraint is usually understood as a restriction on the conduct of the particular agent it seems best understood as taking an agent-relative form. But insofar as that constraint is nothing more than an implication of an agent-neutral principle that prohibits certain kinds of killings (those which prevent more killings) that constraint can be understood in agent-neutral terms, without any reference to the idea that the agent's own killings would have any special significance when compared to the killings of others.

The second worry is that constraintists might get into trouble once they claim that we *should* prevent others from killing in the *Special Prevention Case*. Some might be sceptical about the requirement to prevent *Y* from killing *A* to prevent the killing of *B* and *C*, for if *X* prevents *Y* from killing *A*, *X* comes to play a crucial role in the coming about of the deaths of *B* and *C*. Depending on the precise circumstances of the case, it might look like *X* must choose between letting one of two sets of killings happen, and then must

choose to let the greater set of killings happen. Can this be morally acceptable?[81]

I think agent-neutral constraintists need to take a stance on this issue. However, in order to secure the agent-neutrality of their view it seems that they could simply avoid committing themselves to the claim that in *each instance* of the *Special Prevention Case*, *X* should prevent *Y* from killing *A*. All they need is the idea that if there are times when *X* should not kill even to prevent more killings, then there are also times where *X* should prevent *someone else* from doing this. There might still be times when, *all things considered*, *X* should *not* prevent minimising violations. By way of illustration, imagine the following scene:

> Hesitant Rescuer. Joker has dislodged a boulder that is rolling towards five innocents at the bottom of a hill. It will crush and kill them unless Bruce pushes a massive sixth innocent into the pathway to bring the boulder to a stop. Bruce is hesitant but might decide to do it. Barbara is standing close by watching the scene.[82]

What should Barbara do? In case Bruce goes ahead and tries to push the massive man, should she hold him back?

Dougherty says that constraintists should want Barbara to prevent the killing of the one. But now suppose the story takes the following turn:

> Decisive Rescuer. Bruce looked more hesitant than he really was. Without wasting another breath, he pushed the massive man into the pathway. Barbara is standing on the opposite side. If she acts quickly, she can grab the arm of Bruce's victim and pull him out of the way before he's crushed by the boulder.

---

[81] Thanks to Daniel Elstein for bringing up this problem.

[82] This is my variation on a case used by Otsuka (2011): 40.

Should Barbara do it? The *Decisive Rescuer* case illustrates that by preventing one killing, Barbara comes to play a casual role in the coming about of five other killings—a role she plays also in *Hesitant Rescuer*, but one that seems more crucial in *Decisive Rescuer*. In the latter case, Bruce's victim already occupies the physical space where he will be hit by the boulder, which renders Joker's five victims unthreatened in the present moment. Is it plausible to say that Barbara should pull the massive man out of the way, thereby putting the five back in danger?

In principle, I think constraintists have different options here. They could claim that Barbara should prevent Bruce from killing in both cases. Or they could claim that Barbara should hold Bruce back, but that she should let things go their way if, before she can intervene, the massive man is already situated in the boulder's path. What constraintists say here will depend on their views regarding the morally significant differences, or the lack of such differences, between the two cases. For instance, a constraintist might say that in *Decisive Rescuer* but not in *Hesitant Rescuer*, the five are unthreatened in the present moment and that it would be a moral failure on Barbara's part to put them back in danger.[83] Another constraintist might say that by leaving the massive man in the path of the boulder, Barbara would impermissibly use him for the sake of the five just as she would do if she pushed him herself.

I think whatever constraintists choose to say here has no bearing on the issue of agent-neutrality. Recall that what an agent should do is a function of the balance of her reasons—what she *ought* to do is what she has *most reason* to do. Constraintists do not have to maintain that Barbara has most reason to prevent the killing of the massive man even in *Hesitant Rescuer*. All

---

[83] One might think that whether Barbara should save the massive man or not cannot depend on whether his murderer is hesitant or decisive and that there might not be a morally relevant difference between these two cases. However, I do not think it is absurd to hold that it is a morally relevant piece of information that at t_1 the victim is standing close to Bruce whereas at t_2 he already occupies the physical space between the boulder and the five others. Some pieces of information, although seemingly morally irrelevant, may change what the agent ought to do at different times. Lang and Lawlor (2013) make a similar point in the context of rescue cases.

they need to say is that the existence of a constraint on killing gives Barbara reasons to prevent others from violating that constraint which contribute to the balance of her reasons. Sometimes, she might have most reason to prevent others from violating that constraint. Other times, her contributory reasons might come down on the side of not preventing that violation. But the fact that an agent may be in such a position that she has reasons to do different things and that the balance of her reasons might be as such that she ought *not* to prevent a constraint-violation does not threaten the agent-neutrality of that constraint.

On an agent-neutral constraintist view, all acts of killing in *Decisive Rescuer* and in *Hesitant Rescuer* have the deontic status of being impermissible, and this does not depend in either case on the perspective of any particular agent. Thus, even if agent-neutral constraintists do not say that we should always prevent others from violating a constraint, it is still true that, on their view, everyone should be opposed to an act of killing that prevents more killings, whether they are the agent of that act or not.

2.5. The Perks of Agent-Neutral Constraintism

Why should we prefer an agent-neutral account of deontic constraints to an agent-relative one? Even if constraintists do not need to refer to agent-relativity in order to make sense of deontic constraints, this does not show that they *should* renounce an agent-relative account of constraints or favour an agent-neutral one.

2.5.1. Dirty Hands and the Value Paradox

Perhaps the most common type of objection against agent-relative views in ethics is that they are guilty of a kind of moral egotism. For instance, it has often been argued that agent-relative constraintism would simply require

agents to avoid dirtying their hands with regard to some value. Dougherty voices the dirty hands objection as follows:

> It seems that the agent's only objection to minimizing the number of killings is that it would be *her* dirtying her hands with the business of killing. It is hard to believe that this could be an important moral reason, let alone one that is so important that it requires increasing the number of people who die. (Dougherty 2013: 531–532)

A related worry is that there is a *narcissistic flavour* to agent-relative views. They seem to require us to have, as Moore put it, "a narcissistic preoccupation with your own 'virtue'", by which he means "the 'virtue' you could have if the world were ideal and did not present you with such awful choices" (Moore 1997: 720). Despite their popularity, however, it is not obvious that these objections place agent-relative constraintists in any difficult position. Here is why.

Compare two moral intuitions about the objectionableness of killing in the *Special Case*, call them *Permission* and *Obligation*:

> Permission. Killing is so horrible that it should be permissible to choose not to kill *A* even to prevent the killing of *B* and *C*.
>
> Obligation. Killing is so horrible that it should be impermissible to kill *A* even to prevent the killing of *B* and *C*.

Constraintists are committed to *Obligation*: they think that minimising violations are impermissible. If they advocated *Permission*, i.e., the weaker claim that minimising violations are *not required*, they would not believe that there is a deontic constraint on killing. Instead, they would think that there is an agent-centred permission not to kill. (As we have seen, an agent-centred permission exists where morality encourages us to act in certain ways but leaves

it to the agent to decide whether she is willing to act in this way because doing so would entail a personal sacrifice of some relevant kind.)

However, by stating *Obligation* rather than *Permission*, constraintists reject the idea that the *Special Case* can be considered a case of dirty hands. For it is only possible for an agent to dirty her hands in the *Special Case* if two conditions hold. First, it must be *somehow justifiable* that the agent kills *A* to prevent the killing of *B* and *C*. Second, the requirement not to kill *A* must not lose its normative force, such that if she killed *A* this would result in the "moral remainder" (Nick 2019: 926) of her dirty hands. Someone who states *Permission* might be an appropriate addressee for the dirty hands objection: "you [and the likes of you] would rather that an extra person die than that your hands be dirty" (Hare 2013: 90). By stating *Obligation* rather than *Permission*, however, constraintists reject the first condition. They deny that it is justifiable to kill *A* in the *Special Case*. But if it is not justifiable to kill *A*, then there are no dirty hands to have. For getting one's hands dirty cannot just mean acting in the morally wrong way.

The problem does not go away once we replace the talk about dirty hands with Moore's talk about a narcissistic preoccupation. An agent who thinks that she cannot be required to kill might be preoccupied with her own virtue. But if it is *strictly impermissible* to kill even to prevent more killings, then the agent's so-called narcissistic preoccupation would simply consist in her lack of willingness to act in the morally wrong way. Both the dirty hands objection and Moore's narcissism objection presuppose a moral view according to which, *first*, killing someone to prevent more further killings is somehow justifiable, and *second*, agents may permissibly refrain from doing so for personal objections. Constraintism, even agent-relatively construed, is not a

view of this kind. Constraintists do not argue that deontic constraints would ground in any such personal objections.[84]

It is a mistake, then, to think that agent-relative constraintism is problematic because it invites a kind of moral egotism. As I see it, moral philosophers who state *Obligation* and explain this by reference to agent-relativity face a very different problem. (I say moral philosophers because I think that this problem concerns not only agent-relative *deontological* constraintists but all agent-relative constraintists. More on this in Chapter 5.) In Section 1.6.3, I have introduced this problem as the value paradox: Why should we think that my primary concern should lie with my own acts of killing such that it is impermissible for me to kill even to prevent you from killing more others? How can *my* agency be so significant, if compared to the significance of what would happen to *your* victims?

To be sure, the problem here is not that my *doing* harm is more significant than my *allowing* harm. This is a necessary claim to make for anyone who wants to justify constraints. The problem is that agent-relative accounts of constraints hold that what matters is that the harm done would be done *by me* whereas the harm allowed to be done would be done *by you*. This way, such accounts fail to conceptualise the basic moral commitments expressed by constraints as shared moral endeavours. We would normally think that not killing and not torturing other people are shared moral concerns; concerns about the physical or psychological integrity of persons who matter morally in their own right. How can it be that this concern is relative to each agent in

---

[84] Nagel (1972) has made a similar argument. Nagel's point is that it is unreasonable to ground a prohibition on killing in a concern for clean hands. For if the agent would dirty her hands by killing *A* in the *Special Case*, this could only be because there is *already something wrong* with killing *A*. You can only dirty your hands, so to speak, if you are about to dig in the dirt. Nagel takes this to show that the dirty hands objection rests upon a misunderstanding of the relevant type of moral views (which he calls 'moral absolutism'). While I agree with this conclusion, I think my argument here goes one step further: the mere fact that a moral theorist states *Obligation* rather than *Permission* shows that she is no appropriate addressee of the dirty hands objection.

that it becomes *more important* morally that the particular agent does not kill or torture anyone than that the moral patients are not killed or tortured?[85]

In this context, an agent-neutral account of constraintism indicates significant progress. As Dougherty argues, such an account allows constraintists to maintain that all agents should "be united in their view of a particular action, regardless of whether they are the author of the action or not" (Dougherty 2013: 531). The appeal of an agent-neutral account of constraints is that it "requires us to share a moral vision" and to "form a unified moral community, in which we all have the same goals" (Dougherty 2013: 531).

One might be sceptical about Dougherty's way of expression here. For instance, what does it mean to *share a moral vision*? I take it that all he means to say is that, on an agent-neutral account of constraintism the aims of our action would be shared *substantive* aims. For instance, everyone would have the substantive moral aim that no one commits a killing even to prevent more killings, *regardless* of the identity of the agent who would do the killing.

2.5.2. General and Special Constraints

A related problem with agent-relative accounts—that is properly addressed by an agent-neutral one—is that it is hard to see how such accounts could make sense of the difference between general and special constraints.

There are certain normative domains in which what is rational or good to do varies across different agents such that the aims of their actions cannot be described without indexical references to them. Whether my conduct can count as prudent, for instance, seems to depend on what *my* preferences or

[85] Of course, an agent-relative constraintists might simply deny that on her account *my* non-violation of the killing-constraint is any more important than *your* non-violation of it. But then again, why should I not commit a single killing to prevent you from committing more? The agent-relative constraintist believes that when avoiding killings, I must give priority to those killings potentially committed by me. And I think this has to mean that there is a sense in which *my* killings are, *from my perspective*, more significant or more important than yours.

interests are, and whether my choice of action satisfies those preferences or interests. There may well be a place for agent-relative concerns in the domain of morality, too. Perhaps, I should be specially concerned with keeping *my own* promises or caring for the ones *I* love. But in these cases, my moral commitment stems from the fact that I stand in a special relationship to someone else or that I have previously committed myself to doing something. Absent these facts about the particular agent, the agent's normative situation would be different.

Why should we think that it works in the same way with my *general* moral commitments—my natural duties not to kill, torture, enslave, or abuse others? In contrast to special duties, natural duties are owed to all persons simply *qua* persons (e.g., Jeske 1998). Special ties are absent or irrelevant to natural duties. The identity of the agent acting makes a difference to the existence or non-existence of special duties: I have a duty to keep my promise, but *you* don't have a duty to keep *my* promise. But why should the identity of the agent acting make a difference to what our natural duties are?

As we have seen, special duties can act as constraints. For instance, it might be that I ought to keep my promise even if that means that some other agent will break two of hers. But as we have also seen, the normative force of special constraints is quite different from that of general constraints. Some special constraints, such as a constraint on promise-breaking, are easily defeated by the prospect that I could prevent other kinds of harm, such as killings or tortures. I should not break my promise to prevent you from breaking two, but I *should* break my promise to prevent you from killing even a single person. The best explanation for this seems to be that I cannot keep your promises, I can only keep my own. Thus, even though there might be cases where I can ensure the outcomes of what you've promised, what matters is the *keeping of* the promise itself and no one other than you can ensure that *your* promises are kept. Reference to agent-relativity may therefore justify a constraint on breaking one's promises even to minimise the overall number

of broken promises. But the same justification fails once you could protect something of great agent-neutral value—by preventing killings or tortures, say.

To be sure, my point here is not that agent-relative constraintists cannot hold that I should not kill even if you kill more others because the duty not to kill is relative to each agent acting. Rather, my point is that an agent-neutral account of (general) constraints can track the difference between general and special constraints by holding the first kind to be agent-neutral, the second kind to be agent-relative (like the special relationships they are based on). An agent-relative constraintist understands both kinds of constraints in terms of agent-relativity and is therefore not free to draw the distinction between general and special constraints in this way. How could she account for the different normative forces of special and general constraints?[86]

A defender of the agent-relative view might argue that she can distinguish between special and general constraints simply by referring to the fact that special constraints arise from voluntary commitments whereas general constraints do not. I might face a constraint on promise-breaking only because I voluntarily made that promise. But a general constraint on killing lacks this feature. However, this argument doesn't reach very far since there are special constraints that do not arise from voluntary commitment. For instance, I should not abandon my infant child even if by doing so I could prevent someone else from abandoning three other infants. Duties to care for one's children, to forgive minor wrongs, or to show gratitude are special in the sense that no one other than the agent can discharge them, and yet the

[86] Therefore, the agent-neutral account of constraints I have proposed is strictly speaking an agent-neutral account of *general* constraints. I think agent-relativity provides a perfectly plausible explanation of special constraints. I find it very plausible to say, for instance, that a reason to keep one's promise is robustly agent-relative: The normative force of such a reason cannot be captured without reference to the promisor. Accordingly, it makes perfect sense to me to say that I might have most reason not to break a promise even if by doing so I could prevent more extensive promise-breaking by others since the duty to keep one's promises is relative to each agent.

agent does not take on these duties voluntarily. The best explanation of the difference between such duties on one side and duties not to kill or torture on the other is that the first kind of duties are special whereas the latter kind of duties are general. But this difference remains untraceable on agent-relative accounts.

2.5.3. Maximising and Consequentialising

A final advantage of agent-neutral constraintism that I want to mention is that it allows constraintists to resist the idea that consequentialism provides the more suitable account of constraintism. Agent-relative constraintism, by contrast, makes constraints an easy target for consequentialising. I shall say more about the consequentialising project in Chapter 5. Here, I just want to refer to some of that material in short to show which advantage an agent-neutral account might have in this regard.

Consequentialising is the operation of translating a non-consequentialist theory into a consequentialist theory by accommodating the deontic properties of the original non-consequentialist theory within the consequentialist framework of ranking outcomes from better to worse. Some consequentialists—the *consequentialisers*—claim that this can be done for any originally non-consequentialist moral theory (e.g., Dreier 1993, Portmore 2007, Suikkanen 2009). Constraintism has been a preferred target for consequentialising since it's usually understood as a distinguishing mark of non-consequentialist ethics.

The possibility to consequentialise constraints depends on the question whether we can incorporate constraints into a maximising, teleological structure. The preferred strategy of consequentialisers is to claim that what matters morally is, at least in part, a function of what is good *relative to the agent acting* and that it is better relative-to-the-agent if *she* does not kill than that she prevents more extensive killings by others; obeying a constraint on

killing would produce the better outcome. It would maximise value, not agent-neutrally, but *agent-relatively* construed.

Thus, the possibility to consequentialise constraints presupposes an *agent-relative* understanding of constraints. By embracing the standard view, constraintists agree on the initial premise of the consequentialiser's argument. Even more, they open the doors to a further argument about the preferability of consequentialism as an account of deontic constraints.

To see why, consider how agent-relative constraintists deal with the problem of intrapersonal constraints: if I am to give priority to my own actions, I ought not to kill even to prevent many more killings committed by others. But what if I could kill *now* to prevent myself from killing more others *in the future*? As we have seen, agent-relative constraintists can either embrace the idea that minimising the number of my own killings is precisely what I should do here (Heuer 2011), or they can refer to moment-relativity in order to restrict the agent's value-maximising concern to the present moment (e.g., Johnson 2019). Either way, agent-relative constraintists do not reject the idea that constraints are properly understood as maximising rules. Instead, they seem to be committed to the claim that agents are required to produce outcomes in which the number of killings they might commit are minimised, either across time (Heuer 2011) or in the present moment (Johnson 2019).

In other words, agent-relative constraints seem to require us to focus our maximising concerns on agent-relative rather than agent-neutral values, but they are still value-maximising concerns. This way, constraintists open the doors to the following kind of argument: if constraints are best understood in terms of agent-relative maximising concerns, then the value-maximising structure of consequentialism might just provide the best overall theoretical framework for constraints (Hammerton 2020). Why should we think that deontic constraints are a distinguishing mark of non-consequentialist ethics, if their normative force is best explained on a consequentialist, value-maximising structure? So long as constraints are non-maximising principles, they may

be seen as an important determinant of deontology (Oberdiek 2008: 105) or as giving deontological views "considerable anti-consequentialist force" (Scheffler 1985: 409). However, by embracing the conventional claim that deontic constraints only make sense in agent-relative terms—as rules that require the agent to limit her value-maximising concerns to her own actions (in the present moment)—*deontological* constraintism is vulnerable to the objection that a consequentialised account of constraints provides the best *overall* account of constraintism.

Arguably, some agent-relative constraintists might not be troubled too much by the insight that constraints fit well into a consequentialist theoretical framework. Consequentialisers sometimes claim that non-consequentialists are in the grip of a deep confusion when they declare themselves *non*-consequentialist. And it seems as though constraintists would have a hard time rejecting the idea that they are confused in this way if constraints turned out not to have considerable anti-consequentialist force. But, of course, constraintists do not have to care too much about how they are classified in ethical theory. They might not even care much about the project of proving deontology to be superior to (or even only clearly distinguishable from) consequentialism.

However, I believe that consequentialised constraints that provide an argument in favour of consequentialism would in fact be a considerable defeat on the constraintist's part. This is especially the case for deontologists who find themselves drawn to an absolute version of constraintism. As I will argue in Chapter 5, consequentialised constraintism cannot avoid the value paradox because constraints as maximising rules are based on an implausible conception of value. From this perspective, agent-neutral constraintism offers the promise of a genuinely deontological account of constraints that cannot easily be consequentialised.

2.6. Conclusion

The standard view rests upon the thought that in order to make sense of a constraint on *φ*-ing, constraintists must refer to the idea that when avoiding *φ*-ings I am to give priority to my own *φ*-ings (in the present moment).

I have argued that constraintists need not accept the standard view. They can hold that a moral principle that prohibits *φ*-ings preventing more further *φ*-ings gives all agents a shared moral concern with regard to everyone's *φ*-ings. According to the view I favour here—the inviolability account—the reason why I must not *φ* even to prevent more *φ*-ings is not that I should have a special concern with my own *φ*-ings in the present moment, but that the moral status of my potential victim limits the set of permissible means available to me to prevent those other *φ*-ings.

Not everyone will be convinced that the agent-neutral account of constraintism I have presented is overall plausible. In the following chapters, I will develop the central ideas of the inviolability account more carefully and address various challenges to this account. The easiest way to resist the argument of this chapter, however, is simply to insist on a narrower notion of the agent-relative/agent-neutral distinction as the one I have developed.

Most notably, some define agent-neutral values as the values of certain types of occurrences. Nagel, for instance, calls agent-neutral values "the values of certain occurrences or states of affairs, which give everyone a reason to promote or prevent them" (Nagel 2008: 105). Can inviolability be an agent-neutral value in *this sense*? Perhaps it is possible to show that a world in which we are inviolability is—for the reason that we have that valuable status—morally more desirable than one in which we do not have that status. But since inviolability is not the kind of value one could promote through action it seems unjustified to say that this would give everyone a reason to *promote* a state of affairs in which we are inviolable. I think if agent-neutral values are defined in the narrower sense, it will prove very difficult to convince

sceptics that inviolability is an agent-neutral value that grounds agent-neutral constraints on action.

However, agent-neutral constraintists should remain unimpressed by this objection. They can either insist that their theories are agent-neutral in *some other* plausible sense, i.e., in the sense in which the value of inviolability gives all agents the same moral concern with the impermissibility of certain types of action. Or they could at this point rightly withdraw themselves from the debate. For if we choose the narrower sense of agent-neutrality it is trivial that any moral theory which does not hold that moral rightness is a function of the evaluation of outcomes must fail to exhibit the feature of agent-neutrality. The narrower sense of the agent-relative/agent-neutral distinction rests upon a bias towards consequentialist thinking, and leaves constraintists with the choice to simply reject that the distinction can be applied to their theories at all: the constraints of deontology are then *neither* agent-relative, *nor* agent-neutral in the narrower sense (Mack 1998).

3 *The Idea of Humanity, the Value of Inviolability*

3.1. Introduction

The agent-neutral approach to the paradox of deontology begins with the thought that the rationale for constraints is to be found not in the significance of *my* agency as opposed to that of other agents, but in the moral status of persons as moral patients. The purpose of this chapter is to flesh out this *status rationale* for constraints. The status rationale has its ideological roots in the Kantian idea of persons as ends-in-themselves who may not be used as mere means to other, even morally valuable ends. Already suggested as a possible solution to the paradox by Nozick, the most detailed description of the status rationale is Frances Kamm's *inviolability account*.

Kamm develops the inviolability account in several works on non-consequentialist moral theory. Her main concern in the relevant writings is not primarily with the paradox of deontology, but with developing a non-consequentialist view on several issues related to the ethics of harming. It is common, however, to treat the inviolability account as a systematic solution to the paradox and the present chapter shall do the same. That said, the chapter will first reconstruct the line of thought that leads Kamm to encounter the paradox and will look closer at how she treats the problem there, before

putting together the individual elements of the inviolability account as a systematic response to the paradox.

The core idea of the inviolability account is to justify constraints on the grounds that they give expression to our elevated moral status as inviolable beings. As such, the inviolability account rests upon a simple insight: that there is a relevant difference between the morally significant things that may *happen to* us and our moral significance *itself*; a distinction, that is, between how we *are* treated and how we *may be* treated. Furthermore, it rests upon the assumption that we can assess the value of what it means to have a certain moral status *independently* of the value of what happens to us. For instance, it is certainly a bad thing to be tortured. But it is *another kind* of bad thing to have the status of a being who *may* be tortured. That is, the value of being free from torture is distinct from the value of having the status of a being who *is wronged* by an act of torture.

Based on this insight, a two-stages argument can be formulated to show that constraintism is not paradoxical. First, constraints give expression to an elevated moral status—called *inviolability*—that would be denied by a moral theory that does not include constraints. Constraintism ceases to appear paradoxical once we understand it as a moral theory that gives priority to our moral status over the morally significant things that may happen to us. And second, constraintism makes for a better moral world, preferable to a world without constraints, because the latter is a world where we entirely lack the valuable status constraints give expression to.

I believe this is a promising approach and that once the two-staged argument is developed in sufficient detail, it can clear the air of paradox surrounding the concept of deontic constraints. To a large extent, its promise might stem from the fact that the inviolability account accommodates various intuitions about the nature of the problem at hand and the challenge of solving it. For one thing, the inviolability account incorporates the idea that the expected wrongdoing of some should not make it morally appropriate for

others to act in comparable ways. For another thing, it does not aim to explain the impermissibility of minimising violations by trying to identify a feature or a set of features that assign any *special* significance to the agents or patients of minimising violations.

Section 3.2 begins with some remarks on Nozick's application of the status rationale and on an understanding of inviolability as a moral status. Section 3.3 reconstructs the line of thought that leads Kamm to encounter the paradox. Sections 3.4 and 3.5 put together the individual elements of the inviolability account as a systematic response to the paradox and lay out the two-staged argument to justify constraints.

3.2. Ends and Persons

As we have seen, the paradox of deontology goes back to a passage in *Anarchy, State, and Utopia*, where Nozick raises a bundle of questions about the rationality of constraintist views. His own answer to the paradox is not long in coming. Ensuing the bundle of questions that set up the issue, Nozick continues to claim that:

> constraints upon action reflect the underlying Kantian principle that individuals are ends and not merely means; they may not be sacrificed or used for the achieving of other ends without their consent. Individuals are inviolable. (Nozick 1974: 30–31)

At a first glance, it appears that Nozick refers to not just one but *two* features of persons in the attempt to justify constraints: persons are *ends* and not merely means, and they are *inviolable*. Unfortunately, he omits to explain

how these two features are connected, or which roles they play respectively in justifying constraintism. Thus, a closer look at these two ideas is in order.

To begin with, the Kantian principle Nozick refers to is the second version of the categorical imperative, the *humanity formula* or *idea of humanity*, which says:

> So act that you use humanity, whether in your own person or in the person of any other, always at the same time as an end, never merely as a means. (Kant 1998: 38)

Nozick seems to believe that the humanity formula provides a straightforward explanation as to why we must not commit minimising violations. Recall Mack's *Special Case* where I could kill *A* to prevent the killing of *B* and *C*. If I kill *A*, I would be using *A* as a mere means to save *B* and *C* and thus would fail to treat *A* properly as an end-in-itself. It seems, then, that when Nozick calls *A* inviolable, all he means is that *A* is an end-in-itself. To be an end-in-itself is *to be* inviolable. And to be inviolable means that one may not be treated as a mere means to some other end.

However, being inviolable cannot simply be identical with the property of being an end-in-itself. To see why, consider a commentary by Frances Kamm. According to Kamm, being an end-in-itself entails:

> that one is not to be treated as if one's existence were 'for' the goal of optimizing good. Rather one exists as an *end-in-itself*, even if one does not always serve the greatest good when it could be served. (Kamm 2001: 229)

This passage is insightful because it can be read in two ways. First, if we do not exist solely for the goal of optimising the good then we may sometimes, in our role as moral agents, permissibly act in ways that do *not* optimise the good. Sometimes, we may permissibly choose not to produce the best

available outcomes. In other words, as ends-in-themselves moral agents are not "mere tools" (Kamm 2001: 262) liable to produce the greatest possible good. In this regard, the humanity formula leads to the idea that there are agent-centred permissions not to optimise overall good.

Second, if we do not exist solely for the goal of optimising the good this also means that we must not, in our role as moral patients, be used as mere means for achieving the best available outcomes. As ends-in-themselves persons may not be used as mere means to other, even morally desirable ends. They may not be *violated* in certain ways, even in the pursue of other morally important ends. Inviolability, then, is better understood as an explication of what it means to be an end-in-itself *considering one's role as a moral patient*.

The idea of inviolability, then, does not exhaust the idea of humanity. As we have seen, inviolability is just one among other features of persons as ends-in-themselves. For instance, being an end-in-itself also means something regarding one's role as a moral agent. Moreover, inviolability might not even exhaust what it means to be an end-in-itself *considering one's role as a moral patient*. It is not obvious that being an end-in-itself only says something about the circumstances in which one may not be harmed and nothing more broadly about any other kind of circumstance in which one may not be used.[87]

3.2.1. Inviolability as a Moral Status

Like being an end-in-itself, being inviolable cannot mean that one is in a certain *condition*. Suppose that I am free—i.e., with certain natural limitations I can move wherever I want. This is a condition that applies to me now. But

[87] I will not go into further detail to analyse the idea of what it is to treat someone merely as a means. Some have argued that, despite its being one of the most influential thoughts of Kantian ethics, it is impossible to get a clear grip on its meaning. Most recently, Kleingeld (2020) has defended the Kantian idea of humanity. Instead of going into this debate in further detail, I shall henceforth limit my focus on the idea of inviolability.

one could come around my house, restrain me or lock me in a room, and I would not be free anymore. I would have *lost* my freedom. This is so because conditions apply to a person at certain times and may not apply to the same person at other times. But if I am inviolable, no one can come around my house and make it the case that I am not inviolable anymore. No one can *make me* violable. Being inviolable, as Nagel says, "is not a *condition*, like being happy or free", neither does it mean "that one *will not be violated*" (Nagel 2008: 107). Inviolability defines how a person *should be* treated, not how they *are* or *are not* treated. That is, inviolability is not a condition, it is a *moral status*.

The distinction between condition and status is often described by phrases of the following kind: I might no longer be free once I am restrained, but as a person I am *unrestrainable*. I might get tortured, but I am *untorturable*. And so on. Of course, the suffix *-able* is not to be misread as indicating physical impossibility. If I am inviolable, this does not mean that it is physically impossible to violate me. Instead, the suffix *-able* here indicates *impermissibility*, i.e., morality's notion of impossibility, if you will.

Kamm distinguishes between a broad and a narrow sense of moral status (Kamm 2007: 227–228). In the broad sense, moral status is simply a description of the deontic properties of ways of treating some entity *E*. That means, *E*'s moral status is a description of what it is permissible, impermissible, required, supererogatory, etc. to do, *considering E*.[88] This, of course, would mean that not only persons, but also snails and rocks are appropriate

[88] On a minor terminological note: Kamm speaks of what it is impermissible, permissible, etc. to *do to E*. I prefer to say "to do, *considering E*" as not everything we do that is captured by an individual's moral status is something we do *to* them. For instance, it might be permissible for me not to help a stranger who is struggling to carry a piano on the street, if I am in a hurry to get to a meeting. But not helping the stranger is nothing I do (or do not do) *to* them, in the ordinary meaning of doing something *to* someone. Passing by without offering help is something I *may* do, even considering the stranger's moral status as someone who should be helped under other circumstances. If the stranger collapsed under the weight of the piano and would be at risk of suffocating, for instance, it would *not* be permissible for me to just pass by because I must consider the stranger's status as someone who must be saved in such an emergency.

entities to which we can attribute moral status. Arguably, rocks might have the moral status of entities to which it is permissible to do anything. It is not morally wrong, presumably, to remove a rock from its original place, throw it across the lake making it bounce off the surface of the water, or hit it hard against another rock to start a campfire. But even in this case, since there are things that we may do considering the rock, the rock has moral status in the broad sense (Kamm 2007: 227).

In the narrow sense, moral status is possessed by an entity if and only if that entity in some relevant sense counts morally in its own right (Kamm 2007: 227–228). For instance, an entity has moral status in the narrow sense only if its suffering is at least somewhat morally bad, not because *someone else* has an interest in its non-suffering, but for its own sake (Jaworska and Tannenbaum 2021). In the narrow sense—given a certain common-sense understanding of moral status—rocks are no appropriate entities to which we can attribute moral status. But persons, (certain kinds of) animals, and intelligent extra-terrestrial beings have moral status also in the narrow sense.

Inviolability, then, is a moral status insofar as it is a description of what it is impermissible to do considering some entity *E*. Two terminological junctions open up. At the first junction, inviolability can be understood either as *a* moral status or as *a dimension of* moral status at large. In any case, inviolability does not exhaust the concept of moral status. There are other things that define what is permissible, required, supererogatory, etc. to do considering *E*. Thus, the things that are impermissible to do considering *E* are either only one moral status among many moral statuses possessed by *E*, or they are only one dimension of *E*'s moral status at large. (Quite possibly, *E*'s inviolability does not even exhaust the things that are impermissible to do considering *E*.[89]) To avoid confusions, I shall refer to moral status at large as *moral*

[89] For instance, think again of the stranger struggling to move a piano. It might be impermissible to pass by without saving the stranger from suffocating under the weight of the piano. But it is not obvious that this impermissibility falls within the notion of inviolability.

standing. Thus, according to the terminology I choose, *E*'s moral standing is the set of all of *E*'s moral statuses including *E*'s inviolability.

At the second junction, inviolability can be understood either as a *quantitative* or as a *qualitative* concept. As a quantitative concept, inviolability defines the extent to which we may not violate an entity or treat it in harmful ways (Kamm 2007: 26). This means that entities can be *more* or *less* inviolable. The larger the set of circumstances under which some entity *E* may not be harmed, the greater is *E*'s inviolability.

As a qualitative concept, inviolability defines a quality that entities either have or don't have. I think it is a significant source of confusion in the debate about the inviolability account that no clear distinction is being made between the quantitative and qualitative uses of the concept. Kamm, too, sometimes talks about the ways in which we could be *more* or *less* inviolable, and then speaks of the inviolability that we have or don't have. To avoid such confusions right from the start, I want to introduce the term *hyperinviolability* to refer to the qualitative concept of inviolability.[90] Accordingly, an entity *E* is hyperinviolable if and only if it is impermissible to harm *E* in certain ways even to protect a greater number of other entities of *E*'s kind from the same type of harm.

There are some open questions about the notion of hyperinviolability. Most importantly, is hyperinviolability simply the name for a certain degree of elevated inviolability or does the attribution of hyperinviolability status depend on any further conditions? In Section 4.2.2, I develop a deeper understanding of hyperinviolability according to which that status is reserved to

---

Rather it seems to fall under another moral status possessed by the stranger, which could be called *saveability*; see Chapter 4.

[90] Lippert-Rasmussen introduces another concept to refer to the status "we gain as a result of (and only as a result of) the impermissibility of minimizing violations", which he calls *independence* status (Lippert-Rasmussen 1996: 345). However, he defines the independence status of persons to include, besides the impermissibility to *do* harm even to prevent greater harm of the same type also the impermissibility to *countenance* harm.

those entities who possess the capacity to be wronged.[91] Chapter 4 will also analyse the problems that arise from a mere quantitative notion of inviolability and show how an account based on the concept of hyperinviolability—the *hyperinviolability account*—helps to overcome these problems.[92] However, this chapter focuses on a reconstruction of Kamm's inviolability account. For the moment, I thus confine myself with having a clear distinction between inviolability and hyperinviolability, which will help a better understanding of Kamm's account.

On a final note, as we have seen, inviolability is a concept that can be applied not only to persons but to non-persons as well. It seems that we can speak of the inviolability—i.e., the *degree of* inviolability—of persons just as we can speak of the inviolability of animals or plants (Ross 2016: 71), and perhaps even the inviolability of inanimate objects (Kamm 2007: 228). Kamm suggests that even the moral status expressed by constraints, which I have called hyperinviolability, might not be limited to persons. We think of symbolic entities like flags or items of religious value as inviolable to that extent if we think that they should not be destroyed, for instance, even to save other objects of their type from being destroyed (Kamm 2007: 256). (Whether such entities deserve to be called hyperinviolable depends, as already noted, on a further question about the conditions for hyperinviolability status. In Section 4.2, I will argue that hyperinviolability is the status of entities who *can be wronged*, which precludes non-persons from the realm of hyperinviolable entities.)

---

[91] Another open question is whether inviolability and hyperinviolability are number-sensitive concepts, i.e., if our (hyper)inviolability increases further if the set of violations which may not be prevented increases. I say more about this in Section 4.4.2 (see footnote 130, in particular).

[92] It should be noted that although Kamm often speaks as if taking inviolability to be a mere quantitative concept, she acknowledges early on that "simply talk about inviolability [and here she means talk about inviolability as a quantitative measure of moral worth] cannot be all we need to explain the presence of a constraint" (Kamm 1992: 384). I say more about this in Section 4.2.

3.2.2. Are We Hyperinviolable?

To say that there is a certain valuable status is of course not enough to establish that we *have* that status. Why should we think that we *are* inviolable to the extent that we are protected by constraints only because it would be *good* if we were inviolable to that extent? Kamm's account is often taken to entail the following strange kind of argument: it is better to be hyperinviolable than to lack that status. Therefore, we *are* hyperinviolable.

On the face of it, this looks like a strange kind of argument for two reasons. First, that constraints give expression to our hyperinviolability is a terminological necessity. Kamm essentially defines constraining rights and our status as hyperinviolable beings as "two sides of the same coin" (Burri 2017: 623). To be protected by constraining rights *just means* to have hyperinviolability status. This is true on mere terminological grounds. Thus, the first premise of the argument is a mere definitory commitment, but one on which much of the remaining argument seems to depend.

A second, even stranger feature of the above argument is that it tries to infer the truth of the proposition that we *are* hyperinviolable from the truth of the proposition that it would be *better* if we were hyperinviolable. The argument is an instance of the inference: *It would be better if p, therefore p*. I shall refer to this type of inference as *the better world argument*.

Section 3.5 looks closer at the better world argument and asks how we could make sense of it. I will propose a new interpretation of the better world argument that shows that it can be given the form of a valid argument. But more importantly, the section asks why it should matter at all whether we are in fact hyperinviolable. This is an important question to ask. For the paradox of deontology is not identical to the problem of how we could prove *that there are constraints*. Instead, the paradox asks how we could prove that

the concept of constraints *is not paradoxical*. As I will argue, we can provide the second kind of proof without having to provide the first kind of proof.

The difference here might be a subtle one. But it is an important matter that these are two distinct questions, "Is it true that *x* is *F*?" and "Is it paradoxical to think that *x* is (or could be) *F*?" By way of illustration, a literature theorist might argue that it makes sense to understand the text as a room, i.e., to understand the text in topological terms. She might try to show that this view is an *internally coherent* one. That does—hopefully—not mean that she thinks that the text *is* a room. In Section 3.5, I aim to rephrase the better world argument on the first-order level, as an argument that does not commit us to any metaethical assumptions about the nature of moral truths.

3.2.3. The Separateness of Persons

Before looking at Kamm's account of constraintism in more detail, I should refer to one popular suggestion on how we could show that we *are* in fact hyperinviolable—i.e., by reference to the idea of the *separateness of persons*.

Nozick anticipates the following objection to his inviolability argument: Most of us sometimes undergo a sacrifice for our own greater benefit or to avoid greater harm. For example, we go to the dentist to avoid worse suffering later (Nozick 1974: 32). Call this the *sacrifice-to-benefit argument*. Could we not, on similar grounds, justify the sacrifice of one person for the greater benefit of two or more others? If it is important that we don't use each other merely as means, then why should we not see to it that as few of us as possible are being used in this way?

Nozick's response to the sacrifice-to-benefit argument makes use of a common piece of non-consequentialist thought. Constraints, he says, "reflect the fact of our separate existences", i.e., the fact that "there is no moral

outweighing of one of our lives by others so as to lead to a greater *social* good" (Nozick 1974: 33).

What Nozick seems to be after here is the idea that there is *no such thing* as a greater social good, nothing that would conglomerate from the goodness of what happens to separate persons. Thus, if I save *B* and *C* from being killed by killing *A*, then the good of *B*'s not being killed and *C*'s not being killed do not make up some combined *greater* good that would outweigh the badness of *A*'s being killed. Since *A*, *B*, and *C* are all separate persons, there is no social entity consisting of *A*, *B*, and *C* that would undergo some sacrifice by having *A* killed, for its own greater benefit of having *B* and *C* saved. Since *A* is a separate person, and the life sacrificed is the only life *A* has, *A* does "not get some overbalancing good from his sacrifice" (Nozick 1974: 32–33). The sacrifice-to-benefit argument—a good argument for any of us to go to the dentist—therefore has no weight on the level of the social or moral good. It has no weight where sacrifice and benefit are not combined in the same person.

What Nozick puts forward here is a version of a general argument against utilitarianism that is perhaps most familiar from Rawls' *Theory of Justice*. Famously, Rawls accuses classical utilitarianism of not taking seriously the "distinction of persons, […] the separateness of life and experience" (Rawls 1999: 167). By adding up everyone's happiness, classical utilitarianism would neglect the fact that we lead separate lives. Famously, Taurek (1977) has made the theme of the separateness of persons the basis for his rejection of the idea that the numbers are a significant factor in rescue cases where we can only save one of two groups of unequal size.

I will have something to say about the (in)significance of numbers in the context of a theory about moral status in Chapter 4. But I will generally confine myself to the justification of constraintism. It should be clear that constraintists are not *qua* constraintists committed to *number-scepticism*, i.e., the radical view that the numbers have no moral significance whatsoever.

Constraintist might still believe that we should save the greater number in rescue cases. Moreover, moderate constraintists even believe that the numbers count in paradox cases such that there is a threshold of rights violations I could prevent at which the constraint on violating someone's right gives way.

However, according to Nozick, we could say that we *are* hyperinviolable because we are, as a matter of fact, distinct persons. His argument ultimately rests on the claim that it is *not* worse if two are killed than if one is killed. This is so because there is no social entity that would combine all the lives at stake in one such that it would be adequate to say that two killings are worse—worse *for whom*, we may ask—than one.

However, one does not have to be a utilitarian of rights to reject this justification. *A*, *B*, and *C* may be separate persons; but the separate moral concerns for each of their separate lives—the separate moral reasons for saving them—combine in the moral agent who must decide what to do, considering the lives of everyone involved. Taking the needs of each person seriously and accepting that I have two independent reasons to save *B* and *C* and only one reason to spare *A* seems to be at least one natural way of thinking about my choice in the *Special Case*.

Moreover, even if we accept Nozick's claim that we do not actually promote the greater good by killing *A* because there is no such *greater* good, his answer to the paradox is begging the question. We still lack an argument as to why it would be *impermissible* to resolve the *Special Case* by deciding to save *B* and *C*. *A*, *B*, and *C* are separate persons. But why not saving the lives of two separate persons instead of sparing the live of one? At this point, Nozick's proposed rationale for constraints runs the risk of becoming circular. It seems that the only available answer to the question why it is impermissible to kill *A* is that *A* is hyperinviolable. Nozick's justification of the claim that we are hyperinviolable referred to the idea that *A*, *B*, and *C* are separate persons.

This idea led us to doubt that it could justify a constraint on killing *A* leading back to the initial question; and so forth.

3.3. A Principle of Permissible Harm

As already noted, Kamm suggests a response to the paradox of deontology in the context of developing her non-consequentialist view on the ethics of harming. She first encounters a variant of the paradox when defending a moral principle which she believes captures common moral intuitions about cases in which causing harm is necessary to produce some greater good. She calls it the *principle of permissible ham*:

> **Principle of Permissible Harm (PPH).** It is permissible to cause harm to some in the course of achieving the greater good of saving a greater number of others from comparable harm, if events which produce the greater good are not more intimately causally related to the production of harm than they are to the production of the greater good [...]. (Kamm 1989: 232)

Among other things, Kamm hopes that the *PPH* can help us to answer the trolley problem.[93]

As a reminder, the trolley problem stems from the fact that common-sense morality yields divergent judgements about the following two types of cases:

---

[93] As such, one might wonder what distinguishes the *PPH* from the *principle of double effect* (*PDE*), which states, roughly, that it is permissible to cause harm as an unintended side effect of bringing about a greater good. The *PPH* shifts the focus from the agent's state of mind (her intentions) to the causal role of her actions which results in the two principles having different extensions, permitting or forbidding different kinds of action (Kamm 1989: 245).

> **Switch.** You could save five strangers by pulling a switch that diverts a runaway trolley to a side-track. However, a sixth stranger occupying the side-track would die.
>
> **Footbridge.** You could save five strangers by pushing a massive sixth stranger off a footbridge in front of a runaway trolley; the massive man would die.[94]

Common-sense says that you may cause the death of the one in *Switch*, but not in *Footbridge*.[95] But it seems worth questioning why our intuitions diverge this much in two cases where, on the face of it, you could kill one to save five.

The *PPH* is meant to explain these divergent intuitions. Pulling the switch seems to be at least as intimately causally related to the saving of the five as it is related to the killing of the one. Thus, you may pull the switch. But pushing the massive man off the footbridge seems to be much more intimately causally related to his killing than it is to the saving of the five. That the five are saved in *Footbridge* is an effect further down the causal line, so to speak. The massive man needs to be killed first before his body can stop the trolley and this, eventually, prevents five more deaths. Thus, you may not kill him.

To illustrate a third application of the *PPH* consider another trolley case:

> **Loop.** You could save five strangers by pulling a switch that diverts a runaway trolley to a side-track, occupied by a massive sixth stranger; the massive man would die. The side-track reconnects to the main

---

[94] Both cases have been introduced by Thomson (1985): 1397, 1409.

[95] This is backed by empirical research. Interview studies have shown that around 90% of respondents confronted with these two scenarios agree on this evaluation (Hauser 2006: 139). Of course, such results are to be treated with caution. For something to be morally right it needs more than for it to accord with the majority opinion.

> track such that the trolley would still overrun the five, if the massive man's weight wouldn't stop it.[96]

Pulling the switch in *Loop* is only effective because you use the sixth person's body to slow the trolley down and thus, seems more intimately causally related to the killing of the massive man than to the saving of the five. Thus, the *PPH* forbids pulling the switch in the *Loop* case (Kamm 2007: 25).

According to Kamm, the notion of an event or action being more or less *intimately causally related to* something is an explication of what it means to achieve something *by* doing something else. More precisely, it should help us to spell out the different senses of the *by*-relation (Kamm 1989: 243).[97] In *Footbridge*, you would cause the death of one *by* pushing the man off the footbridge in a stronger sense than you would be producing the greater good *by* pushing him. In *Switch*, on the other hand, you would produce the greater good *by* pulling the switch in a stronger sense than you would be causing the death of one *by* pulling it.

Most importantly, Kamm's *PPH* entails that there are deontic constraints on action. In paradox cases, it is almost always true that the agent's interference is more intimately causally related to the harming of the few than it is to the saving of the many.[98] For instance, recall the paradox-version

---

[96] Also this case was introduced by Thomson (1985): 1402.

[97] Thus, Kamm agrees with Thomson, who also argued that when judging these kinds of cases, we "should be attending to the means by which they [the agents]" harm some and save others (Thomson 1985: 1407). See also Woollard (2008).

[98] What about a paradox-version of the *Switch* case? Suppose again that the trolley has been send on its way by a villain trying to kill five innocents. It seems that if this is the only difference between *Switch* and *Switch Paradox*, then Kamm's *PPH* should yield the same verdict for both cases: Pulling the switch is at least as intimately causally related to the saving of the five than it is to the killing of the one. Thus, you may pull the switch in both cases. A proponent of *PPH* has two options now: She can either argue that *Switch Paradox* is not actually a paradox case because your killing and each of the villain's killings aren't equally significant. Yours is an *indirect* killing whereas the villain's killings are *direct* killings or intentional murders; or she could allow for *Switch Paradox* to be a paradox case and then embrace the idea that the *PPH* generates constraints on action in only some, but not in all paradox cases. I think the first option is the preferable one.

of the *Footbridge* case—*Footbridge Paradox*—where the trolley is not out-of-control but has been brought on its way by a villain trying to kill five innocents. In *Footbridge Paradox*, just like in *Footbridge*, pushing the massive man off the bridge is more intimately causally related to the coming about of his death than it is to the saving of the five. Thus, the *PPH* forbids that you kill the massive man.

Or recall *Bomb Paradox*—the intrapersonal paradox case—where pushing a stranger onto it is the only way to prevent that your bomb kills five others. Pushing the sixth person onto the bomb is more intimately causally related to the coming about of his death than to the saving of the five who are situated. Thus, it is impermissible according to the *PPH* to push the stranger onto the bomb (Kamm 1989: 255). According to the *PPH*, the constraint on killing exists even though it was the agent herself who set up the bomb in the first place. In both cases—*Footbridge Paradox* and *Bomb Paradox*—that the five are saved is a result further down the causal line.

Kamm acknowledges that any moral principle that, like the *PPH*, prohibits the causing of harm even where this produces a greater good must seem problematic if we accept that we have reasons to promote the good. The challenge of how to defend the *PPH* comes in two parts.

3.3.1. The Priority Argument

Here is the first part of the challenge—followed by Kamm's response to it. It is uncontroversial that not causing others harm is morally important. But isn't it also morally important to do good, to aid, or to benefit them? Why, then,

---

There seems to be something about the mode of your killing in *Switch Paradox* that should prevent us from saying that you would harm someone to prevent greater harm *of the same type*.

should we think that it is ever impermissible to do harm if one could thereby do comparably *greater* good?

Kamm's response is that it would be "incorrect merely to assume that when 'do' […] modifies harm, it has the same weight as when it modifies good" (Kamm 1992: 381). Instead, morality would give priority to the negative of doing harm over the positive of doing good. It gives priority, as Kamm says, "to the inviolability of the person over his status as recipient of […] benefits" (Kamm 1992: 382). Call this the *priority argument*.

Here, Kamm already employs the idea of inviolability which will come to play a central role in her response to the second part of the challenge. The *PPH* reflects the idea that doing harm is morally more significant than doing good. For the *PPH* defines a set of circumstances under which it is impermissible to cause harm *even if* by doing so the agent would bring about comparably greater good. In other words, that there should be any such circumstances means that doing no harm is so important that doing harm sometimes cannot be justified even by doing *greater* good.

Arguably, however, one might rightfully ask what reason we should have to think that doing harm *is* morally more significant than doing good (Kagan 1989: 122). The most straightforward answer might be this. Perhaps there are cases in which doing harm is not, simply qua mode of action, more significant than allowing harm. But the cases in question are cases in which we would harmfully use an innocent person to prevent greater harm to others. And our intuitions on the worseness of harmfully using others if compared to allowing harm to happen are stable. We do not normally think that it can be right to use the massive man's body to stop the trolley in the *Footbridge* case, nor that it can be right to harvest the organs of the healthy patient in the *Transplant* case, even though in both cases this means that we would have to allow comparably greater harms to happen to others.

When it comes to our negative duties not to harmfully use other people, it seems true what Foot says: that "even where the strictest duty of

positive aid exists, this still does not weigh as if a negative duty were involved" (Foot 1967: 8). This would also explain our intuition that the agent should minimise harm caused by herself in the present moment. For instance, consider the *Original Trolley* case introduced by Foot (1967):

> Original Trolley. You could either steer the out-of-control trolley to the right track occupied by five strangers, or to the left track occupied by a single stranger.

In this case, you are confronted with a conflict between only negative duties, and it seems plausible to say that should resolve that conflict in favour of transgressing the smaller number of these duties.

3.3.2. The Futilitarianism Argument

Now for the second part of the challenge to defending the *PPH*. A critic of the priority argument might agree that in the *Footbridge* and *Transplant* cases, doing harm is morally more significant than doing good such that the agent should not kill one even to save five. Yet she might argue that this is so only because the intentional, direct killing of another person is a greater moral evil than a person's accidental or natural death. In neither of the two cases would the killing of the five involve any moral wrongdoing. They would die either by a trolley accident or from organ failure.

But what reason do we have to think that doing harm is more significant than doing good, where doing good means preventing *greater harm of the very same kind*? For instance, the *PPH* prohibits you from killing the healthy patient in the *Transplant* case. But it also prohibits your colleague from killing two healthy patients in the *Transplant Paradox* case. If you could prevent the killing of two healthy patients by killing a single healthy patient, why should you not do it? How can any moral principle that tells us what we

should not do to individuals "exclude minimizing as a justification for not abiding by [that very principle]" (Kamm 1989: 251)?

This is the second part of the challenge and a variant of the rationality paradox. Can we show that it is "not irrational or paradoxical to be concerned about rights and yet not minimize rights violations by transgressing rights" (Kamm 2007: 269)? Kamm's first response to the rationality paradox is this:

> a moral system—where a moral system is our attempt to represent moral truth—that permits minimization of the violations of a certain right by transgression of that very right essentially eliminates that right from the system, hence it would be futile as a way of showing respect for rights; it would be a 'futilitarianism' of rights. (Kamm 1989: 252)

In short, Kamm's argument seems to be this. Making it permissible to commit minimising violations of rights is a futile way of showing respect for rights. This is so because on a moral system[99] that permits minimising violations of rights we lack the relevant kind of rights altogether—such a system *denies the existence* of the relevant kind of rights in everyone. I shall refer to this as the *futilitarianism argument*.

The futilitarianism argument essentially provides an answer to Nozick's original question:

> How can a concern for the nonviolation of *C* lead to refusal to violate *C* even when this would prevent other more extensive violations of *C*? (Nozick 1974: 30)

[99] Kamm speaks of moral systems but could as well be speaking of moral theories. To me, there is no visible distinction between these two terms in her work, thus I shall use the terms interchangeably.

Taking *C* to stand for some right, the answer is simply that such concern *can only lead* to refusal to violate *C* because once morality allows for its violation it denies the existence of the relevant right altogether.

The futilitarianism argument depends on a far-reaching assumption: that by "rights" we must mean *constraining* rights. As we have seen, Nozick's utilitarian of rights believes in a moral system that accommodates rights—rights, even, that sometimes restrict our options to achieve the greater good in cases where the greater good is to maximise overall welfare. But these rights are not thought to act as constraints on the minimisation of rights violations overall; they are *non-constraining* rights. If one is ready to call such properties rights, then strictly speaking utilitarianism of rights is a theory that eliminates constraining rights, but not rights altogether from the moral system.

More accurately, then, the futilitarianism argument says the following: making it permissible to commit minimising violations of rights is a futile way of showing respect for *constraining* rights. This is so because on a moral system that permits minimising violations of rights we lack *constraining* rights altogether—such a system *denies the existence* of the relevant kind of rights.

However, whether constraining rights *are* the relevant kind of rights is precisely what is at issue. The question at the heart of the paradox of deontology, as we have seen, is how we can make sense of constraining rights given that *an alternative conception* of rights is available—rights as *non-constraining* rights—which would moreover allow us to minimise violations of these rights overall. This is a pretty good prospect from at least one plausible perspective in ethics. Thus, the futilitarianism argument alone does not seem to be a sufficient answer to the rationality paradox.

3.3.3. The Degradationism Argument

What we need is an additional argument that explains why the futilitarianism of constraining rights would indeed be a worrisome feature of a moral system.

Fortunately, Kamm provides such an argument: once we eliminate constraining rights from the moral system, she argues, we also eliminate the *concept of the person* that such rights give expression to from that moral system (Kamm 2001: 263). Since constraining rights give expression to a concept of the person as a more valuable type of being, a moral system without constraining rights might permit us to "save more people, but they would, in a sense, be less worth saving in our eyes" (Kamm 1989: 254). I shall refer to this as the *degradationism argument*.

Let us look at this argument more closely. The degradationism argument proceeds in three steps. First, suppose that it is permissible to kill *A* in the *Special Case* where this would prevent the killing of *B* and *C*. This would have to mean that *A*'s right not to be killed is "weaker" (Kamm 2007: 269) than it would be if it were impermissible to kill *A*. At this point, Kamm expands on the idea of inviolability that has already played a role in her answer to the first part of the challenge but has not been spelled out there. If *A* is protected by weaker rights, this indicates that *A* is *less inviolable*. *A*'s inviolability is a description of what it is impermissible to do to *A* and, as a moral status, increases or decreases with the set of harmful things that we may not do to *A*.

The second step generalises from the status of *A* to the status of everyone. Inviolability, as we have seen, is a *moral* status. It is a status of *A* qua person, thus "a function of what is true of any one person" (Kamm 2007: 254). If *A* is less inviolable, then so are *B* and *C*. If they were in *A*'s place, it would be equally permissible to kill *B* or *C* to prevent the killing of the two others. Thus, if it is permissible to kill *A* in the *Special Case* where this is the

only way to prevent the killing of *B* and *C*, then *everyone's* inviolability decreases by the same factor.

At the same time, if I have no morally permissible means available to save *B* and *C* (because I would have to kill *A* to save them), then this does not mean that *B* and *C* are any less inviolable. I do not fail to respect *their* inviolability status because letting it happen that they are killed does not deny that their status renders it *impermissible* that they are killed. (However, there is the worry that it might prove *B* and *C* to be less morally important in some other sense that is not captured by the notion of inviolability. As I argue later, this gives rise to two variants of the value paradox one of which I address in Section 3.5; the other one will be addressed in Chapter 4.)

Note that the degradationism does therefore not depend on any conception of the victims of minimising violations as having a *special* moral significance. As Kamm points out, the victims of such treatment need not refer to their special status. In fact, they do not have any special status—they "need only say that no one should be treated in this way" (Kamm 1992: 385). Rather, the victim's status constitutes a moral barrier that the agent must not cross even to enforce the interests of others who have the same status. To refer to a particularly vivid metaphor: the fact that I cannot pass through a solid wall does not mean that my physical incapability of doing so, nor the wall itself have any special significance compared to the physical incapability of others to pass through other equally solid walls.[100]

The third and final step of the degradationism argument extrapolates from the decreased inviolability of persons to their decreased moral worth.[101] Inviolability is an indicator of the inviolable entity's moral importance. If we are more inviolable, all this means is that morality attaches *greater moral value* to any one of us, protecting us from harmful treatments even at the

---

[100] Kamm has used this metaphor in a talk presented at the conference *Rethinking Moral Status* at the University of Oxford in June 2019.

[101] I will treat the terms moral worth, moral value, moral significance, and moral importance as largely interchangeable.

expense that other desirable ends cannot permissibly be achieved. As Kamm says, "individuals whose rights stand as a barrier to action are more potent individuals than they would be otherwise" (Kamm 1989: 254). In turn, if it is permissible to violate us in the pursue of other morally important ends, then we as individuals seem *less* important. Thus, if it is permissible to kill *A* to prevent the killing of *B* and *C*, then all persons are of less moral value than they would be if it were impermissible to kill *A*.

Note that spelled out in this way, the degradationism argument uses the *quantitative* notion of inviolability. Kamm however alternates between the quantitative and qualitative notions. Sometimes, for instance, she says that the permissibility of minimising violations would "deny the existence of the valuable status in everyone" (Kamm 2001: 307). This only makes sense if she means *hyperinviolability* status, as the denial of constraints does *not* deny that we are inviolable *to any, even so menial extent*. It denies only that we are hyperinviolable. In Section 4.2.3, I will argue that we should rephrase the inviolability account in terms of the concept of hyperinviolability.

3.4. Clearing the Air of Paradox

Kamm's inviolability account develops its full potential only once we combine the futilitarianism argument with the degradationism argument. This might work as follows.

A moral system that makes it permissible to commit minimising violations of rights—by endorsing the permissibility of *this type* of violations—is bound to express the view that each of us is a less inviolable and thus a less valuable, less important type of being. In contrast, a moral system that accommodates constraining rights reflects a concept of the person as a more inviolable, and thus more valuable, more important type of being (Kamm 2001: 310–311, Kamm 2007: 269). In a word, constraintism is the theory of our elevated moral worth. Eliminativism about constraints entails an

eliminativism of our elevated moral worth, a *degradationism* with regard to our moral standing.

This way, the inviolability account seems to avoid the rationality paradox that if something is valuable it seems irrational not to maximise the presence of that value overall. Recall that, on the reason-focus interpretation, the paradox of deontology exists because constraintists seem committed to accepting the truth of each of the following statements:

(1) I have reasons not to violate *R.*
(2) I have reasons to prevent violations of *R.*
(3) I can only either refuse to violate *R* or prevent a greater number of violations of *R*. (Introduction of paradox cases)
(4) I ought to do what I have most reason to do. (Introduction of the *PBR*)
(5) Balancing my reasons against rights violations, I have most reason to prevent the greater number of violations of *R*.
(6) Thus, I ought to prevent the greater number of violations of *R.*
(7) I ought not to prevent the greater number of violations of *R*. (Introduction of a deontic constraint)

The key to resisting the rationality paradox is to resist (5). The agent-centred approach aims to do so by rejecting (2) which however makes the approach particularly vulnerable to the charge of the value paradox. The inviolability account, in contrast, does not deny that we have reasons to prevent violations of *R* by others or our future or past selves. Instead, it resists (5) by showing that *although* we have such reasons, we have most reason not to violate *R* to prevent more violations of *R*. The inviolability account focuses on the value of moral status, a value that cannot possibly be furthered or maximised by committing a minimising violation. A concern for the moral status expressed by constraining rights "could not rationally be served by violation of that right" (Kamm 2001: 310). Instead, a permission to minimise rights violations simply "destroys [that] component of the moral system itself and alters

its essential structure" (Kamm 1989: 252). We would get a moral system on which we lack that status, i.e., the relevant value we were supposed to maximise.

Thus, the inviolability account holds that not all moral values can be furthered through action. Inviolability is a value which we can be *respected* in our action, but which we cannot promote through acting in one rather than another.[102] As Kamm puts it, inviolability is *not* "a consequentialist value that we promote by bringing about something through action or omission" (Kamm 2007: 29). Instead, the central thought is that there are different modes of engaging with what is valuable. Not all values can be *produced* and thus maximised; some values simply guide our action. This gives the inviolability account considerable momentum as a solution to the rationality paradox. For as Daniel Muñoz comments:

> If some acts essentially fit the world as is, rather than trying to change it, then it makes perfect sense why optimific changes could be wrong. (Muñoz 2021: 90)

As such, the idea of inviolability is again closely related to the Kantian idea of humanity as an end-in-itself. In Kantian thought, the idea of humanity plays an action-guiding role in practical deliberation: that persons are to be treated as ends, never merely as means, functions an *ideal* to guide our action, not a state to be achieved (Dean 2013: 173–174).

In sum, the fact that constraints do not allow for minimising violations of rights does not render them paradoxical once we understand that the value of having constraining rights lies in the moral status that they give

---

[102] Note that the distinction I use here is not identical to the commonly used distinction between *honouring* and *promoting* values (Pettit 1989, McNaughton and Rawling 1992). At a closer look, what is meant by honouring a value is usually nothing more than promoting it in one's own life. For instance, to honour the value of honesty is to be as honest as possible oneself. My point here is rather that certain kinds of values cannot be *produced* (not even in one's own life).

expression to—a value that cannot persist in a moral system that allows for minimising violations of these rights. This seems to avoid the rationality paradox.

One might find that there is a certain glitch in Kamm's account. As Kamm notes, it is "the permission to [φ], not any [φ-ing], that eliminates the right [against φ-ing] from the moral system" (Kamm 1989: 254). Thus, if I choose to commit a minimising violation of *R*, this has no effect on the right holder's inviolability status. Why, then, should I refrain from violating *R*? It looks as though by doing what is impermissible to do, I could combine the best of two worlds: I could make sure that fewer people are harmed by anyone's φ-ing without removing their valuable status as inviolable beings.

It is trivial that any moral theory which holds that it is sometimes impermissible to do what leads to the best outcome has this feature: that by doing what is impermissible, we could sometimes achieve better outcomes. I don't think this is a problem or speaks against this type of moral theory. As soon as the aim of our action is more complex than simply to always produce the greatest good, there will be situations in which doing the right thing comes with a sort of 'moral aftertaste'—the knowledge that there were other morally important things we couldn't protect. It is worth noting, however, that the inviolability account asks us to buy into a specific understanding of moral status as something that our actions might *come up against* but that we cannot spoil or destroy through action. Even by murdering another person, we do not destroy their status as someone who may not be murdered because having murdered them does not change the fact that it was impermissible to do so.[103]

[103] Note that this does not mean that we would have to attribute moral status to non-existent entities. For instance, we do not have to be able to attribute moral status to the dead body in order to claim that murdering the person whose body it is was impermissible. What allows us to say in the present moment that it *was* impermissible to murder them one minute ago is simply that at the moment when they were murdered, they *had* the status of a being who was wronged by that act.

3.5. Better Moral Worlds

As I have argued, addressing the rationality paradox is just one part of the challenge of justifying constraints. We also need to answer the second part of the paradox of deontology, i.e., the value paradox.

Moreover, I have noted that the value paradox has many faces and that, as an objection against the inviolability account, it may take the form of an external and the form of an internal criticism (see Section 1.6.4). The *external* value paradox says that we cannot take for granted, without further argument, that we should care more about our moral status as inviolable beings than about the things that may happen to us. The *internal* value paradox says that we cannot take for granted, without further argument, that we should care more about our inviolability than about other dimensions of our moral worth. Chapter 4 is dedicated to addressing the internal version. Here, I want to address the external version first.

So, why should we care about our moral standing? This question might not be too difficult to answer. The thought that it is good if there are things that others may not do to us because we matter morally in our own right is not only directly accessible to us, without extensive explanation or justification of its content; it is also an *appealing* kind of thought. But why should we care *more* about our moral standing than about the things that might happen to us? That we *should,* is a strong assumption, to say the least.

Imagine a cruel lottery which will assign each of three persons randomly to one of the three possible positions of *A*, *B*, and *C* in the *Special Case*. They are required to take a secret vote before to decide whether the moral system that applies should permit minimising violations or not.[104] While there is always a minor chance that people will not do what is morally right, the

[104] The thought experiment I use here has obvious structural similarities to Rawls's *original position*. But this should be seen as an innocent resemblance as the bigger picture that Rawls is after has no bearing on my further argument.

three voters live in a world in which most people at most times do as morality says. Each of the three could now follow a particular line of thinking:

'Surely it would be great if we were hyperinviolable. It would mean that we were morally more important creatures, and I see that having this status would be valuable in itself. However, since I could end up in any position and it is more likely that I end up in the position of *B* or *C* than that I end up in *A*'s place, I am more likely to live if we are not hyperinviolable. I am happy to trade the valuable status for a greater chance to live. After all, once I lose my life, what does it matter which elevated status I had at the time when I was killed?'

This line of thought does not seem absurd. A world without constraints is one in which we must believe, as Kamm said, "in a less sublime and elevated conception of ourselves" (Kamm 1989: 254). But, at the same time, it is a world in which fewer of us end up being violated. Even if having the status of someone who *may not* be violated is valuable in itself, this doesn't show that we should care more about having that status than about how likely we are to *be* violated. How can we just assume that morality should give priority to the moral worth of persons over what happens to them, especially if the inattention to what happens to them seems to undermine the very idea that they have been assigned a type of moral worth which is worth having?

Kamm provides us with *some* argumentative material to address the external value paradox, although these materials are rather scarce in comparison to the three arguments discussed earlier.

In a nutshell, Kamm argues that "it is better to have a world populated by more important entities" (Kamm 2007: 227). For example, it is better to have a world populated by the kind of entities we are if we are protected by constraints. And thus, constraints make for a better moral world, a world populated by more valuable entities. Elsewhere, she says that "we do not make people inviolable. They either are or are not inviolable. If they are, we should act in accord with this" (Kamm 2007: 269). The bigger picture of what Kamm

is saying in these passages is often taken to be this: a world in which we are inviolable to the extent that we are protected by constraints is the better moral world, compared to one in which we lack that protection. Thus, we *are* inviolable to the extent that we are protected by constraints. In short: it is better if we are hyperinviolable. Therefore, we are hyperinviolable. Call this the *better world argument*.

3.5.1. A Suspicious Form of Reasoning

On my interpretation, the better world argument is Kamm's response to the external value paradox. Constraintism is correct in giving priority to our inviolability over what happens to us *because we are hyperinviolable*, and a moral theory is nothing more than "our attempt to represent moral truth" (Kamm 1989: 252). As it stands, however, the better world argument looks like a suspicious form of reasoning. After all, I cannot infer the truth of a proposition *p* from the fact—even if it is a fact—that it would be *better* if *p* were true.

Compare the better world argument to the following piece of reasoning: *It would be better for me if it wasn't raining, and therefore, it is not raining*. It is easy to imagine a desperate hiker who employs this line of reasoning perfectly, and yet is standing in the rain with no shelter, miles away from home. Sometimes, the world is not such that what would better be true *is* true. This is the vice of wishful thinking. As a matter of fact, sometimes raindrops are falling on my head even though they would better not be. *This* argument, call it the *desperate hiker's argument*, is clearly a non-starter.

Why should it be any different with the better world argument? Sure, it would be nice to have hyperinviolability status. But how could this ever show that we do *in fact* have that status?

Yet Kamm is not the only one to appeal to this form of argument. Thomas Nagel and Warren Quinn both refer to the better world argument in

the justification of rights (Nagel 2002: 37–40, Quinn 1993: 149–174).[105] Moreover, Nagel gives a reason why we should think that the better world argument is not just a non-starter like the desperate hiker's argument:

> The argument is that we would all be worse off if there were no rights […]—ergo, there are rights. This is a curious type of argument […]. However it may have a place in ethical theory, where its conclusion is not factual but moral. It may be suitable to argue that one morality is more likely to be true than another, because the former makes for a better world than the latter […]. (Nagel 2002: 39)

Here, Nagel suspects that the better world argument might have a place in moral theory; that the inference *It would be better if p, therefore p* is a valid form of argument for a *moral p*. In fact, moral instances of the better world argument do not sound "so obviously ridiculous" as its non-moral instances (Enoch 2009: 222). There seems to be some difference between Kamm's better world argument and the desperate hiker's argument.

David McNaughton and Piers Rawling raise two further questions about Kamm's argument (McNaughton and Rawling 1998: 53). Compare two possible moral worlds: in ω_1 we are hyperinviolable, whereas in ω_2 we are not. A first question is in what sense ω_1 and ω_2 are *possible* moral worlds? If, as Kamm seems to believe, we simply live in the world that turns out to be

[105] Instances of the better world argument can be found outside of the debate about constraints. For example, Michael Slote seems to appeal to this form of argument in order to justify agent-centred permissions (Slote 2020: ch. 2). His argument is that if there are such permissions, then we have a sort of autonomy that we lack otherwise, which he calls *moral autonomy*. Moral autonomy is the freedom to choose what sort of life one wants to live in morally relevant ways. It is better if we have moral autonomy, and therefore, we have moral autonomy. Another curious example might be Richard Swinburne's argument for the existence of God (Swinburne 2008). In essence, Swinburne seems to argue that if there is a God, the world is a better place because there are additional moral truths, and therefore, there is a God (although Swinburne does not make the conclusion explicit in this case).

the better world (ω_1), then this *particular* world is the actual moral world, and ω_2 is not possible.

The second question is in what sense ω_1 would be *better* than ω_2? Certainly, the utilitarian of rights has a plausible argument as to why ω_2—not ω_1—is the better moral world. Given the wrongdoing of some, we still have permissible means available in ω_2 to achieve the best available outcomes in which only as few rights violations as possible occur. Whether we think that ω_1 or ω_2 is the better moral world surely is not an entirely objective matter.

I shall address both these points in Section 3.5.3. In general, I believe Nagel is right when he claims that the better world argument has a rightful place in moral theory. Yet I also agree that anyone who wishes to defend it will have to admit that there remains something suspicious about the form of the argument, so understood. My proposal is that we look at what could distinguish the better world argument from its non-moral counterparts, like the desperate hiker's argument. As far as I can see, there is something to say about the sense in which these different arguments talk about worlds that are possible. I will then propose a rephrased version of the better world argument—the *better theory argument*—and show how it avoids the problems of the original argument.

3.5.2. Genuinely Possible Worlds

We can think about possible worlds, evaluate, and compare them in terms of how things are in those worlds *in morally relevant ways*. That is, we can treat them as possible *moral* worlds.[106] For instance, suppose that there are two possible worlds, ω_3 and ω_4. The world ω_3 contains the occurrence of a

[106] While it is always a risk to use a much-discussed philosophical term without proper acknowledgement of the discussions around its use and meaning, I shall use *possible world* as a helpful notion to talk about how things *could be* or *could have been*; a notion whose meaning is directly accessible to us without deeper knowledge of its use as a technical term in modal logic.

terrorist attack tomorrow at noon at Gotham City Library[107] that kills thousands of people, whereas the world ω_4 does not contain the occurrence of a terrorist attack tomorrow at noon at Gotham City Library. Otherwise, the two worlds are identical. It seems plausible that we can make the following kind of statements about these two worlds: other things being equal, ω_4 is the better moral world, compared to ω_3, because it is morally very bad if thousands of people are killed in a terrorist attack. World ω_4 is preferable to world ω_3. And so on. I take these to be perfectly meaningful things to say.

Now, ω_3 and ω_4 are *possible* worlds in a specific sense of the word. There is a moment in time, t_1, at which none of these worlds is actual, and a second point in time, t_2, at which one of these worlds is *actual* and the other one is *not*. Tomorrow at noon, Gotham City Library will either be the scene of a terrorist attack or not. In other words, the truth value of the proposition *that a terrorist attack takes place on that particular day at noon at Gotham City Library* depends on the moment of evaluation. By the moment of evaluation I mean the moment in time at which the subject or speaker evaluates the truth or falsity of a proposition. There is a possible history where the attack occurs and one where it does not occur. Thus, I shall call ω_3 and ω_4 *contingently* possible. Two worlds that are contingently possible are *possible* in the sense that, depending on the moment of evaluation, either *could be* or *could have been* the actual world.[108]

The better world argument does not seem to be talking about moral worlds that are *possible* in the above sense. Compare ω_1 in which we are hyperinviolable to ω_2 in which we are not. It seems that there is *no* moment in

---

[107] I use a fictional place here so as not to assume any terrorist attacks at real places. For the sake of argument, I ask the reader to imagine that Gotham City Library were a real place.

[108] Note that what I call genuine possibility is the usual notion of possibility in modal logic. Worlds are possible in this sense if they are candidates for the actual world. To be clear, all I mean by contingency here is that it is *not necessary* that some *x* has feature *F*. It is not necessary that ω_3 and ω_4 have the feature of being possible because there are times when they are not possible.

time, t_1, at which none of those worlds is actual, and no second moment in time, t_2, at which one is actual, the other is not. What even would the condition be under which it is true that either ω_1 or ω_2 have been actualised? If by the end of the year 2025 all moral theorists finally agree that we are hyperinviolable, is that the moment in time at which ω_1 has been actualised? Or is it the moment when all states in the world have codified a positive law that reflects our hyperinviolability status? Neither seems to be the case. Worlds like ω_1 or ω_2 are never actualised. They are *just always possible*. They are possible so long as someone can claim or argue that we are or are not hyperinviolable. They are, as I shall say, *genuinely* possible moral worlds. Two worlds that are genuinely possible are thus *possible* in the sense that, *independently of the moment of evaluation*, either *could be* the actual world.[109]

The distinction between contingently and genuinely possible worlds provides a conceptual tool for distinguishing the better world argument from the non-moral instances of the same inference, such as the desperate hiker's argument. The desperate hiker's argument talks about worlds that are *contingently* possible. Whether the proposition *that it is not raining* is true or false depends on the moment of evaluation (and, of course, on the location of the speaker). At that moment, it either is or is not raining. It is not genuinely possible that it is raining in the same way that it is genuinely possible that we are hyperinviolable. *That* is why the desperate hiker's argument is a non-

[109] One could argue that also genuinely possible worlds are not possible in a sense that is completely independent of the moment of evaluation. That is, one could argue that before it is possible that there are human rights, for instance, such values must first somehow *find their way into existence*. It is plausible to think that many moral values depend on social practices at least in the way they come into being. This is roughly Joseph Raz's *social dependence thesis* (Raz 2003). Thus, the possibility of there being a human right against torture, say, is not completely independent of the moment of evaluation. It depends on whether such values have already come into being. This does not undermine my argument. From this perspective, once the concept of human rights is in the world and has been assigned meaning, whether it applies to us in the sense that we *in fact have* human rights is a matter of *genuine* possibility.

starter: whether it is better if it is not raining has no impact on contingently possible worlds.

On a side note, the same should apply to the following kind of argument: *It would be better if the terrorist attack didn't take place, and therefore, it didn't take place*. This argument, too, is clearly a non-starter because whether or not a terrorist attack takes place at Gotham City Library on a particular day at noon is a matter of contingent possibility. There is a moment in time (before the specified date and time) when the attack *could* take place, and another moment in time (after the specified date and time) when it either *did* or *didn't* take place.

This is good news. We have identified a feature of the better world argument that the obvious non-starter arguments like the desperate hiker's argument lack. But with a piece of good news comes a piece of bad news. The aspiration was that the better world argument could establish something. It was supposed to establish the truth of some moral proposition, namely that we *are* hyperinviolable. But if the truth of that proposition is *genuinely* possible, then it is also *genuinely* possible that it is false. If it is genuinely possible that we are hyperinviolable, then it is also genuinely possible that we are not. What, then, could an argument establish that talks about genuinely possible moral worlds?

3.5.3. What the Argument Can Establish

I believe the better world argument can establish something, not that we are in fact hyperinviolable, but something about moral theories. Note that on my interpretation, no moral theory can establish the truth of some moral proposition. No theory can show conclusively that we are hyperinviolable, nor that we aren't, that we have human rights, that the right thing to do is to promote good outcomes, etc. I will not defend this interpretation any further here. To

me, it seems like an unreasonable expectation to think that moral theories could do any such thing.

Instead, moral theories are in the business of *representing the truth* of moral propositions. And they do so not with any kind of moral propositions, but with those whose truth is *genuinely* possible. This is the same as saying that moral theories are in the business of representing genuinely possible moral worlds. For instance, for a moral theory *H* to hold that we are hyperinviolable *just means* that *H* represents a genuinely possible world in which we are hyperinviolable. For an alternative theory *M* to deny that we are hyperinviolable just means that *M* represents a genuinely possible world in which we lack the relevant status.

I think this puts us in a position to rephrase the better world argument as follows:

(1) It is genuinely possible that we are hyperinviolable.
(2) Other things being equal, it would be morally preferable[110] if we were hyperinviolable than if we did not have that status.
(3) Thus, other things being equal, a moral theory *H* that holds that we are hyperinviolable is (on the very same grounds) morally preferable to an alternative theory *M* that denies us that status.
(4) We should accept the moral theory which is morally preferable to its alternatives (as long as what the theory holds is genuinely possible).
(5) Thus, we should accept *H*.

[110] I say morally preferable instead of "better" because the adjective better is often understood as limited to outcomes. But it is important for the argument that moral worlds can be preferable on some other grounds than better outcomes.

In short, if it is morally preferable that *p* is true, and *p* is genuinely possible, then we should accept a moral theory that represents the truth of *p*. Call this the *better theory argument*.[111]

The better theory argument avoids the problems faced by the original better world argument. For one thing, unlike the better world argument it exhibits a cogent form of reasoning. The better theory argument is not an instance of wishful thinking but an argument about the preferability of moral theories. For another thing, the better theory argument helps to address the first point raised by McNaughton and Rawling, namely that it would be unclear in which sense the relevant worlds are *possible*. As they argue, if we simply live in the world that turns out to be the better world, then this *particular* world is the actual moral world, and all alternative moral worlds are not possible. We can now see that this objection is based on the misunderstanding that the argument would talk about worlds that are contingently possible. But instead, the argument talks about possible worlds that are just always possible. It does not try to establish that we actually are hyperinviolable, just that we should think that we are.

Whether the better theory argument is plausible then comes down to what we think about its second premise: is a world in which we are hyperinviolable *really* morally preferable to a world in which we are not? This was McNaughton and Rawling's second point. Does the better theory argument help to address *this* point?

On the face of it, it doesn't seem so. The better theory argument itself includes no further justification of the claim that it would be preferable if we

---

[111] Preston-Roedder (2014) suggests an interpretation of the better world argument that goes into a similar direction. Preston-Roedder, too, proposes that we shift the focus to the preferability of moral theories. But there are two important differences between his argument and mine. First, the centrepiece of my argument is the insight that plausible moral theories must talk about moral propositions the truths of which are genuinely possible; there is no equivalent piece of thought in Preston-Roedder's argument. And second, his argument eventually takes a different route because he argues that the better world argument helps us to understand that certain moral theories are *self-defeating* once we understand that their truth would make the world worse.

were hyperinviolable. But even so, the better theory argument does not establish anything less than the better world argument. The idea that it is better to be hyperinviolable than to lack that status was one of the premises of the original better world argument for which the argument itself offered no further support or justification.

But where does this leave us? The truth of the conclusion that a moral theory which represents that we are hyperinviolable is preferable to one that represents that we lack that status depends on the truth of the claim that it *is* preferable to have that status. But if the argument does not show this, how can it show that we should accept constraintism?

I think at this point it is important to recall what it is that the inviolability account is supposed to establish. It is not supposed to show that constraintism *is true*, not even that we *should accept* constraintism *period*. To address the paradox of deontology we need not show that constraintism *is true*, only that constraintism is internally coherent. There seems to be a plausible perspective in ethics that focuses on the moral significance of persons rather than the moral significance of what happens to them. From *this* perspective, since constraints express (or seem to express[112]) a concept of the person as a more valuable kind of being, we should accept constraintism rather than eliminativism. For constraintism represents a genuinely possible moral world in which we are beings of greater moral significance.

3.5.4. Wider Implications of the Argument

The better theory argument might be said to have wide implications in moral theory. Most importantly, it seems that the *form* of argument could be used to establish the preferability of many different normative views. To begin with, all instances of the original better world argument can be rephrased in

[112] I say *seem to* because this claim stands pending the need to address the internal value paradox, which I shall do in the next chapter.

terms of the better theory argument. For instance, take Michael Slote's argument for the existence of agent-centred permissions (Slote 2020: ch. 2). In short, if there are such permissions, then we have a sort of autonomy that we lack otherwise, i.e., *moral autonomy*. Moral autonomy is the freedom to choose what sort of life one wants to live in morally relevant ways. It is better if we have moral autonomy, and therefore, we have moral autonomy. Rephrased as a better theory argument we get: if it is morally preferable, other things being equal, to have moral autonomy than to lack that kind of autonomy, and it is genuinely possible that we have moral autonomy (which seems to be the case), then we should accept a moral theory that represents the truth of the proposition that we have moral autonomy. Again, the question whether we should accept this argument depends on the premise that it is in fact preferable to have moral autonomy.

But it seems that the better theory argument could be used to argue for the preferability of all kinds of other normative views. For instance, what if we apply the argument to compatibilism as a solution to the free will problem? Roughly, the idea would be that if determinism is true, then it might be morally preferable that compatibilism is true. For if determinism is *incompatible* with free will, then we do not seem to be morally responsible for our actions. Does this mean that we should accept compatibilism, i.e., the view that represents the truth of the proposition that determinism does not threaten free will?[113]

I think the answer to this question depends on two further questions. First, is it *genuinely* possible that compatibilism is true? One could argue that whether free will—in any appropriate sense of the term—and determinism are compatible depends on physical circumstances. Without cashing out this view in any more detail, it is important to note that the better theory argument requires that the propositions one argues for are *genuinely* possible. Thus, if it would turn out that the truth of compatibilism is *contingently*

[113] I owe this critical question to Shaun Nichols.

possible, then the better theory cannot be used to show that we should accept compatibilism.

Second, is it morally preferable that compatibilism is true? This is the same question as we can ask about the better theory argument regarding the hyperinviolability of persons. If the premise which states the preferability of the truth of some moral proposition is false, then the conclusion about the preferability of the relevant moral view will come out as false, too. Of course, one might claim that if determinism is true, it would be preferable if we could still be held responsible for our actions, for instance, for the possibility of cooperation and the management of our social lives. But there is also the counterclaim that once determinism is true, it does no longer matter whether we are responsible for our actions or not (as what will happen will happen either way). Thus, I don't think it is obvious that it is morally preferable that compatibilism is true.

However, if the answer to both questions is 'Yes'—i.e., if it is genuinely possible that compatibilism is true and that it is preferable if it is true—then I think the better theory argument could be used to show that we should accept compatibilism. I don't see this as a problem. The central insight of the better theory argument is that when accepting one or the other moral theory has an impact on how morally good the world is, we should accept the theory which makes for a better world. This is the case whenever the world a moral theory represents is *genuinely* possible. All the better theory argument shows us is one way of how to justify moral theories.

3.6. Conclusion

Recall the formulation of the paradox of deontology developed in the first chapter:

(1) I have reasons not to violate *R.*

(2) I have reasons to prevent violations of *R.*

(3) I can only either refuse to violate *R* or prevent a greater number of violations of *R*. (Introduction of paradox cases)
(4) I ought to do what I have most reason to do. (Introduction of the *PBR*)
(5) Balancing my reasons against rights violations, I have most reason to prevent the greater number of violations of *R*.
(6) Thus, I ought to prevent the greater number of violations of *R.*
(7) I ought not to prevent the greater number of violations of *R*. (Introduction of a deontic constraint)

The inviolability account does not deny that we have reasons to prevent rights violations by others (or ourselves at different times). Instead, it solves the paradox by denying (6) via a rejection of (5). If the value governing my action is the inviolability status of persons, not the disvalue of the actual violation of their rights, then I do not, as (5) states, have most reason in paradox cases to prevent the greater number of rights violations. On a moral view that gives priority to the moral worth of persons over what happens to them, I have most reason to respect the inviolability status of the person whose right I would have to violate.

In a word, constraints cease to appear paradoxical once we understand them as part of a moral theory that gives priority to the moral worth of persons. Considering what the permission to commit minimising violations would mean, it is *not* irrational for such a theory to endorse constraintism.

4 *The Dimensions of Moral Standing*

4.1. Introduction

As we have seen, the value paradox can take the form either of an external or an internal criticism against the inviolability account. The main purpose of this chapter is to address the internal value paradox. The first version of this line of criticism has been raised by Shelly Kagan and is sometimes referred to as the *saveability objection* or *saveability challenge*. Kasper Lippert-Rasmussen and Michael Otsuka have presented similar challenges.[114] All three confront the inviolability account with a particular circumstance: that there are various dimensions to our moral standing, not all of which can be captured by the notion of personal inviolability. Our moral standing is indicated not only by our inviolability but also, for instance, by how *saveable* we are.

The criticism based on this insight is an *internal* kind of criticism in the sense that it accepts the initial premise of the inviolability account, namely that the relevant value we should care about is not the significance of what

[114] Kagan raised the saveability objection in a short paper replying to Kamm and other of his critics (Kagan 1991). According to my interpretation, which I will argue for later, Lippert-Rasmussen (1996), (2009), and Otsuka (2011) both put forward versions of Kagan's criticism. I shall use the name *saveability challenge* to refer to the objection in all three versions.

happens to us but our moral significance itself, i.e., our moral standing as persons. In contrast, in the last chapter I have discussed the *external* criticism that asked why, where the minimising of harm is concerned, we should care that much about our moral standing at all.

For the central argument of this chapter, the distinction between internal and external criticism is an important one. An internal criticism, as I understand it, is a way of criticising an argument *on its own terms*. It means to attack the argument based on one or more premises which are part of the original argument, aiming to show that the argument is inconsistent or insufficient. In general, an internal criticism might be understood to be more threatening than an external one because it threatens to undermine the coherence of the target moral view itself. However, as I argue in this chapter, the saveability challenge in all its versions fails as an *internal* criticism. It leaves the constraint sceptic with a weak version of an external criticism that poses no serious threat to the inviolability account.

Kamm's account is vulnerable to the saveability challenge because it combines two features: (1) an underlying *gradual* understanding of moral standing, and (2) a focus on *quantitative measures* regarding the factors which have an impact on the degree of overall moral standing. The saveability challenge deserves special attention but as we will see, the combination of these two features is also the source of another problem which I shall call the *problem of source-plurality*. Section 4.2. will expand on the discussion of quantitative and qualitative measures of inviolability and address the problem of source-plurality. Section 4.3 raises the saveability challenge in its various versions. Finally, Section 4.4 shows how the challenge fails as an internal criticism against the inviolability account.

4.2. Quantitative and Qualitative Inviolability

Recall that, on Kamm's view, inviolability is indicated by the set of types of circumstances in which it is impermissible to harm some entity (e.g., Kamm

1992: 383). If there are more types of such circumstances, then the entity in question and all entities of its kind are more inviolable. If there are fewer types of such circumstances, then the entity and all entities of its kind are less inviolable.

That means, Kamm defines inviolability in terms of the quantity of ways in which it is impermissible to treat some entity. She therefore commits herself to an understanding of inviolability, first and foremost, as a *quantitative* concept (Lippert-Rasmussen 2009: 175). Furthermore, the quantitative notion of inviolability corresponds to a gradual conception of moral standing. Although Kamm explicitly says that we should not conceive of moral status as equivalent to inviolability (Kamm 2001: 275–277), a higher level of inviolability, other things being equal, *indicates* or *corresponds to* a higher level of moral standing. The higher our inviolability, the higher our moral standing.

One implication of this view is that, in principle, inviolability is no exclusive feature of the moral standing of persons, human beings, or rational beings. Instead, a great many kinds of entities turn out to be inviolable on Kamm's account. It seems that we can speak of the inviolability (i.e., the *degree of* inviolability) of persons just as we can speak of the inviolability of non-persons. As Andrew Ross points out:

> if inviolability is understood as a list of restrictions detailing how we may treat an entity, then animals and plants also possess inviolability insofar as they too cannot be treated in certain ways. (Ross 2016: 71)

As Kamm suggests, even inanimate objects like artworks might be inviolable to some extent (Kamm 2007: 228). Yet since the number of restrictions associated with persons seems higher than the number of restrictions associated with animals, plants, or artworks, persons possess a more elevated level of

inviolability than those other types of entities.[115] But the difference between the inviolability of plants, animals, artworks, and persons is, first and foremost, a difference *in the degree of* inviolability possessed by these kinds of entities.

Thus, on Kamm's account, differences between the moral importance of different types of entities are due to differences in the levels of inviolability associated with those types of entities. If persons are inviolable to the extent that there are constraints on harming them, this only means that persons possess a certain elevated level of inviolability.

This quantitative understanding of inviolability invites the following kind of worry. Recall Kamm's *degradationism argument*. In short, if we are not protected by constraining rights, then we are less inviolable and thus, less valuable beings than we would otherwise be. By contrast, constraining rights give expression to our elevated moral importance. However, it seems that mere quantitative considerations about the different levels of moral standing, indicated by the levels of inviolability we possibly have, cannot show this (Lippert-Rasmussen 2009: 175). For it could be that we have an elevated moral standing *even if* we lack constraining rights—either *qua being highly inviolable in other ways* or *qua being highly important in another sense*. The first option gives rise to what I shall call the *problem of source-plurality* (see Section 4.2.1). The second option leads to Kagan's saveability challenge (see Section 4.3).

---

[115] Note, however, that Kamm suggests that symbolic entities like flags or items of religious value might have a status such that they may not be readily destroyed even to save other objects of their type from being destroyed (Kamm 2007: 256). If this is true, then the fact that persons have a higher moral standing than flags—which I take for granted—needs to be established through a direct comparison between those entities. For instance, it is *not* impermissible to sacrifice any number of flags to save the life of only one person. But it is impermissible to sacrifice only one person even to prevent the destruction of any number of flags.

4.2.1. The Problem of Source-Plurality

Let me begin with the first option. That is, if minimising violations were permissible, we could still possess a high level of inviolability—comparable to the level of inviolability we would possess if such violations were impermissible—so long as there are *other things* that may not be done to us.

Kamm explicitly approves of the idea that our elevated inviolability may come from a variety of sources. For instance, most of us think that it is permissible to turn the trolley in the *Switch* case. But suppose that this is false. Consider a moral theory, *S*, on which it is *impermissible* to turn the trolley in *Switch*. In general, *S* holds it to be impermissible to redirect a deadly threat caused naturally or by accident from five onto one. On Kamm's view, *S* holds that we are more inviolable than we would be if it were permissible to redirect such threats (Kamm 1992: 383).

Now suppose that, at the same time, *S* does *not* make it impermissible to commit minimising violations of rights. Thus, *S* does not include constraining rights. Could we not say that we are comparably inviolable on *S* than on an alternative, constraintist theory *H* that says that we have constraining rights but that holds it *permissible* to turn the trolley in *Switch*? On *S*, our elevated inviolability would simply derive from a source *other than* the impermissibility of minimising violations (Lippert-Rasmussen 1996: 339). If *any* types of circumstances in which it is impermissible to harm us would make us more inviolable, then this would not seem to matter, at least not for the question whether *S* or *H* give expression to a higher level of moral standing. Within *S*, we are *less* inviolable because minimising violations are permissible, but we are also in a sense *more* inviolable than we are on *H* because it is impermissible to harm us in cases like *Switch*. From this perspective, *S* might be said to *compensate us* for the permissibility of minimising violations by balancing

the decrease in our inviolability with an increase in our inviolability that comes from another source.

I shall refer to this as the *problem of source-plurality*. In short: if inviolability is a mere quantitative measure of moral worth defined in terms of the harmful things that others may not do to us, then there is no reason to think that a moral theory that entails constraints gives expression to a higher level of inviolability than one which does not include constraints. And thus, there is no reason to think that the second kind of theory holds us to have a lower moral standing overall. On the second kind of theory, our high level of inviolability—and hence, our high moral standing—could just be due to other impermissibilities than the ones described by constraints.

4.2.2. Inviolability and the Capacity to be Wronged

The problem of source-plurality shows that a plausible account of the status rationale for constraints cannot rest solely upon considerations about the quantity of harmful things that are impermissible to do to persons. As Kasper Lippert-Rasmussen says:

> Any explanation couched exclusively in such quantitative terms will be a bad one because it will be possible for the explanandum—the impermissibility of [minimising violations]—to obtain whether or not the proposed explanans—that we have a certain high degree of moral status or inviolability—does. (Lippert-Rasmussen 2009: 175).

Kamm is not unaware of the problem. She admits that "simple talk about inviolability"—here she means talk about inviolability as a quantitative measure of moral worth—"cannot be all we need to explain the presence of a constraint" (Kamm 1992: 384). Early on, she thus flirts with the idea that the inviolability represented by constraints could constitute a specific *sort* of inviolability (Kamm 1992: 384). Elsewhere, she speaks of the respect we are

worthy of if it is impermissible to minimise violations of rights as "owed respect" (Kamm 2001: 321). And in her later comments on the inviolability account, she characterises the duties we have not to violate constraints as taking the form of *directed duties* towards persons such that we "wrong the person if we violate the constraint, as we owe it to him not to do it" (Kamm 2007: 231).

Yet Kamm does not develop these thoughts in more detail. However, I believe the inviolability account develops its full potential when we combine the idea of inviolability with the capacity of persons to be wronged. Let me elaborate on this.

Not all entities that possess some degree of inviolability also possess the capacity to be wronged. For instance, it might be wrong to trample a bed of sunflowers for no reason—and in so far sunflowers can be said to be inviolable to some extent. But it does not seem to be true that I owe it to the individual sunflower not to trample it (Ross 2016: 69).

The same may or may not apply to animals. If animals—I should say non-human animals—are like plants, just a bit more morally important than those, then the set of things we may not do to, say, a tiger is larger than the set of things I may not do to a sunflower. And thus, tigers are more inviolable than sunflowers. But the difference between the two would merely be a difference in the level of inviolability possessed by each type of entity. This fits well with the intuition that the set of restrictions associated with at least most animals is significantly smaller than that associated with persons. It seems permissible, for instance, to kill a small number of deer in order to reduce the

deer population in a forest, if this is necessary to prevent overpopulation and the dying of many more deer (Otsuka 2011: 48).[116]

Thus, the inviolability of persons is not only higher than those of other entities such as plants and tigers. Rather, it is a *specific sort* of inviolability. I have earlier referred to this sort of inviolability as *hyperinviolability*. On the face of it, it seemed that hyperinviolability could just be the name for the elevated *degree of* inviolability possessed by those entities who morality protects from minimising violations. But if only entities who have the capacity to be wronged can count as hyperinviolable, we have a *qualitative* criterion that reserves that particular status to those kinds of entities who, like persons, can be wronged.

Once we connect constraints to the capacity to be wronged, we also have the conceptual resources to distinguish between the moral status expressed by constraints and the constraint-like restrictions which derive from the *unsubstitutability* of certain inanimate objects. As we have seen, Kamm suggests that items of symbolic, religious, or aesthetic value should not be destroyed even to save other entities of their type from being destroyed (Kamm 2007: 256). However, so long as we do not believe that we *owe it to* those entities themselves not to destroy them, such items are *not* hyperinviolable.[117] In any case, the concept of hyperinviolability helps us to resist the impression that the fact that some entities possess both the capacity to be

---

[116] Of course, this is merely an intuition. If animal rights theorists are correct in claiming that some non-human animals "resemble normal humans in morally relevant ways" (Regan 2004: xvi), then perhaps also those animals are protected by constraints. In this case, however, it is also reasonable to attribute to animals the capacity to be wrong. Animals would just be much more like persons, from a moral perspective, than like sunflowers.

[117] We would then have to find another way of explaining the unsubstitutability of such objects. I think a quite natural way of doing this would be to refer to the interests of persons who attach value to those objects and care about their persistence. That is, it seems reasonable to think that some objects are not substitutable because someone values them in a certain way. It is hard to imagine unsubstitutable objects in a word without persons, for instance.

wronged and an elevated level of inviolability is nothing but a "metaphysical accident" (Ross 2016: 80).

4.2.3. The Hyperinviolability Account

The inviolability account is vulnerable to the objection that it cannot establish that we are more important if we have constraining rights because we could be just as important, even *more* important, if we were inviolable in other ways. In order to avoid the problem of source-plurality, I propose to rephrase the inviolability account based on the qualitative notion of hyperinviolability. This might work as follows.

First, let me begin with the degradationism argument: if we are not protected by constraints, then we lack hyperinviolability status. We have that status if and only if we are protected by constraints. Other things being equal, the person as hyperinviolable is a more valuable type of being than the person as lacking that status. Thus, constraints give expression to a concept of the person as a more valuable type of being, a being of higher moral worth.

Second, combine this with Kamm's futilitarianism argument. A moral system that denies constraints also denies the concept of the person they give expression to. That is, it denies the concept of the person as a hyperinviolable being. Such a moral system is bound to express the view that persons are less valuable, beings of lesser moral worth, than they would otherwise be.

Therefore, constraints cease to appear paradoxical once we understand them as part of a moral theory that gives priority to our moral value over the value of what happens to us. Once we focus on the value of moral standing, it is *not* irrational to think that there are constraints because constraints give expression to our elevated moral standing as hyperinviolable beings. Call this the *hyperinviolability account*.

The hyperinviolability account can avoid the problem of source-plurality. Quantitatively speaking, we might be just as inviolable if minimising

violations were permissible as we would be if such violations were impermissible, if only we were more inviolable in other ways. But we are hyperinviolable only if there are constraints. Since hyperinviolability is qualitatively defined as the status that we have if and only if minimising violations are impermissible, it cannot be true that we could be just as *hyperinviolable* if minimising violations were permissible.

One might object that the hyperinviolability account circumvents the problem of source-plurality only in terms. To resurrect the problem, it could simply be rephrased as follows: it could be that we possess a comparably high *moral standing* if we possessed a sufficiently high level of inviolability, without being hyperinviolable. After all, what Kamm's argument was supposed to show, ultimately, was that we have a higher moral standing if there are constraints. That is, reformulating the problem of source-plurality in terms of the plurality of sources for high moral standing (rather than the plurality of sources for high levels of inviolability) should be enough to challenge the hyperinviolability account.

I believe that the proponent of the hyperinviolability account has a plausible response at hand. First, it seems that what the inviolability account as an approach to the paradox of deontology was supposed to establish is that a moral system that does not allow for minimising violations, *other things being equal*, grants us a higher moral standing than one that makes such violations permissible. That means, provided that the fact that minimising violations are permissible within the moral system *S* and impermissible within the system *H* is *the only relevant difference* between *S* and *H*, *H* holds that we are more valuable beings, morally speaking, because it holds that we are hyperinviolable. The scenario mentioned earlier in which *S* protects us from being killed in the *Switch* case whereas *H* does not was therefore begging the question. For the question was whether *S* reflects a higher moral standing than *H*

if the only difference between them is that *H* includes constraints whereas *S* does not. And here, the answer seems to be *Yes*.[118]

Second, there is a sense in which the moral theory *S*, as it was described earlier, does not seem to constitute a coherent kind of view. Recall that S holds that it is impermissible to redirect a trolley set off by accident from five to one and that, at the same time, it is permissible to directly kill one to prevent five killings. But it seems much worse to harmfully use someone, directly killing them for the sake of others, than to kill that person as an unintended side effect of directing a threat away from five. How can it be permissible to harmfully use and yet impermissible to kill the one as an unintended side effect? A defender of *S* could respond that since killings are worse than accidental deaths, we should go to greater length to prevent them. Thus, it would make perfect sense to think that we should do more, maybe even harmfully using someone, to prevent killings. However, this would only explain why we should not harmfully use one to prevent five accidental deaths, not why we should not even redirect a threat to prevent five accidental deaths.

In both cases, the *Switch* case and the paradox case, the agent would do something to one innocent person thereby preventing something comparably bad happening to a greater number of other innocents. The only difference is that killing the one in *Switch* is an unintended side effect but killing the one in the paradox case means harmfully using them. Since harmfully using someone is worse than killing them as an unintended side effect, there is no rationale for holding the first kind of treatment impermissible but the second kind of treatment permissible.

This illustrates that the ways in which a moral theory holds us to be inviolable are not subject to random choice. There must be some rationale for which treatments are permitted and which are not. If we are inviolable to

---

[118] Note that this answer is preliminary. The saveability challenge I will address shortly puts precisely this answer in question.

the extent that it is impermissible to even redirect threats to us and kill us as an unintended side effect of saving many, then it seems we must also be inviolable to the extent that it is impermissible to harmfully use us to save many from being harmed in a similar way. The impermissibility of redirecting a threat in *Switch* indicates that *S* is a highly restrictive moral theory. It can only be a plausible kind of theory if it is also restrictive in the sense that it makes killing in paradox cases impermissible. If we are not hyperinviolable, how can we be inviolable in conceptually even more demanding ways?

Thus, in order to justify constraints, all we have to show is that, *other things being equal*, persons have higher moral value if their rights are constraining. All we have to argue for is that, *other things being equal*, being hyperinviolable makes one more valuable than if one lacks that status. The hyperinviolability account provides precisely this kind of argument.

4.3. Measures of Moral Importance

The hyperinviolability account explains why, just considering the ways in which we were inviolable, we would have a more elevated moral standing if we were hyperinviolable than if we were not hyperinviolable. Yet, as already mentioned, there is another way in which we might turn out to be more important, morally speaking, if we did *not* possess hyperinviolability status.

Before turning to the saveability challenge, let me mention one other point Kagan makes against the inviolability account—and Kamm's response to it. Kagan argues that if my action were not limited by constraints, "then there would be a sense in which I would be revealed to be a more important type of creature" (Kagan 1991: 920). For if it were permissible to perform minimising violations, we would have greater freedom to act according to our own conception of the good. That this greater freedom would come at the

expense of limiting our inviolability might, in and by itself, not show that we would be less valuable overall if we were free to that greater extent.

Kamm's response is that it does not seem that I would gain anything *in terms of my moral standing* if I had this kind of greater freedom (Kamm 2001: 324–325). For one thing, it comes at the expense that I have lost my hyperinviolability. Others, too, would gain the same freedom to harm me. For another, I would not have to be *respected by others* to any greater extent because the freedom to harm would still countenance the freedom of others to resist being harmed. Simply put, having greater moral freedom to harm others does not mean that it is impermissible for those others to do anything other than resigning themselves to their fate. It does not seem true, then, that a greater freedom to harm others would mean an increase in our moral standing as individuals.

I now turn to the saveability challenge. The challenge sets off from the thought that considering other dimensions of our moral standing than inviolability, it is unclear why we should think that we are morally more important if there are constraints than if there aren't any constraints. As mentioned earlier, the saveability challenge has been put forward in three different versions.

4.3.1. Saveability, Unignorability, Enforceability

The first version of the saveability challenge has been formulated by Shelly Kagan. Kagan admits that if "I am protected by a constraint, then there is a sense in which I must be a more important type of creature" (Kagan 1991: 919). However, he also notes that there are other measures of moral importance. For instance, we are revealed to be more important creatures if others must *save us* from harm (Kagan 1991: 920). If there are constraints, then we are moral *inviolable*. But since constraints make it impermissible under certain circumstances to save us from harm, we are less *saveable* if there are constraints. Thus, constraints would "express a greater importance in one

way, but a lesser importance in another" (Kagan 1991: 920). They would reflect greater *inviolability*, but lesser *saveability*.[119]

This, as Kagan notes, might render the inviolability account—and consequently also the hyperinviolability account[120]—insufficient to show that constraints would give expression to an overall higher moral standing because constraints would limit the set of circumstances under which others must save us from harm. Assuming that both inviolability and saveability are moral statuses and, as such, measures of our moral importance, it seems inadequate to claim that greater inviolability would mean greater *overall* moral importance since it would come at the expense of lesser saveability. Which of these two dimensions of moral standing—inviolability or saveability—would reveal us to be *most* important, so Kagan, remains an open question (Kagan 1991: 920).

Kasper Lippert-Rasmussen has followed up on Kagan's criticism. He introduces a moral status he calls *unignorability*. Unignorability is the status we possess if and only if "there are circumstances in which it is impermissible for others to *allow* [us] to be harmed" (Lippert-Rasmussen 1996: 340). Essentially, his following argument runs parallel to Kagan's: If we are protected by constraints, then we are less unignorable because the set of circumstances in which others must not allow us to be harmed is smaller than it would be if there were no constraints. Thus, constraints do not seem to reflect an *overall* higher moral standing.

However, there are two ways in which the notion of unignorability sharpens Kagan's challenge. First, unignorability represents a more specific status than saveability, defined broadly as the status of beings whom other

---

[119] Although the saveability objection is commonly attributed to Kagan, the term *saveability* seems to have been coined by Kagan's recipients (Otsuka 1997: 204, Kamm 2001: 275).

[120] Of course, Kagan (and other constrains-sceptics) have raised this as an objection to Kamm's inviolability account, and I will describe it in that way. However, the objection targets my hyperinviolability account to the same degree.

must aid. Individuals can be said to be saveable in a great many ways, for instance, if it is required to provide first aid to them after an accident. In contrast, the concept of unignorability mirrors the concept of inviolability: the latter concerns the ways in which other *must not harm us* and the former the ways in which others *must not allow us to be harmed*. In other words, inviolability and unignorability are parallel concepts, both concern harmful actions with the only difference that one uses the modifier *do*, the other uses the modifier *allow*. As such, the concept of unignorability provides the basis for a more targeted criticism against the inviolability account.

Second, Lippert-Rasmussen defines the alternative moral status not in terms of what others are *permitted* to do but what they are *required* to do. Again, compare the moral systems *M* and *H*. *M* holds that committing minimising violations is permissible whereas *H* makes them impermissible. Just considering this feature, *M* holds that saving us is sufficiently valuable so as to justify some killings. But *M* does not hold that we are so important that we *must* be saved. Lippert-Rasmussen argues that the lack of constraints on *M* is not sufficient to make us more important in any significant sense. Instead, we are more important only if it is *required* to save us even through minimising violations (Lippert-Rasmussen 2009: 174).

As it stands, this claim might be false. After all, if compared to *H*, *M* holds that saving us is a sufficiently valuable cause to justify the killing of another person, and this must mean that *M* holds us to be important in some sense. Even though it is true that a moral theory, *M'*, which would make saving us in the relevant type of circumstances a requirement would hold us to be even more important than *M*, it doesn't seem to be true that *M* would not, as Lippert-Rasmussen seems to think, hold us to be more important to *any* notable extend. However, I agree with Lippert-Rasmussen that for the saveability challenge to arise, we need to assume that persons could be saveable in a very specific sense. As I will argue in Section 4.3.2, the saveability challenge requires that we take persons to be saveable in the sense that they

should at least sometimes be saved even where this would require the agent to commit a minimising violation.

Finally, Michael Otsuka has reformulated Kagan's criticism using the terminology of rights. Otsuka begins with the idea that rights against interference—for instance, a right not to be killed—come with a sort of meta-rights which entitle the right holder to prevent these rights from being violated (Otsuka 2011: 52–53). For instance, I have not only a right not to be killed but also a right to defend myself against an attack on my life. According to Otsuka, rights of the second kind can be understood as *enforcement* rights: they are rights *to enforce* one's rights against interference.

In general, enforcement rights do not seem to be less important than the rights they permit to enforce. Suppose that I have a right against torture, but I have no enforcement right that comes with it. That is, I am not permitted to resist someone else's attempt to torture me. Essentially, I must keep still and let it happen or else, if I defend myself, I do what morally I have no right to do. This would be strange. But even more, it would seem to put the value of my right against torture itself into question. For what is the value of a right against being tortured if such a right does not entitle the right holder to enforce that right to any, even so menial extent?

Moreover, Otsuka argues that enforcement rights should be transferable. If I am restrained and cannot defend myself against torture, it seems that the right to enforce my right could fall into the hands of any capable bystander who can prevent the torturer from going ahead. There is no reason to think that others may never enforce our rights, or even be required to do so, where they can do so with little or no risk to themselves. The idea that there are enforcement rights of this kind gives rise to the following challenge:

> Why, then, should not the right to enforce our rights against interference in a manner that involves the harming of innocent bystanders also constitute an increase [in our moral standing]? (Otsuka 2011: 53)

In a nutshell, one dimension of our moral importance is embodied in the *enforceability*[121] of our rights such that if minimising violations are permissible or required, then this means an increase in *this* dimension of our moral importance. Just like the saveability and unignorability challenges, the enforceability challenge is supposed to render the inviolability and hyperinviolability accounts insufficient: it is not obvious why constraints should reflect an *overall* higher moral worth once we understand that our constraining rights are less enforceable.

4.3.2. Hypersaveability and Hyperenforceable Rights

I take it that the three challenges set out above are versions of one and the same criticism. What Lippert-Rasmussen and Otsuka propose are different ways to cash out Kagan's internal criticism against the inviolability account. Admittedly, saveability, unignorability, and enforceability clearly describe different moral statuses. They are not identical. But they are all *specifications* of one and the same dimension of our moral worth, which describes the extent to which others must go to prevent the bad things that might happen or be done to us.[122]

I shall focus on Otsuka's version of the criticism because to me this seems to be the strongest version. Whereas saveability and unignorability can be clearly separated from inviolability, inviolability and enforceability seem to be *two sides of the same coin*. If we have rights against interference, it seems unreasonable to think that these rights are inviolable to some extent, but not

---

[121] Again, the term seems to have been coined later; see Burri (2017): 621.

[122] As such, it should be noted that the challenges set out in Section 4.3.1 cannot be combined to form the following kind of super-challenge: If minimising violations are required, then we are more saveable, more unignorable, *and* our rights are more enforceable, which would mean that we are morally more important in three different ways. Although saveability, unignorability, and enforceability are defined distinctly, when it comes to the lack of constraints, they all mean the same: the status we have if minimising violations are required.

enforceable to any extent. Inviolable rights *are* enforceable rights. From this perspective, the initial premise of the enforceability challenge seems hard to resist.

Yet as a matter of terminological choice, I henceforth use the term saveability to refer to the dimension of our moral importance that increases or decreases with the extent to which our rights are enforceable. This might seem like an odd terminological choice. However, it is necessary to use some other term than enforceability to talk about the relevant moral status. Whereas it makes sense to speak both of the inviolability *of rights* and the inviolability *of persons*, it does not make sense to speak of the enforceability of persons, only of the enforceability of rights. Thus, if we want to be able to talk about the moral status we have if our rights are enforceable, we need an alternative notion. And here, I choose Kagan's notion of saveability.

Moreover, note that this terminological choice is justified only because I focus on the moral standing *of persons*. There are a great many entities who are saveable, but who lack the moral status defined in terms of the enforceability of rights. For instance, a sunflower might be said to be saveable if we must sometimes water it. (It may be said to be unignorable if we must sometimes prevent others from trampling it.) But sunflowers lack the status defined in terms of rights-enforceability so long as we do not want to say that the sunflower *has rights*. For an entity to have *enforceable* rights, it is of course necessary that the particular entity has rights in the first place. In contrast, persons have rights. And I take it that it makes perfect sense to say that our saveability increases with the extent to which our rights are more enforceable.

Note that rights can be *enforceable* in various senses and not in all these senses does hyperinviolability conflict with greater enforceability. For instance, suppose that *X* is required to treat *A* as follows:

Normal Case. *X* should not kill *A.*

Special Case. *X* should not kill *A* even to prevent the killing of *B* and *C.*

Normal Prevention Case. *X* should prevent *Y* from killing *A.*

A's right not to be killed is inviolable to the extent that *X* must not kill *A* in the *Normal Case*. Even more, *A*'s right is a *constraining* right since it is impermissible to kill *A* in the *Special Case*—from here on, I will also refer to constraining rights as *hyperinviolable* rights.[123]

However, as the *Normal Prevention Case* tells us, *A*'s right is also enforceable in a certain sense. If *Y* is about to kill *A*, *X* should enforce *A*'s right by preventing *A*'s killing. The sense in which *A*'s right is enforceable does not conflict with the fact that *A*'s right is a constraining or hyperinviolable right.

Thus, for the saveability challenge to arise we need a very specific sense in which our rights could be enforceable. They could be enforceable in the sense that in order to enforce them, we should sometimes even commit minimising violations of these rights. I shall call rights that are enforceable in this sense *hyperenforceable* rights. Thus, a right *R* is *hyperenforceable* if and only if we should sometimes violate *R* to prevent more extensive violations of *R* in everyone.

Hyperenforceable rights are the counter concept to constraining or hyperinviolable rights. If hyperinviolable rights give expression to a particular moral status—which I have called hyperinviolability—then hyperenforceable rights give expression to another moral status, which I shall call *hypersaveability*. Thus, we are *hypersaveable* if and only if others should sometimes commit minimising violations of rights to save us from harm.

For Otsuka's challenge to arise, we need to assume that our rights could be hyperenforceable. The challenge then goes like this: if we have *hyperinviolable* rights, then we are beings of higher moral worth in terms of the

---

[123] This is a mere stylistic choice because very shortly, I will introduce the concept of *hyperenforceable* rights and contrast them with *hyperinviolable* rights. But it should be noted that hyperinviolable rights is nothing more than another name I use for constraining rights.

things that others may not *do* to us. But if we have *hyperenforceable* rights, then we are beings of higher worth in terms of the things that others may not *allow to be done* to us. Since rights cannot be both hyperinviolable and hyperenforceable, and both properties of rights make us more important in *some* sense, it is not obvious that having hyperinviolable rights would mean an *overall* increase in our moral standing.

4.4. The Saveability Dilemma

I think that critics of the inviolability account rightly insist that there is a dimension to our moral worth that cannot be captured by the notion of inviolability, and which is defined in terms of what others may not *allow to be done* to us. In particular, I believe Otsuka rightly insists on the enforceability of our rights as being an important signifier of our moral worth.

However, I believe that the saveability challenge ultimately fails.[124] I think that the claim that being hypersaveable (or having hyperenforceable rights) would mean an increase in some dimension of moral worth leads into a dilemma: either it is false, or it is based on a misconception of the moral worth of persons that is unintelligible as an account of individual moral status. On the first horn of the dilemma, the saveability challenge simply fails. On the second horn, it fails as an *internal* criticism for, as we have seen, the saveability challenge was meant to accept the inviolability account's initial premise: that constraints are grounded in the value of having a certain moral status. My argument begins with a comparison between two morally important individuals called *Jen* and *Joe*.

---

[124] Burri (2017) also argues for this conclusion. She uses an analogy between personal sovereignty and state sovereignty to show that the authority each of us enjoys over the territory of her body and mind—to stay within Burri's metaphor—leaves conceptual room only for enforcement rights held against liable attackers, but not for enforcement rights held against innocent third parties (Burri 2017: 628-29). I shall not discuss Burri's proposal in any more detail here. My response to the saveability challenge will take a different route.

4.4.1. Who Is More Important?

Imagine two separate worlds. In the first world lives Jen. Jen is so important morally that others should not violate Jen's basic rights even to minimise violations of the same rights in everyone. In the other world lives Joe. Joe is also morally important, but in a different way. He is so important that others should even perform fewer rights violations to prevent the same right from being violated in Joe.[125] In short, Jen is *hyperinviolable*, whereas Joe is *hypersaveable*.

Who is morally *more* important? The upshot of Otsuka's challenge—as well as Kagan's and Lippert-Rasmussen's points—is that we cannot say for certain. The question who is more important depends on a further question about the comparative importance of having hyperinviolable or hyperenforceable rights. At least we cannot say for certain that Jen is morally more important than Joe; that she would have a higher moral standing than he does. I think this is false. We *are* justified in saying that Jen is morally more important than Joe. Here is why.

Suppose that *X* is required to kill *A* in the *Special Case* where this prevents the killing of *B* and *C*. As far as the saveability challenge goes, this would have to mean that there is a sense in which *B* and *C* are morally more important than they would be if it was not required to kill *A*. *B* and *C* would be *hypersaveable*. And to be hypersaveable means that one is more important in terms of the things that others may not allow to be done to us.

But is this true? Suppose that Jen's and Joe's worlds are as (much like in ours) such that each of them could potentially end up in a scenario that resembles the *Special Case*. And they could end up in any of the three

---

[125] This might sound a little awkward. In fact, there is no less awkward formulation of the sense in which Joe is morally important. This is so because in order for the rights violations others might be required to commit to be *fewer in number*, the set containing the violation of Joe's right must contain at least one other violation. So, talking about Joe's moral importance is always also talking about the moral importance of at least one other person like Joe. In Section 4.4.2, we will see that this invites some doubts about hypersaveability as a candidate for *moral status*.

positions of *A*, *B*, or *C*. Considering this, the saveability challenge suggests that the ways in which *X* may or may not treat Jen and Joe are as follows:

	Jen	Joe
1.	*X* must not kill *A*-Jen.	*X* must kill *A*-Joe.
2.	*X* may not save *B*/*C*-Jen.	*X* must save *B*/*C*-Joe.[126]

Jen is hyperinviolable and thus must not be killed if she ends up in *A*'s position. But this comes at the expense of it being impermissible to save her if she ends up in the position of either *B* or *C* (since saving them is only achievable if *A* is killed). By contrast, *X* is required to kill Joe if he ends up in *A*'s position. But if he ends up in the position of either *B* or *C* he must be saved. According to this picture, it really seems hard to see why Jen should be morally more important than Joe. Solely focusing on the ways in which the two may be treated by *X*, it seems that morality assigns importance to both in some relevant sense.

However, the above picture is incomplete. I believe there is more information on how Jen and Joe may be treated. The complete list should look like this:

	Jen	Joe
1.	*X* must not kill *A*-Jen.	*X* must kill *A*-Joe.
2.	*X* may not save *B*/*C*-Jen.	*X* must save *B*/*C*-Joe.
3.	*X* must save *A*-Jen.	*X* may not save *A*-Joe.

[126] All these deontic properties are limited and thus each line should be read as containing the word 'sometimes'—it might be that there are instances of the *Special Case* in which *A*-Jen may be killed. For instance, moderate constraintists might argue that *A*-Jen may be killed to save one hundred others. Similarly, there might be instances of the *Special Case* in which it is permissible, at least, not to save *B*/*C*-Joe. For instance, one might think that Joe may not be saved as part of a group of one hundred individuals if this requires that other ninety-nine individuals are harmed.

The added third line contains an interesting piece of information. I will argue that once we take this piece of information seriously, we must come to realise that *there is a sense* in which Jen is *more* saveable than Joe and that Joe's hypersaveability does not seem to make him *overall* more saveable than Jen. This is so because if *A*-Jen must not be killed in the *Special Case*, then she must be saved. And if *A*-Joe must be killed in the *Special Case*, then he may not be saved.

The truth of these claims about the saveability of Jen and Joe might not be evident at first glance. But recall the *Special Prevention Case* from the discussion around agent-neutral constraints in Chapter 2. Imagine that Bruce is about to kill *A* to prevent Joker from killing *B* and *C*, and you must decide whether to hold Bruce back or allow for him to kill *A*. Imagine further that the person referred to as "*A*" here is hyperinviolable Jen. In this case, Bruce may not kill *A*. *A*-Jen is protected by a constraint on killing that renders Bruce's potential act of killing impermissible. As we have seen in Chapter 2, agent-neutral constraintists should want to say that you should (or that you have strong reasons, at least) to prevent Bruce from using *A*-Jen to save *B* and *C*. Using *A*-Jen in this way is wrong, and it is within your power to prevent this wrong.

Now imagine the person behind "*A*" is hypersaveable Joe. In Joe's world, people are not protected by constraints, but it is more often permissible or even required to save them. You know that Joe is hypersaveable, thus should be saved if he is in the position of *B* or *C*. But what does Joe's saveability demand you do if he is in the position of *A*? What does *A*-Joe's saveability imply you should you do if Bruce is about to kill him for the sake of *B* and *C*?

I think there are two options here. Either morality allows (or even requires) that you save Joe from Bruce, or it doesn't. In the first case the moral

system in Joe's world is directly self-defeating.[127] It would be a moral system that tells you to achieve some goal *G* and, at the same time, to prevent the achievement of *G*. In the second case, if morality tells you *not* to save Joe, the moral system in Joe's world is not self-defeating. But there is something else that is worrisome about that moral system: it embraces the *impermissibility* of saving *A*-Joe. As such, it holds that there is a sense in which Joe is *less* saveable than Jen. *A*-Jen must be saved, but *A*-Joe may *not* be saved.

To make the point more generally: there seem to be two dimensions to our saveability. One is expressed in the status of hypersaveability. In one sense, we are *more* saveable if it is required to commit minimising violations of rights. But there is a second dimension to our saveability, one which *decreases* once it is required to commit minimising violations. This is so because such a requirement entails the impermissibility of saving the victims of minimising violations. A moral system that demands someone's killing, if it should not be directly self-defeating, must hold that someone to be *less* saveable in this sense.

Note that these considerations do not solely depend on the presence of the bystander in the *Special Prevention Case*. Imagine the following scenario:

> Time Bomb. A villain has tied two strangers to the tracks, started an electric trolley that is now heading towards them, and afterwards fled the scene. You could save them, but only by blowing up a house close to the tracks such that the debris would block the tracks. A third stranger is sleeping inside the house and would die in the explosion. You only have a time-action bomb at hand. But if you plant the bomb

[127] In Parfit's sense, a moral theory that tells you to save Joe would be *directly individually self-defeating*. A theory *T* is directly individually self-defeating if "it is certain that, if someone successfully follows *T*, he will thereby cause his own *T*-given aims to be worse achieved than they would have been if he had not successfully followed *T*" (Parfit 1984: 55).

> right now and put the timer on the lowest possible value (ten seconds), the house will blow up just in time for the debris to block the tracks.

The important detail of *Time Bomb* is that you have ten seconds to reconsider. Suppose the person sleeping in the house is hyperinviolable Jen. In this case, you may not plant the bomb. You may not kill *A*-Jen even to prevent two further killings. But suppose that you plant the bomb anyway. You now have ten seconds to defuse it—and that is exactly what you should do. Given your initial wrongdoing, you should *save A*-Jen from certain death.

But now suppose the person sleeping in the house is hypersaveable Joe. You should then plant the bomb—morality demands that you kill *A*-Joe to prevent two further killings. But after you plant the bomb, you get to think. Should you really let it happen that *A*-Joe is killed in his sleep? Or should you save him? Unless the moral system in Joe's world is directly self-defeating, it seems that you should let it happen that Joe is killed. And as far as I can see, this would have to mean that there is a sense in which *A*-Joe is less saveable than *A*-Jen because *A*-Jen should be saved but *A*-Joe should not.

What these cases show, in other words, is that *A*-Jen possesses a valuable kind of saveability that *A*-Joe lacks. Call this kind of saveability *constraints-saveability*. If there are constraints, then others must sometimes save us from becoming the victims of minimising violations of rights. Joe might be hypersaveable, but he lacks constraints-saveability. Jen might lack hypersaveability; but she has constraints-saveability.

To sum up the discussion, the impermissibility to allow for *B* and *C* to be killed would not reflect merely positively on the saveability dimension of our moral worth. Provided that it is true that the impermissibility of allowing for *B* and *C* to be killed *entails the permissibility* of allowing for *A* to be killed, we would be more saveable in *one sense*, but less saveable in *another sense*. For the case introduces both a type of circumstance when it is impermissible

to allow us to be harmed and a type of circumstance when it is permissible to allow us to be harmed.[128]

If this picture is plausible, then Jen's and Joe's moral standing can be compared as follows:

Jen	Joe
A: more inviolable	*A*: less inviolable
A: more saveable	*A*: less saveable
B/*C*: less saveable	*B*/*C*: more saveable

We can now see that the only sense in which Joe seems clearly more important than Jen is that Joe must be saved if he ends up in the position of *B* or *C*. Jen is overall more inviolable than Joe because, unlike him, she must not be killed if she ends up in the position of *A*. (Recall that the fact that *B* and *C* will be killed if *A* is not killed does not entail that killing them is permissible.)

The saveability challenge stated that Joe must be overall more saveable than Jen, and that it is therefore not obvious that Jen is overall more important than Joe. But we can now see that the question whether Joe is *in fact* overall more saveable than Jen depends on a further question about the relative importance of two measures of saveability—hypersaveability and constraints-saveability.

Although Joe is hypersaveable, it is not obvious that Joe is *more* saveable than Jen, i.e., more important in terms of the things that others may not allow to be done to him. Once we consider not only *B*'s and *C*'s saveability but also the saveability of *A*, we see that there is one sense in which Joe is

[128] It is worth noting that there is no plausible argument about hyperinviolability that would run along the same lines, showing that if we are hyperinviolable, we are more inviolable in one sense, but less inviolable in another sense. If *A*-Jen is hyperinviolable, then this does not mean that *B*/*C*-Jen is any less inviolable because the impermissibility of violating *A* does not entail the permissibility of violating *B* or *C*.

more saveable than Jen, and another sense in which he is *less* saveable than her.

It is not obvious, then, that hypersaveability entails an increase in our *overall* saveability, and thus not obvious that being hypersaveable makes us more important in terms of the things that others may not allow to be done to us. If we are hypersaveable, however, we clearly are less important in terms of the things that others may do to us. Thus, the claim that hypersaveability might mean an increase in our moral worth comparable to the level of moral worth we had if we were hyperinviolable, it seems now, is false.[129] This is the first horn of what I shall refer to as the *saveability dilemma*.

4.4.2. Saveability in Numbers

It seems that there is only one way to escape the first horn of the saveability dilemma. That is, there is only one way to maintain that Joe is *overall more* saveable than Jen. The proponent of the saveability challenge would have to refer to the idea that Joe is saveable *in the position of B* and saveable *in the position of C*. *This* is what makes Joe *more* saveable: that he must be saved in

[129] Note that the argument of this section takes the same form as the saveability challenge itself. The central insight of the saveability challenge was that once we broaden our perspective on moral worth, we see that the impermissibility of minimising violations has further implications; it would only make us more valuable in one sense, but less valuable in another. Parallel to this, the central insight of my argument is that once we broaden our perspective on personal *saveability*, we see that the requirement to minimise violations of rights has further implications; it would only make us more saveable in one sense, but less saveable in another. The parallel structure of my argument should make it more difficult for advocates of the saveability challenge to refute it without questioning the validity of their own challenge.

two possible positions but may not be saved in only one possible position. In other words, we are *more* saveable if *more of us* are saveable.[130]

Can this be true? Suppose we would try to establish the same idea about the concept of inviolability. Imagine that both Jen and Jas would have to be killed to prevent two sets of other killings, containing two killings each. If it is impermissible to kill either of them, does this mean that Jen is any *more* inviolable than she was before? Is she any more morally important than she was when it was impermissible to kill her to save two others? It does not seem so. The thought that Jas, too, may not be killed does not make Jen any more inviolable, and vice versa. So long as Jas inhabits the same world as Jen, the fact that Jas may not be killed is already implied in the idea that Jen may not be killed. For hyperinviolability is a moral status that Jen has qua being a person and not for any contingent facts about herself. Thus, if Jen is hyperinviolable, then so is Jas.

Now suppose Joe and Jon must be saved from being killed even if this means that the agent must kill a third person. Why would the thought that Jon, too, must be saved make Joe any *more* saveable? Again, so long as Joe inhabits the same world as Jon, that Jon must be saved under this circumstance is already implied in the idea that Joe must be saved under this circumstance. Hypersaveability—if it should be a moral status—is a status that Joe has qua being a person and not for any contingent facts about himself. Thus, if Joe is hypersaveable, then so is Jon. No additional type of circumstance has been introduced that would increase the set of circumstances under which Joe must be saved.

Thus, the fact that Jon, too, must be saved does not seem to make Joe any more saveable, or vice versa. We are simply not more saveable *as*

---

[130] This must be the thought at the heart of the saveability challenge (although it is not made fully explicit by Kagan and the others): that Joe is overall more saveable than Jen because it is a significant fact that Joe is more likely to end up in the position of *B* or *C* than in the position of *A*. He is more saveable because he is saveable in a greater number of possible positions.

individuals, if only more of us are saveable.[131] Using the enforceability-terminology: our rights are simply not more *enforceable*, if only they must be enforced *for a larger number of people*.

This illustrates that there is something puzzling about the moral status called hypersaveability. In Joe's world, it seems that—as individuals—each of us could be *called* hypersaveable; but whether that means that we must be saved or not is conditional on the further question whether we are part of the smaller or the larger group. Effectively, we are hypersaveable only as *members of the larger group*. That Joe is part of the larger group in a paradox case is a precondition for it to be true that Joe must be saved. If he were on his own, I assume proponents of the saveability challenge would not want to say that he must be saved. That is, I assume they would not want to say that we are required to kill someone to prevent a single other killing.

In fact, Kamm voices a similar worry about saveability as a moral status:

> The status of persons qua persons is a function of what is true of any one person. If you should be saved simply because you are in a group with more people, this does not indicate that you or the others *as individuals* have higher saveability, but only that the numbers of people could affect what we should do. Strictly speaking, a status of high saveability would have to show up as a duty to do a great deal to save *any one* person. (Kamm 2007: 254–255 [emphases added]).

---

[131] Just the same, we are not increasingly inviolable—according to how I understand the concept of inviolability—if it is impermissible to violate any one of us to prevent an increasing number of other violations. There is a sense in which the numbers matter for the question how inviolable we are. For we are inviolable (to a certain extent) if it is impermissible to violate any one person to prevent the violation of *one other* person. But if it is impermissible to violate any one person to prevent the violation of *two or more others*, then, by definition, we are hyperinviolable. However, it would be a mistake to think that this means that our (hyper)inviolability would increase further if the set of violations which may not be prevented increases. Hyperinviolability is the status of individuals who may not be sacrificed for the (supposedly) greater good of saving others. It is *not* the status of being more important than *so and so many others*.

If hypersaveability should be a moral status—i.e., a value that we have *qua* persons but *as* individuals—it seems puzzling that our hypersaveability should depend, in one way or another, on whether we are part of the larger group. To say that Joe has a certain kind of moral status but that it *effectively* entitles him to some form of respect only in the case that he shares his fate with a sufficiently great number of others comes up against the very idea of moral status itself.

Proponent of the saveability challenge might interject that our hypersaveability—the fact that we have that particular moral status—does not depend on any facts about groups. I am hypersaveable, even if I am the only person on the planet. In this case, no requirement to save me will ever ensue *from that status* (because it is impossible for there to be a situation in which someone could save me—I am the only person on the planet). But this does not show that I am not hypersaveable, neither does it show that hypersaveability is not a suitable candidate for a moral status.

In short: I agree. I still believe that there is something puzzling about the idea of there being a moral status such that prescriptions or prohibitions would ensue from that status only if the person having it is a member of the larger group. But my argument does not depend on this puzzle. Instead, what I am arguing is merely that it cannot be true that we are *more* saveable, if only *more of us* are saveable. This is the idea that proponents of the saveability challenge need in order to avoid the first horn of the saveability dilemma. But it is an idea that renders the notion of saveability unintelligible as an account of individual moral status. Just suppose that it is required to kill *A* not only to save *B* and *C*, but also to save *B*, *C*, and *D*, and to save *B*, *C*, *D*, and *E*; and so on. It does not seem plausible that anyone's saveability increases with the number of individuals we could add to this case.

4.4.3. The Complaint Model

Before concluding the present chapter, I want to point to one last way in which proponents of the saveability challenge might be able to make sense of hypersaveability as an account of individual moral status. They could argue that if we are hypersaveable, what this means is that our additional presence in the *Special Case* would make a difference to the normative situation, which it does not make if we lack hypersaveability status. Hypersaveability is the moral status of individuals whose moral importance shows up in the change of the deontic property of killing *A* from impermissible to permissible or required.

One possible route to argue in this way would be to suggest a Scanlonian picture of hypersaveability. Scanlon has famously tried to account for the significance of numbers in rescue cases, without using any consequentialist methods of aggregation. He intends to do so by employing a *complaint model of moral justification*:

> [The complaint model] calls attention to a central feature of contractualism [...]: its insistence that the justifiability of a moral principle depends only on various *individuals'* reasons for objecting to that principle and alternatives to it. (Scanlon 1998: 229)

Imagine that you could save Jim from serious harm, at little or no cost to yourself. Presumably, you *should* save Jim. Now imagine that you could save either Jim or Jay, but not both, from serious harm. Again, you could do so at little or no cost to yourself and the harms they would suffer are roughly equal in magnitude. Adding Jay to the situation turns a situation in which you should save Jim into a situation where it is permissible to save either of them (Scanlon 1998: 232), or—perhaps more plausibly argued—into a situation where you

should apply a fair method to decide which of the two to save (Timmermann 2004: 108).

Now, in a third step, imagine you could either save Jim or save Jay *and Joe* from suffering serious harms roughly equal in magnitude and at little or no cost to yourself. It seems now that the fact that adding Jay to the situation gave you a reason that balanced the reason to save Jim, adding Joe should have some *outbalancing* effect. The additional reason presented by the needs of a second person in one of the two groups "must at least have the power to break [the] tie" between the conflicting reasons you have for saving Jim or Jay (Scanlon 1998: 232). Joe might reasonably reject any moral principle that implies that his additional presence does not change the normative situation because if there is no normative difference between the situation in which you could save either Jim or Jay and the situation in which you could save either Jim or Jay *and Joe*, then it would seem that his claim to be saved had no moral weight at all.[132]

Scanlon focuses on rescue cases. However, a similar argument could be made for paradox cases. That is, if you may not kill Jim to prevent the killing of Jay and you may not kill Jim to prevent the killing of Jay *and Joe*, then Joe might complain that his additional presence made no difference to the

[132] By changing the normative situation I do not mean that adding Joe to the case must make it *required* to save the two rather than the one. What Scanlon seems to be getting at is rather that adding Joe must change *something* about the normative situation; for instance, it must at least change the procedure we should employ to decide whom to save such that Joe's additional presence weighs in and makes it more likely that the two are saved.

normative situation—that although he is morally important, his importance does not show up in this case.[133]

On the other hand, if Joe (and Jay) must be saved, once we add Joe to the situation, then it seems that there is a sense in which Joe's moral importance shows up in the requirement to have someone else killed. Hypersaveability, then, is not the moral status Jay and Joe enjoy only as a group but the status of individuals whose additional presence in a paradox case has the power to change the normative situation. Of course, Jay and Joe both are hypersaveable—but they have hyperinviolability status *as individuals*.[134] And it does seem that there is a sense in which Joe is *more* important if his additional presence in the *Special Case* has the power to make it required to save him and Jay.

Does this resurrect the saveability challenge? One could understand Scanlon's model of trying to give the claims of all individuals proper weight as a minimal requirement to respect their moral value. To say that we have given the claims of all individuals equal and proper weight is a minimal requirement for saying that they have each counted in their own right.[135]

Yet even if we can make sense of hypersaveability as an account of individual moral status in this way, this would not help to escape the saveability dilemma. We do not seem to be *more* saveable simply because *more of*

---

[133] Whether this is a reasonable kind of complaint is a different question. Can Joe really expect that others are killed to prevent his killing, even as part of the larger group in a paradox case? It seems much less obvious that he can reasonably expect *this* than that he can reasonably expect it to be more likely that he is saved as part of the larger group in a rescue case.

[134] Of course, we could also think of Jay as the one added last to the situation. Note that, strictly speaking, it would not be appropriate to call hypersaveability status the status of the individual *as a tiebreaker* because there is no tie in the case where you may not kill Jim to save Jay.

[135] Note, however, that as several authors have argued Scanlon's proposal fails to deliver on this promise. Scanlon's model seems to simply rest upon an implicit appeal to the idea that the claim of the larger group outweighs the claim of the smaller group (Otsuka 2000: 291). Just like on a consequentialist account of aggregation, the members of the smaller group have little or no chance to being saved whereas the members of the larger group enjoy relative or absolute safety in numbers (Lang 2005: 329).

us are saveable. Joe is not any more saveable only because he is more likely to end up in the position of *B* and *C*. This draws the saveability challenge back into the grip of the first horn of the dilemma. For even if *B*/*C*-Joe is more saveable than *B*/*C*-Jen in the sense that his additional presence in the *Special Case* has the power to change the normative situation, *A*-Joe is less saveable than *A*-Jen because he must be killed to prevent two further killings and thus, in this sense, may *not* be saved. Again, the question who is *overall* more saveable—Joe or Jen—depends on a further question about the comparative importance of these two senses of saveability. And since Jen is overall more *inviolable* than Joe, it seems appropriate to think that she is overall more important than him.[136]

4.5. Conclusion

I do not see how the sense in which Joe is more saveable than Jen could be shown to be more significant than the sense in which Jen's saveability exceeds that of Joe. The only way to establish this, I have argued, is through an appeal to the numbers—an appeal to the idea that Joe is more saveable because hypersaveable beings may be saved in larger numbers. However, as I have argued, moral status is not number-sensitive in this way. Our individual rights are simply not more enforceable if they are enforceable in a greater number of individuals.

This does not mean, of course, that the appeal to numbers has no place in moral theory. The thought that it is good to enforce rights naturally seems to lead to the thought that it is *better* to enforce the rights of *more*

[136] I would like to mention, at least, that the argument of this section is born of necessity, not belief. I myself find the idea strange that we could calculate our overall moral worth by comparing the factors which seem to make us more or less morally important. To me, the idea that hyperinviolable individuals have an elevated moral worth has a more direct credibility. However, in order to counter the saveability challenge which is based on the idea that moral worth is calculable in this way, I have agreed to this idea merely to show that the proposed criticism fails on its own terms.

people. As we have seen, even on constraintist views we might be required to enforce the rights of a greater rather than a smaller number of people. The argument I have presented in this chapter is therefore not to be misread as an argument for number-scepticism.[137] Instead, I want to again draw attention to the fact that the saveability challenge set off as an *internal* criticism against the (hyper)inviolability account. Kagan and others accept the initial premise of the account that the value that justifies constraints is not the moral significance of what happens to us, but our moral significance itself, i.e., our moral standing as individuals. As an *internal* criticism, the saveability challenge leads into the dilemma I have presented: either its central claim is false, or it talks about something other than individual moral status.

Constraints-sceptics are free, of course, to retreat to an *external* criticism against the (hyper)inviolability account. They could argue that while moral standing is important, it is more important in paradox cases to prevent the greater number of rights violations. At the very least, it is not obvious that we should care more about our moral significance than about the significant things that could happen to us, especially where these things are *more likely* to happen to us. As we have seen in Chapter 3, this is a valid form of criticism. And it is a form of criticism which this thesis cannot fully dispel. Attempting to do so is, as I have argued, not a promising kind of task.

The much humbler task that I have taken on is to show that constraintism is a coherent kind of moral view—a view that is free of paradox. Once we understand that the morally significant things that might happen to us are not all that matters, we see that plausible moral theories may focus on other things that matter. Constraints can be justified if how we *should* act depends, first and foremost, on how *morally valuable* we were if we should act in certain ways. There is a plausible kind of moral theory that gives priority

[137] As we have seen, constraintists are not committed to number-scepticism; see Sections 1.3 and 3.2.3.

to our moral worth over the things that may happen to us. Constraintism is precisely this kind of moral view.

5 *The Constraints of Consequentialising*

5.1. Introduction

This chapter focuses on the relationship between constraintism and consequentialism. Some consequentialists—from here on referred to as *consequentialisers*—have argued that all originally non-consequentialist theories can be given a consequentialist reinterpretation. What is referred to as *consequentialising* is the operation of translating a non-consequentialist theory into a consequentialist theory by accommodating the deontic properties of the original non-consequentialist theory within the consequentialist framework of ranking outcomes from better to worse.

One might think that it is an interesting question whether the consequentialisers' claim holds true for constraintist views. As laid out before, constraintism is the view that the right action is one that produces the best possible outcome. Constraintism is the view that some actions are wrong *even though* they seem to produce the best possible outcome. It is therefore not obvious how consequentialism could make sense of constraintism. However, at a closer look, the question whether constraintism—or any other non-consequentialist theory for that matter—*can* be consequentialised will turn out to be of minor importance. Once we take the idea of consequentialising at

face value, it is trivial that all possible non-consequentialist theories *can* be given such a consequentialist reinterpretation.[138]

However, there is a difference between the claim that something *can be done* and the claim that it *should be done*. Consequentialisers generally do not think that we should consequentialise moral theories just for fun. Instead, they think that there is something *so compelling* about consequentialism that it is worth going through the trouble of consequentialising originally non-consequentialist view—instead of simply adopting these views, for example. On any charitable interpretation of what the consequentialisers are after, their suggestion that we should consequentialise moral theories is not simply an expression of a form of consequentialist fetishism. Instead, they must believe that consequentialists (and perhaps non-consequentialists) have *something to gain* from following that suggestion.

So, what could be gained if constraintism was consequentialised? As we will see, the consequentialising project holds certain promises for consequentialists. But in the context of this thesis, the most relevant promise is one that it holds for consequentialists and non-consequentialists alike. At first glance, constraints must appear paradoxical from the perspective of the good. For how could it ever be *better* not to kill or torture, where this would minimise killings or tortures overall? A plausible consequentialist account of constraintism holds the promise that it could clear the air of paradox surrounding constraintism.

And more than that: consequentialised constraintism holds the promise that, contrary to initial appearances, deontic constraints *can* be reconciled with the part of ethics that concerns the good and thus, with the powerful idea that what we should do, morally speaking, is to always make the world a better place. In this regard if there is a plausible consequentialist account of

---

[138] Consequentialisers sometimes say that all "remotely plausible" (Portmore 2007: 39) or "non-crazy" theories (Suikkanen 2009: 2) can be consequentialised. We will see later that this limitation is unjustified.

constraintism, then it might be preferable to the hyperinviolability account I have presented, which can avoid the paradox of deontology only at the expense that it departs from that powerful idea to some extent. For the hyperinviolability account holds that sometimes our actions should not change the world for the better, but instead fit a certain ideal of what kind of moral world we could live in.

It is therefore necessary that I examine the possibility of consequentialised constraintism as an alternative to the hyperinviolability account. I will argue that although it is *technically* possible to consequentialise constraintism, the resulting moral views are unacceptable. They are unacceptable, I argue, because once we implement constraints into a consequentialist-type theory, that theory refers to an implausible conception of value, and thus cannot avoid the value paradox.

Section 5.2 begins with some remarks on the compelling force of consequentialism and then examines the technical side of the consequentialising project, i.e., how consequentialisers intend to give non-consequentialist theories a consequentialist reinterpretation. Section 5.3 investigates the consequences of consequentialising for ethical theory and, in particular, for the idea that there *different kinds* of moral theories. Finally, Sections 5.4 and 5.5 demonstrate how the consequentialising operation would work for the different versions of constraintism: agent-relative, agent-neutral, moderate, and absolute constraintism, showing that none of the consequentialised views can escape the value paradox. Thus, consequentialised constraintism cannot answer the paradox of deontology.

5.2. The Force of Consequentialism

As we have seen, consequentialism is a family of ethical theories that concern the good. Traditional consequentialist views (like classical act-utilitarianism) are committed to three central ideas. First, the good can be defined in terms of a single core value (like pleasure, well-being, or happiness). Second, the

right act is one that produces the best possible outcome, where the *best* outcome is one in which the core value is optimally realised. And third, whether an act produces the best possible outcome does not depend on the agent's perspective. Everyone has reason to prefer the same kind of outcomes. What I have reason to prefer, morally speaking, is the same as what anyone else has reason to prefer. Thus, traditional consequentialism is committed to (1) *value-monism*, (2) *value-maximisation*, and (3) *agent-neutrality*.

More recent versions of consequentialism are best understood in virtue of the ways in which they diverge from traditional consequentialism. For instance, some consequentialists have recognised the implausibility of the value-monist claim that all values could only matter indirectly, i.e., insofar as they relate to a single core value. By contrast, *value-pluralist* consequentialism holds that we can identify a set of fundamental values and take all these into account when determining the best possible outcomes. *Agent-relative* (or evaluator-relative) consequentialism is the response to the objection that traditional consequentialism would not be able to account for agent-relative values, such as friendship. *Satisficing* and *progressive* consequentialism move away from traditional consequentialism's commitment to value-maximisation and hold that we should do what creates *enough* good or what makes the world a better place (compared to the place it would be if we did nothing); and so on.

For each of the three central commitments of traditional consequentialism, there is a novel version of consequentialism that does not include the relevant commitment. It is therefore useful to ask what unifies all these diverse accounts of morality to a single consequentialist tradition.

5.2.1. The Priority of Good

Traditionally, the distinguishing mark of consequentialism, as opposed to non-consequentialism, has been taken to be the focus on consequences or outcomes rather than the act itself, whereby an act's consequences or

outcome were understood as a description of what happened, *minus the fact* that the agent made it happen (Thomson 2001: 8, Muñoz 2021: 79).

However, consequentialists have pointed out that there is also a *broad* notion of an act's outcome.[139] An act's broad outcome is a description of what the world would be like if the act would be performed (e.g., Portmore 2007: 39, Portmore 2009: 330). Understood in this broad sense, there is no way of drawing a clear distinction between an act and its outcome. Instead, a full description of an act's outcome would, besides statements about its narrow consequences, also feature statements about the fact that the act itself is performed, that it constitutes a rights violation, or that it is done for a morally assessable reason. If this is a plausible notion of outcomes, then a minimal definition of consequentialism had better not focus on the idea that consequentialist theories concern consequences or outcomes *rather than* the act itself as the act itself can be understood as part of its own outcome.

Instead, it is more promising to distinguish consequentialism by virtue of its commitment to a certain conception of rightness, i.e., the conception of rightness *as a function of goodness*. Portmore, for instance, explains that:

> the ranking of outcomes is prior to the determination of an act's deontic status in that one must first determine how a set of alternative outcomes are ranked before one can determine whether the acts that produce them are permissible or not. (Portmore 2007: 44)

What unifies all consequentialist views is that they conceive of moral rightness as a function of a ranking of acts from better to worse. Whether an act

---

[139] Traditionally, consequentialisers have introduced the notion of broad outcomes as a response to the following type of criticism: It seems that there are acts that are clearly wrong, although they produce the best outcomes. For instance, killing the healthy patient in the *Transplant* case leads to one death instead of five. Yet who would want to deny that killing him is clearly wrong? Once consequentialists include the act itself in a description of its outcome, they could at least hope to be able to account for the badness of killing as opposed to letting five other patients die.

is right is determined solely by the relative goodness of that act—or its outcome, broadly construed—in comparison to the relative goodness of all alternative acts. Consequentialism, then, is constrained only by "the *structure of good*" (Broome 1991: 11). To be able to say that *φ*-ing is *right*, the consequentialist must be able—at the very least—to say that *φ*-ing is not worse than any alternative.

On this view, what distinguishes one version of consequentialism from another is simply how they each specify the ordering we should use to rank acts. It is "their differing commitments concerning what agents have most reason to desire, all things considered" (Portmore 2009: 334).

On a side note, sometimes consequentialists disagree on what the agent ought to do regarding the proposed ranking of acts. For instance, classical act-utilitarianism requires us to perform the highest-ranking act—or one of the highest-ranking acts, if more than one act is best. By contrast, satisficing and progressive consequentialism might hold that there is some threshold of goodness and that we are required only to perform an act that lies above that threshold (Slote 2020). We might be required to give to charity, for instance, but there might be a threshold above which contributing *more* becomes supererogatory.

5.2.2. Why Consequentialise Constraintism?

Not all ethical theories are called consequentialist. Some ethical theories are called *non*-consequentialist because they deny consequentialism's key claim that moral rightness is a matter of the relative goodness of act-outcomes.

However, thinking of the above minimal definition of consequentialism, one might be sceptical that there is any account of morality that *cannot* be formulated in consequentialist terms. After all, nothing keeps us from saying that what a (supposedly) non-consequentialist view does when it tells us which act is *right*, is—essentially—telling us which act is *best*. This thought

gives some initial credibility to idea that even so-called non-consequentialist theories can be expressed in consequentialist terms.

As already noted, with regard to constraints the most relevant promise of the consequentialising project might be that it could reconcile constraints with maximising rationality: if there is a plausible way of arguing that it is *better* not to commit minimising violations of rights, then it is intelligible even from the perspective of the powerful idea of maximising rationality why she should not commit any such violations.[140] While this is a promise the project holds for both consequentialists and non-consequentialists alike, it should hold a couple of additional promises for consequentialists.

First, constraints are often considered as a distinguishing mark of non-consequentialist ethics (Scheffler 1985: 409, Oberdiek 2008: 105). The thesis that there are deontic constraints seems to amount to a denial of consequentialism's key claim that we should always do what is best. Sometimes, it is wrong to act in certain ways, even if doing so would lead to better outcomes. Thus, if even constraintism could be shown to be consequentialisable, this would give considerable momentum to the consequentialisers' project. Arguably, this might be an attractive prospect only to *consequentialisers*, but not to other consequentialists who do not care much for the consequentialising project.

But the project holds a second promise that should be relevant to all consequentialists. As already noted, constraintism is an *appealing* kind of view. Its appeal is, first and foremost, a *moral* one. Not many people—I assume—might be ready to say that it can be right, say, to rape another person to prevent two further rapes, or that it can be right to torture twenty people to prevent the torture of twenty-one others. At least for some action types

---

[140] I take for granted here that the only *possible* consequentialist account of constraintism is one that makes this claim: that there is a sense in which not committing a minimising violation is *better* than doing it. Any constraintist account that denies this claim also denies consequentialism criterion of rightness, namely that rightness is a function of which acts produce better outcomes.

or small differences in the numbers, the idea that the prevention of greater harm is sometimes constrained by a moral duty not to perform certain kinds of acts seems hard to resist. This is one reason why consequentialists should care about the question whether constraints—or at least a minimal subset of constraints—can be given a place within a consequentialist ethical framework.

5.2.3. The Compelling Idea

Before turning to the details of how consequentialising works, it will be useful to distinguish between two dimensions of the consequentialising project.

The consequentialisers' primary aim is to show that all moral theories *can* be consequentialised. Certainly, consequentialisers will have to keep to a certain level of generality here. As Portmore explains, the aim cannot be to prove each and every possible non-consequentialist theory to be consequentialisable (Portmore 2014: 87). Instead, consequentialisers should aim to show how consequentialism can accommodate the various general features of non-consequentialist ethics, such as constraints, agent-centred permissions, supererogation, and moral dilemmas. This is the *technical dimension* of the consequentialising project.

However, as already noted, there is a difference between the claim that we *can* consequentialise moral theories and the claim that we *should* do so. One idea behind the consequentialising project is that consequentialism is such a compelling kind of theory that much could be gained for ethical theory if all other theories could be expressed in consequentialist terms. What would make consequentialising worth doing would be that it generates *compelling* moral theories. This is the *semantic dimension* of the consequentialising project, the level at which consequentialising becomes a *meaningful* practice. Only due to its semantic dimension the project goes beyond the mere technical side of *how* we could consequentialise moral theories and gives

us—including consequentialists—reasons to accept the resulting consequentialised views.

What is so compelling about consequentialism? According to Foot, what gives consequentialism its "spellbinding" force is the "rather simple thought that it can never be right to prefer a worse state of affairs to a better" (Foot 1983: 275). However, there are two necessary adjustments here.

First, in order to include agent-relative consequentialism, we need to take preferability to include what *the agent*, as opposed to others, has most reason to prefer. For instance, suppose that the agent must choose to save either one person, or two other people. All three are strangers to Max, but the first person happens to be Chloe's best friend. Agent-relative consequentialism allows for the ranking of outcomes to change according to the identity of the agent acting. Whereas it might be true that it cannot be right *for Max* to prefer to save one rather than two, it might be right *for Chloe* to prefer to save the one since the one is her best friend, the two others are strangers to her. (If Foot's simple thought requires an *agent-neutral* understanding of what we have reason to prefer, however, then agent-relative consequentialism cannot preserve what is so compelling about consequentialism; more on this in Section 5.4.2.)

The second adjustment is necessary if we want to include *satisficing* and *progressive* consequentialism. In order to attribute Foot's simple thought to these forms of consequentialism, we need to understand the claim that it can never be right to prefer a worse state to a better as limited to those states that are less than sufficiently good. As we have seen, satisficing and progressive consequentialism hold that it is sometimes permissible to produce a sub-

optimal outcome so long as they meet a predefined threshold of sufficient goodness.

Looking ahead to the central argument of this chapter, the relevance of the distinction between a technical and a semantic dimension[141] of the consequentialising project is this. It will turn out that, notwithstanding occasional technical difficulties, there are no restrictions on the technical possibilities to consequentialise moral theories. All theories are consequentialis*able*. However, I will argue that this does not guarantee the *plausibility* of the resulting consequentialist views. In particular, consequentialised constraintism accommodates constraints to the effect that it gives up on a plausible kind of value theory such that different versions of consequentialised constraintism each face particularly severe variants of the value paradox. On this basis, I will reject consequentialised constraintism as a convincing answer to the paradox of deontology. This should give considerable momentum to the hyperinviolability account.

5.2.4. How Consequentialising Works

How do you consequentialise a moral theory? Consequentialisers have said a great deal about the technical aspects of the project. In general, the consequentialising operation is described in terms of a simple recipe:

> Take whatever considerations that the non-consequentialist theory holds to be relevant to determining the deontic status of an action and insist that those considerations are relevant to determining the proper ranking of outcomes. (Portmore 2007: 39)

[141] Baumann (2019) draws a similar distinction in terms of the *technical* and the *interpretive* sides of the consequentialising project.

To be sure, there is a second step to the recipe: claim that once the relevant considerations are built into the outcomes of acts, the ranking of outcomes identifies whatever act the non-consequentialist theory deems to be right as the act that is best (or sufficiently good).

For instance, take a non-consequentialist theory *F* that says that it is permissible to save one's best friend instead of two strangers in a rescue situation. Classical act-utilitarianism, by contrast, tells me that giving priority to my best friend would be wrong and that I should save the greater number of people. *F* thus captures an important moral conviction that classical act-utilitarianism does not. To the consequentialiser, this is what makes *F* a desired target for consequentialising.

Now, all the consequentialiser has to do is to follow the recipe. She should claim that, first, considerations about the agent's relationships to others are morally relevant when determining the possible outcomes of her act. And second, the consequentialiser should claim that an outcome in which I save my friend instead of two strangers ranks more highly than one in which I save two strangers but fail to save my best friend. (This could be done, in particular, by appealing to the importance of agent-relative values; more on this in Section 5.4.) This way, consequentialising may yield a counterpart theory, *F**, which holds that it is right for me to save my best friend instead of two strangers and thus, has the same deontic properties as the original theory *F*. However, in contrast to *F*, *F** is a consequentialist theory because it holds that I should produce the best available outcomes.

What has been done by consequentialising the given moral theory is, first and foremost, that the theory has been given a consequentialist *structure*. *F** is thus a substantive moral view only insofar as *F* is a substantive moral view. As we will see shortly, this fact has led some consequentialisers to claim that *F* and *F** essentially constitute the same moral theory and others to advocate the slightly different view that they were just notational variants of the same theory. However, as we will see, there remain substantive

differences between a consequentialised theory and the original target theory in terms of how they *explain* the deontic verdicts on which they agree.

5.3. The Semantic Dimension of Consequentialising

Consequentialisers maintain that the recipe for consequentialising is universally applicable. It may be used to consequentialise *any* target non-consequentialist view. Once we understand the terms consequentialism and outcomes broadly, it would turn out that all non-consequentialist theories can be consequentialised. That is, it would turn out that for any non-consequentialist ethical theory, there would be a "consequentialist extensional equivalent", i.e., a consequentialist theory that agrees with the original non-consequentialist theory "on the deontic status of every act" (Dreier 2011: 98). I call this thesis *Extensional Equivalence*.

All consequentialisers are committed to *Extensional Equivalence*. Once we grant that it is possible to arrive at a consequentialist theory merely by means of making claims about how outcomes are to be ranked in order to mimic the verdicts of some non-consequentialist theory, we must grant that this strategy can be used to consequentialise *any* non-consequentialist theory, at least so long as there are no restrictions on *how* outcomes may be ranked. Thus, consequentialising would truly provide, as Peterson (2010) puts it, a "royal road to consequentialism".

I have already talked about the promises of the consequentialising project. We can now see that, first and foremost, consequentialising should present itself as a universal means for defending consequentialism against its critics. The consequentialising project starts with the insight that non-consequentialists are sometimes right in insisting on the relevance of some right- or wrong-making features that are not typically considered features of an act's outcome or consequences. A critic may argue that some more traditional version of consequentialism fails because it cannot account for *f*, whereby *f* is what the critic takes to be one relevant factor when determining

the deontic status of an act. The consequentialist now has a way of responding to this criticism: all she needs to do—or at least, this is the idea—is to claim that *f* can be accounted for just as well in consequentialist terms.

However, it should be noted that this makes consequentialising futile as a means for defending a *specific* version of consequentialism. Instead, the consequentialising operation generates a novel version of consequentialism that can evade the criticism against the specific consequentialist view. If some version of consequentialism needs defence qua consequentialising in the first place, all this means is that such a view must be overcome for the sake of a more sophisticated version of consequentialism. Thus, the primary objective of the consequentialising project might be to come up with a theory that preserves consequentialism's compelling idea while providing "a more sophisticated account of how outcomes are to be ranked" (Portmore 2007: 41).

5.3.1. The One-Theory and Two-Variants Views

Consequentialisers have come to wide agreement on the technicalities of how various non-consequentialist theories are to be consequentialised. However, moving away from the technical side of things, consequentialisers and their critics have put forward very different ideas about the *semantic* side of the project. That is, they have suggested very different things about what the consequentialising project *means*.

On the most common view, consequentialism and non-consequentialism are distinguishable not only in terms of their deontic extensions—the sets of things they hold to be right or wrong—but also in terms of *how* they come to hold certain acts right or wrong. That is, extensional disagreement is often understood as a symptom of a more substantive dispute between the two traditions.

Many consequentialisers contradict this view and claim that consequentialising would remove the original disputes between consequentialists and non-consequentialists. Once we grant that there is a consequentialist

extensional equivalent to any non-consequentialist theory, it would turn out that all theories are consequentialist theories. All moral theorists would be, as Jennie Louise puts it, united "under the consequentialist umbrella" (Louise 2004: 536). Accordingly, people who took themselves to be advocating rival moral theories would instead "just make slightly different claims about how to evaluate consequences" (Peterson 2010: 155). I call this the *one-theory view* on the consequences of consequentialising.[142]

The one-theory view is popular also with critics of the project. Paul Hurley, for example, thinks that the consequentialising project, if successful, would have to lead to the conclusion that theorists who take themselves to be non-consequentialists are merely "in the grips of a deep confusion" (Hurley 2013b: 123–124).

A less radical view holds that although consequentialising settles the substantive disputes between both kinds of views, there is still room for the distinction between consequentialism and non-consequentialism. For instance, Jamie Dreier first claimed that consequentialising would eventually prove every moral view to be consequentialist (Dreier 1993: 24); but he later advocates the weaker thesis that a non-consequentialist theory and its consequentialist equivalent are "really just notational variants of one another" (Dreier 2011: 97). As we have seen, consequentialisers are committed to *Extensional Equivalence*, the thesis that for every non-consequentialist theory there is an extensionally equivalent, consequentialist counterpart. In addition, Dreier evokes a thesis that I will refer to as *Extensionality*. *Extensionality* holds that "nothing but [deontic] extension matters in a moral view" (Dreier 2011: 98).

If *Extensionality* is plausible—if extensional equivalence between two views *is* all that matters—then there remains no real dispute between those

---

[142] Tenenbaum (2014) attributes this interpretation to the *dismissive* consequentialisers and distinguishes them from the *earnest* consequentialisers who think that the consequentialising project vindicates consequentialism.

two views. Extensional equivalence would guarantee that the two views agree on everything *that matters*. Yet a non-consequentialist view and its consequentialist equivalent would still be distinguishable by virtue of the different notational frameworks they used. I call this the *two-variants view*.

Note that also the one-theory view rests upon Dreier's *Extensionality* thesis. The difference between this more radical view and the two-variants view is merely that the first concludes that consequentialism is the only game in town, whereas Dreier admits to the more modest position that non-consequentialism and consequentialism are just notational frameworks that can be translated into one another.[143]

Importantly, this translatability relation is a symmetrical one. If non-consequentialism and consequentialism are just notational frameworks, different ways of phrasing the same substantive content, then we can expect that it is possible not only to express any non-consequentialist view using the language of consequentialism, but that it is also possible to express any consequentialist view using the language of non-consequentialism.

Portmore, for instance, argues that we can come up with a Kantian counterpart to any non-Kantian theory that yields, in every possible world, the exact same set of deontic verdicts as the non-Kantian theory. He calls this fictional project *Kantianising*.[144] For instance, if we would like to Kantianise traditional act-utilitarianism, all we had to do is to insist "that we treat humanity as an end-in-itself if and only if we give equal consideration to everyone's interests in maximizing aggregate utility" (Portmore 2007: 59–60).

[143] For a more detailed examination of these diverging interpretations of the consequentialising project see e.g., Schroeder (2017) and Baumann (2019).

[144] I call Kantianising a fictional project because, unlike with the consequentialising project, there do not seem to be any earnest attempts to Kantianise non-Kantian moral theories.

Mirroring the *Extensional Equivalence* thesis there might be a non-consequentialist (or deontological) counterpart to every consequentialist theory.

I do not intend to go further into the idea of Kantianising, nor into the idea of *deontologising*, as the mirror operation to consequentialising is sometimes called.[145] It seems reasonable, however, to think that once we understand consequentialism and non-consequentialism as mere notational variants of the same substantive moral views, it is possible to translate and re-translate the theories of either tradition into one another, just as we can translate between different languages.

5.3.2. The Distinct-Theories View

The idea that consequentialising removes the substantive disputes normally assumed to divide consequentialists and non-consequentialists rests upon the assumption that moral theories do not say anything substantial beyond which acts are right or wrong. It is this assumption—*Extensionality*—which allows the consequentialiser to claim that the substantive disputes between two views are settled once extensional agreement between them is achieved. It is not surprising that *Extensionality* has been questioned, primarily by the critics of the consequentialising project. Its implausibility, however, has been noted by some consequentialisers as well.

It is widely acknowledged that ethical theories aim to do more than just tell us which acts are right or wrong. Also Portmore agrees that ethical theories are in the business of doing more than that; among other things, they also specify *what makes* these acts right or wrong (Portmore 2014: 109). Two theories that are extensionally equivalent in their deontic verdicts can therefore still constitute distinct theories: distinct theories about *what explains* these verdicts. This suggests that consequentialising, even if it grants

[145] The argument to the effect that the possibility of consequentialising entails the possibility of a counter-operation of deontologising is demonstrated in more detail by Hurley (2013b).

extensional equivalence between two views, leaves room for the thought that there is "something substantive at issue between them" (Portmore 2007: 60).

By way of illustration, imagine a moral theory *P* which holds that breaking my promise is wrong because by promising I enter a social contract and it is a law of nature that contracts must be kept. Now take an alternative theory *P** that holds that breaking my promise is wrong because breaking my promises makes the angels cry and I should not make the angels cry. *P* and *P** may yield the exact same verdicts in all relevant types of cases. They may be in perfect agreement that each time I break a promise, I do *something wrong*. They only differ in how the fill in the blank in the statement *Breaking a promise is wrong in virtue of* ____.[146]

Thus, *P* and *P** agree extensionally but disagree on what the right explanation of that extension should be. If extensional equivalence was all that mattered in a moral theory, there would really be no *relevant* difference between these theories. But it seems strange to think that only because *P* and *P** support the same set of deontic verdicts, they would be one and the same moral view or notational variants of one and the same moral view.

Thus, two moral theories may support the same set of deontic verdicts and still be *explanatorily* incompatible. Baumann (2019) proposes that we understand this phenomenon analogous to a well-known phenomenon in scientific theory. Sometimes, empirical evidence is not enough to decide between two scientific theories. At some point, for instance, all the available empirical data equally supported both the Ptolemaic and the Copernican theory of planetary motion, although it was out of question that they offered

---

[146] It is very common to think that acts have their normative or evaluative properties (their being right, wrong, etc.) *because* or *in virtue of* their non-normative properties (their causing pain, constituting rights violations, etc.). From this perspective, a *normative explanation* aims to explain in virtue of *what* acts are right or wrong. I will presuppose this view on normative explanations without discussing it in any further depth. For further discussion see e.g., Hooker (2002), Leibowitz (2011), and Väyrynen (2013), (2021).

incompatible explanations of why the planets move the way they do. Sometimes, scientific theories are *underdetermined* by empirical evidence.

In an analogue way, Bauman says, moral theories are underdetermined by deontic verdicts. Two moral views may yield the exact same set of deontic verdicts and still offer incompatible explanations as to why we should act in those ways. Seen in this way, it does not appear miraculous that we have always regarded consequentialism and non-consequentialism as mutually exclusive theories, even if it would turn out that extensional equivalence between them could easily be achieved (Baumann 2019: 521).

Thus, consequentialism and non-consequentialism can be mutually exclusive, even if they may host extensionally equivalent views, due to the incompatible explanations these views would give for the relevant deontic verdicts. Extensionally equivalent moral views can be explanatorily incompatible. This means, however, that we should reject Dreier's *Extensionality* thesis. Deontic extension is *not* all that matters in a moral view. It also matters how that view explains the deontic verdicts contained in its extension.

The rejection of Dreier's *Extensionality* thesis has two major consequences. First, consequentialising must be seen to have a much more moderate impact on ethical theory than the one-theory and two-variants views suggest. In a way, the finding that all theories can be consequentialised would invite everyone to stand under the consequentialist umbrella. But if extensional equivalence isn't all that matters, the original substantive disputes about the *right kind of explanation* in ethics would simply continue under consequentialism's umbrella. Consequentialism and non-consequentialism are distinct theories, not just notational variants of the same moral view. I call this the *distinct-theories view* on the consequences of consequentialising.

Second, once we understand that moral theories with the same deontic properties can be substantively incompatible in terms of their explanatory properties, we obtain an additional criterion for deciding between these theories. Whereas it seems hard to decide between two extensionally

equivalent theories if extension is all that matters, it seems not so hard to decide between them if extension is *not* all that matters. This is so because their explanatory properties can make theories more or less *plausible*. For instance, we might have reasons to accept the moral theory *P* rather than the theory *P**, not based on their extensions (they are extensionally equivalent) but based on the explanations they each give to support the relevant deontic verdicts.

Theory-plausibility should be an important concern to consequentialisers. If consequentialising generates moral views that lack plausibility due to the explanatory properties which ground their extensional properties, then consequentialising is futile both as a means of defending consequentialism and as an invitation for non-consequentialists to join under the consequentialist umbrella. For one thing, consequentialists will not want to advocate some consequentialised theory if it is overall less plausible than their original consequentialist view. For another, non-consequentialists will have no good reason to adopt a consequentialised theory if it is overall less plausible than their original non-consequentialist view.

Others have argued that the success of the consequentialising project depends on the question whether consequentialising creates *plausible* moral views (e.g., Woodard 2013, Betzler and Schroth 2019, Muñoz 2021). However, what makes a moral theory plausible may be a matter of debate. The point I want to make is a more specific one: I want to argue that consequentialising generates versions of constraintism that are unavoidably paradoxical. Consequentialised constraintism might avoid the rationality paradox. But it cannot avoid the value paradox and thus, fails to justify constraints.[147]

[147] Thus, I take for granted that for a constraintist view to be unable to justify constraints is something that makes that theory lack plausibility which, at the very least, should give us reasons to accept an alternative constraintist account that *is* able to justify constraints—so long as, of course, such an account is available.

5.4. Agent-Relative Consequentialising

Consequentialisers widely agree that when it comes to constraintism, the consequentialising project will have to make use of considerations that feature in *agent-relative* consequentialist views. This is so because consequentialisers generally subscribe to the standard view according to which constraintism is an agent-relative kind of theory.

I have proposed an alternative, agent-neutral interpretation of constraintism in Chapter 2. And in Section 5.5, we will see that the project of trying to agent-neutrally consequentialise constraintism might not be a lost cause entirely. However, under the assumption that constraintism only makes sense in agent-relative terms, consequentialisers have focused their efforts on showing how constraintism could be represented as a version of agent-relative consequentialism. I therefore start with the agent-relative consequentialising project.

On an agent-relative account of consequentialism, the ordering of acts (or outcomes) may change according to the identity of the agent acting. For instance, whereas Max should prefer the death of one to the death of two strangers, Chloe may have reason to prefer the death of the two to the death of one, if the one is her best friend. What matters is, at least in part, what is good *relative to the agent acting*.

As we have seen, in order to accommodate deontic constraints, agent-relative consequentialists will have to arrive at the claim that it is *better* if the agent does not kill anyone, even if by doing so she could prevent more further killings. In the context of an agent-relative account of consequentialism, the idea must therefore be that for any act type subject to a deontic constraint, acts of that type performed by the particular agent in the present moment are worse-*relative-to-that-agent* than acts of the same type performed by other agents or by the same agent at a different time.

To illustrate, Portmore asks us to imagine a model constraintist theory *D* that holds "that it is wrong for an agent to dirty her hands and violate

someone's rights even for the sake of minimizing rights violations overall" (Portmore 2007: 54–55).[148] Next, Portmore claims that we should come up with a consequentialist theory *D** that holds that an outcome where the agent dirties her hands by violating someone's rights ranks lower on the proposed ordering that any outcome where she does not do so, even if by violating someone's right she could prevent more violations overall. *D* and *D** have the same deontic properties regarding minimising violations of rights—they forbid them. Thus, *D** is a consequentialised version of constraintism. Note that to account for intrapersonal constraints on minimising one's own rights violations, the ranking of outcomes should also be moment-relative (Portmore 2009: 330–331).

5.4.1. Preserving the Deontic Properties

The proposal of agent-relative consequentialising seems easy enough to understand. However, it faces a certain technical difficulty, to begin with. It is not obvious how agent-relative consequentialising could preserve the deontic properties of constraintism. In particular, it is unclear how it could make it *impermissible* (instead of *not required*) to commit minimising violations.

To see why, suppose that it is preferable from my standpoint that *I* do not commit one killing than that someone else commits two, and that this preference is a morally significant one. Even so, it should be possible for me to understand that there is *something bad* about the killings that I could prevent. The mere reason that those acts would be committed by someone else cannot make up, at least not entirely, for the badness of two killings. Agent-relative consequentialised constraintism must assign *some* negative weight

[148] Note that Portmore speaks of a very specific kind of non-consequentialist theory here. In Chapter 2, I have argued that it is not obvious that non-consequentialists would accept such a view. For one thing, non-consequentialists do not typically think that constraints are grounded in a refusal to dirty one's hands. For another, I have argued that reference to dirty hands cannot in fact ground an obligation to obey constraints.

to the killings I could prevent. It will certainly not encourage me to be indifferent towards them.

In other words, the agent-neutral disvalue of killings is still assessable on an agent-relative account of consequentialism. The only thing that makes it right *for me* to prefer two killings to one is the great agent-relative disvalue assigned to killings that *I* would commit.[149] If we took away the agent-relative disvalue of my act, what would remain is the agent-neutral disvalue of two killings versus one, and it should be clear what consequentialists must say for this case.

What, then, could keep me from choosing to sacrifice something of agent-relative value—being free from committing any acts of killing—for the sake of the greater agent-neutral good? Strictly speaking, I would not do anything wrong if I chose to minimise killings overall. I don't *have to* do it, but why shouldn't I be allowed to? How can agent-relative consequentialism ground an *obligation*—not just an agent-centred permission—not to commit minimising violations? It would no longer be *impermissible* to minimise killings but *permissible* not to kill (or even supererogatory to kill).

Arguably, the problem just sketched is one not concerns not exclusively a consequentialist account of agent-relative constraintism but also the original non-consequentialist view. Thus, I am not saying that agent-relative constraintism cannot agent-relatively consequentialised. It makes perfect sense that to consequentialise a moral theory also means to adopt at least some of the problems of the original view. All I am making is the rather trivial

[149] The great disvalue of my violations might be understood either as being weightier than the agent-neutral disvalue of the violations I could prevent, or it might be understood as taking lexical priority over the latter. In either case, since the relevant disvalue of my violations is agent-relative, it is unclear why I should not have a certain sovereignty over the decision to preserve *that* value or to protect something of greater *agent-neutral* value instead.

point that in order to accommodate *constraints*, it is not enough to accommodate agent-centred restrictions.

5.4.2. Preserving the Compelling Idea

Even if agent-relative consequentialism could accommodate constraints, there is a second worry concerning the semantic side of the consequentialising project: consequentialising was supposed to be justified on the grounds that it preserves consequentialism's compelling idea. However, it is not obvious that agent-relative consequentialism can do this.

For instance, Mark Schroeder has argued that agent-relative consequentialism cannot perverse what is so compelling about traditional consequentialism. On Schroeder's view, the appeal of traditional consequentialism lies with the idea that it is "always permissible for every agent to do what will lead to the outcome that is best" (Schroeder 2007: 279). This is sufficiently close to Foot's simple thought. But Schroeder also seems to assume that what is best must be defined in agent-neutral terms. Consequentialism's *genuine* appeal lies with the thought that we should do what is best *from an agent-neutral standpoint*. And since deontic constraints make it impermissible sometimes to produce the agent-neutrally best outcomes, any view allowing for constraints "would be inconsistent with the Compelling Idea, and hence inconsistent with consequentialism" (Schroeder 2007: 279). In short, consequentialism cannot accommodate constraints because constraints are simply

counterexamples of the only compelling form of consequentialism, i.e., *agent-neutral* consequentialism.[150]

Admittedly, as we have seen, it is not necessary to agree with Schroeder that only an agent-neutral account of consequentialism could preserve its compelling force. Consequentialisers may try to argue that the thought that we should do what is best-relative-to-the-agent is sufficiently compelling so that agent-relative consequentialism can accommodate the compelling idea.

However, at the very least, consequentialisers will have to say more in order to make this claim plausible. Even if agent-relative consequentialism might be seen as somewhat compelling, its new compelling idea seems "far less compelling than [the compelling idea] as proposed by standard consequentialism" (Betzler and Schroth 2019: 130). Agent-relative consequentialism may come with the advantage that it allows for agents to be good in a bad world. But this comes at the cost that the world would be an overall worse place, both in terms of how many agents would perform wrong acts and in terms of the number of people killed, tortured, or otherwise seriously harmed.

Thus, agent-relative consequentialism, once it aims to incorporate constraints, faces a variant of the value paradox that also troubled the agent-centred approach. (This is unsurprising as agent-relative consequentialism's

---

[150] Recently, Muñoz (2021) has put forward a similar argument against consequentialising, which attacks all forms of consequentialism, not just the agent-relative one. He argues that consequentialising undercuts the consequentialist position because it denies the theory of action central to consequentialism. Consequentialism holds that action is production; that to act *is* to produce an outcome. Muñoz argues that consequentialism is only compelling because of this view on action. However, even if consequentialisers only argue that killing is worse than letting die, this already implies that action is *not* just production but that, sometimes, acts should fit the world as it is instead of changing it. Thus, consequentialising destroys the component of consequentialism which made it compelling and justified the consequentialising project. While I agree with the general direction of this argument, I do not want to commit myself to the claim that consequentialisers cannot even account for the difference between killing and letting die without undercutting their own position. I want to stay focused on the question whether there is a plausible consequentialised version of constraintism.

solution to the paradox takes the agent-centred approach.) How can it be appropriate to think that the agent should act upon the agent-relative disvalue of rights violations instead of acting upon the greater agent-neutral disvalue of rights violations committed by others (or herself at other times)? How can *her own* agency be so significant, when compared to the significance of what would happen to the greater number of victims? But more than that, agent-relative consequentialism faces a *particularly severe* variant of the value paradox because, unlike non-consequentialists, consequentialists are committed to the claim that what the agent should do is produce better outcomes. How does an outcome in which the agent does not commit any killings herself deserve to be called *better* than an outcome in which she does, if the first outcome contains many more killings? I return to this point in the discussion of absolute and moderate consequentialised constraintism (Section 5.6).

5.5. Agent-Neutral Consequentialising

As already mentioned, consequentialisers have had little hope that agent-neutral consequentialism could accommodate deontic constraints. The problem is that accepting that we should agent-neutrally consequentialise constraintism seems to amount to a denial of the consequentialist claim that the right act is one that produces the best outcome overall. As Paul Hurley puts it, the consequential*ising* claim about constraints would seem to defeat the consequential*ist* claim about moral rightness (Hurley 2013b: 128). Maintaining that it is appropriate to consequentialise the "impartial but agent-relative values" of deontic constraints, would mean to be "committed to the rejection of an [agent-neutral] consequentialist account of moral value" (Hurley 2013b: 132).

Hurley seems to think that this problem arises only for agent-relative views. The way he sees it, by recognising the need to consequentialise constraintism, what the consequentialiser would recognise is the existence of agent-relative restrictions on the production of the (agent-neutrally) best

outcomes. Evidently, however, such restrictions cannot be accommodated within the framework of a consequentialist theory that holds that we should always produce the best outcomes, agent-neutrally understood.

But what happens if constraintism, as I have argued in Chapter 2, is *not* committed to agent-relativity? If we can make sense of deontic constraints in solely agent-neutral terms, does this help the consequentialising project? On the face of it, it does not seem so. I will later discuss an argument proposed by Setiya (2018) that might help to resurrect the agent-neutral consequentialising project. For now, it will be useful to say why agent-neutral consequentialism might struggle to accommodate not only *agent-relative*, but also *agent-neutral* constraints.

5.5.1. The Road to Utilitarianism of Rights

Suppose that the consequentialist agrees that any plausible moral theory should account for rights. People have rights, for instance, against severe forms of interference with their mental and physical well-being constituted by acts of killing, torture, enslavement, abuse, and so on. Constraintism, understood as a theory of rights, holds that these rights act as constraints on the conduct of others.

It might be possible to consequentialise rights. The consequentialiser could argue that something that adds to the badness of killing, for instance, is that such an act constitutes a rights violation. Acts that constitute rights violations rank lower than otherwise identical acts that do not constitute rights violations. These claims are open to the consequentialiser because she may choose to recognise additional sorts of good and bad to count towards the consequentialist calculation of the relative goodness of act-outcomes.

However, if the consequentialiser is committed to agent-neutrality, it seems that she is also committed to the claim that the action that takes rights seriously is one that maximises the extent to which they are respected. She will be committed, it seems, to saying that we *should* kill *A* in the *Special Case*

where this prevents the killing of *B* and *C*. This way, the consequentialiser would have consequentialised rights, but not rights-*as-constraints*. In other words, she would have failed to consequentialise rights-as-constraints and instead become what Nozick had called a *utilitarian of rights*.

The problem here is not that rights would place *agent-relative* constraints on action. Rather, the problem is that in order to accommodate rights-as-constraints, the consequentialiser would need to get to the claim that an outcome in which you do not kill *A* but two others get killed is *better* than an outcome where you kill *A* but the two live. It is hard to see how she could get to this claim so long as her view commits her to the idea that we should do what is best, *agent-neutrally* speaking. From an agent-neutral consequentialist perspective, the trade-off is one rights violation against two. And it is hard to see how the consequentialist calculus could yield the verdict that *one* rights violation is *worse* than *two*.

In short, the reason why agent-neutral consequentialism does not seem to be able to accommodate constraintism as a theory of rights is not that constraintism is agent-relative, but that consequentialism, it seems, cannot make sense of deontic constraints in agent-neutral terms. On a maximising, teleological structure, it seems that any explanation as to why you must not kill in the *Special Case* must feature the thought that killing *A* is a very bad thing to do *for you*. As we will see shortly, this is an important acknowledgement.

5.5.2. Can It Be Worse to Kill?

There might be a way to resurrect the agent-neutral consequentialising project. Kieran Setiya (2018) has argued against the common idea that agent-neutral consequentialists must prefer that we kill to prevent more further killings. His argument runs as follows.

First, Setiya asks us to compare two cases (Setiya 2018: 96)—his versions of the *Normal Case* and the *Special Case*:

> **Five Killings.** A villain is using a trolley to attempt to kill five strangers.
>
> **One Killing to Prevent Five.** A villain is using a trolley to attempt to kill five strangers. You can stop the trolley by pushing a button that drops a sixth stranger off a bridge into its path.

Setiya argues that things are going equally bad in both scenarios up to a certain point. In each case, five innocents are going to get killed. If we should prefer *One Killing to Prevent Five* to *Five Killings*, then we should think that, at the point where you push the button, things *improve* morally.

However, Setiya claims that this is not the case. Quite the contrary. Things get *worse* from there. By pushing the button, you add to the situation yet another insult on an innocent person's life, making *One Killing to Prevent Five* worse in comparison to *Five Killings* (Setiya 2018: 104). Moreover, Setiya claims that this verdict does not depend on the perspective of the agent who would commit the sixth killing. Suppose that someone *other than you* would have to push the button. Would that change anything? Setiya thinks it wouldn't. Just like you, so would *anyone* add yet another insult on another person's life by pushing the button and would thereby make the situation impersonally worse. You should want everyone, not just yourself, to prefer *Five Killings* to *One Killing to Prevent Five* (Setiya 2018: 98).

Setiya's point is that focusing on the temporal unfolding of events in *One Killing to Prevent Five*, agent-neutral consequentialists could claim that killing a sixth stranger where five other strangers are already going to get killed if not saved makes the situation *impersonally* worse.

As a side note, Setiya does not explicitly claim to have provided an argument for consequentialising.[151] This, of course, is just a dialectical point.

---

[151] It is not obvious what Setiya thinks his argument amounts to. He distinguishes *special* from *general* constraints and argues that while special constraints conflict with agent-neutrality, general constraints (such as a constraint on killing) do not. Yet he thinks of general constraints as a subset of *agent-centred restrictions* (Setiya 2018: 95), thus as

If consequentialising means to reproduce the deontic verdicts of an originally non-consequentialist theory on a consequentialist structure, and if constraintism is understood as an agent-neutral theory, then the upshot of Setiya's argument is that we can consequentialise constraintism on an agent-neutral consequentialist structure.

Setiya's argument rests upon a rather momentous assumption: that *actual* killings are just as bad as *prevented* killings. This is not an interpretation of Setiya's view. He explicitly endorses this idea:

> The situation in which someone is going to be killed unless they are saved… is as bad as the situation in which they are going to be killed. Ethically speaking, the damage has been done. (Setiya 2018: 104)

I will say more about whether this is a convincing piece of thought. For now, I just want to take note of some of its wider implications that should make it a somewhat unattractive claim for consequentialists. Let me refer to the idea that actual and prevented killings have an equal negative disvalue as the *Equality* thesis. *Equality* has several disconcerting implications in other types of cases.

First, there are situations where an act of rescue should—but according to *Equality* does not—make a moral difference. Suppose that you are walking past a shallow pond and witness how Gregg throws Mae, a young child, into the water. In the next moment, Gregg has fled the scene. You know you could easily save Mae, at basically no risk to your own life. You might as well do it. But does it make a difference whether you save Mae, ethically speaking? According to *Equality*, the situation in which Gregg kills Mae is as bad as the situation in which Gregg *would* kill Mae if you did not save her.

"restrictions, which give special weight to whether *you* kill anyone *now*" (Setiya 2018: 92). One result of these definitory commitments is that Setiya thinks of his view as accommodating the impermissibility of minimising violations without accommodating even general constraints (because such constraints would be agent-relative and his view is an agent-neutral one).

Does this not mean that your act of saving Mae would fail to add any positive value to the consequentialist calculation?

Admittedly, the *Equality* thesis is not logically incompatible with the idea that the situation in which you fail to save Mae is worse than the one in which you save Mae. After all, *Equality* only says that the badness of Gregg's act should be associated with the attempt to kill Mae, not with her successful killing. It says that the situation where Mae is *threatened* to be killed bears already the full negative value that the situation would bear were Mae *actually* killed. This is not an absurd claim to make. Non-consequentialists often make parallel claims about moral wrongdoing (e.g., Brook 1991). However, if we take this claim seriously, *Equality* implies that since the full badness of the situation already manifests before Mae dies, saving her does not actually prevent anything bad from happening. If consequentialists want to argue that it would be *right* to save Mae (and I assume they would), then they are short an explanation as to why it is *better* if you save Mae than if you just do nothing.

Second, *Equality* has (potentially) disconcerting implications for a popular type of trolley case where pulling a switch will save five but cause the (unintended) death of one. Compare two cases—the second case is identical to the *Trolley Switch* case but has been given a new name here in the spirit of Setiya's argument:

> Five Deaths. A runaway trolley is about to kill five strangers.
>
> One Death Instead of Five. A runaway trolley is about to kill five strangers. You can save them by pulling a switch that will redirect the trolley to a side-track where it will kill a sixth stranger instead.

Note that there are two main differences between *One Killing to Prevent Five* and *One Death Instead of Five*. Firstly, you would prevent intentional killings in the first case and accidental deaths in the second case. And secondly, you

do so by *directly* killing the sixth person in the first and *indirectly* killing them in the second case.

Again, up to a certain point, things seem to be going equally bad in both scenarios because five innocents will be killed by a runaway trolley. Many people think—and consequentialists should feel particularly drawn towards this belief—that we should prefer *One Death Instead of Five* to *Five Deaths*. We should thus be able to claim that at the point where you pull the switch the situation improves compared to *Five Deaths*. However, if we take Setiya's argument about *One Killing to Prevent Five* at face value, it turns out that we are committed to the opposite claim. In fact, Setiya's view commits us to the idea that from the point where you pull the switch the situation gets *worse*. It is quite bad if five people are going to be killed by a runaway trolley. But it must be *worse*, on Setiya's view, if in order to prevent those five deaths, you kill a sixth person. By doing this, you add something bad (an indirect killing) to an already bad situation.

Thus, if we should prefer *Five Killings* to *One Killing to Prevent Five*, then it seems that we should also prefer *Five Deaths* to *One Death Instead of Five*. If you would make *One Killing to Prevent Five* worse than *Five Killings* by killing the sixth person, then it seems your killing them would also make *One Death Instead of Five* worse than *Five Deaths*.

Could Setiya resist this conclusion? Is there a way of arguing that we should prefer *Five Killings* to *One Killing to Prevent Five* and yet, at the same time, prefer *One Death Instead of Five* to *Five Deaths*?

One possible response might be this. Killing the sixth person by pushing a button that drops them in front of the trolley is *pretty* bad. In fact, it is much worse than indirectly killing someone by redirecting a trolley away from five other innocents. In *One Death Instead of Five*, the sixth person just happens to be on the side-track. It is just worse if you directly kill someone (like

in *One Killing to Prevent Five*) than if you indirectly cause their death (like in *One Death Instead of Five*).

However, this cannot explain why five *accidental deaths* should be worse than one (indirect) *killing*. Setiya must explain how it can be better to kill in *One Death Instead of Five* but worse to kill in *One Killing to Prevent Five*. The worseness of direct killings as opposed to indirect killings provides no such explanation. It is reasonable to think that if *Equality* implies that actual and prevented killings are equally significant, then it also implies that actual and prevented deaths are. Thus, if you prevent five deaths by indirectly killing one, what you would do is adding the negative value of one (indirect) killing to an outcome that already entails the negative value of five prevented deaths.

Admittedly, these might be obstacles on the technical side of the agent-neutral consequentialising project. And I don't intend to argue that there is no way in which consequentialisers could handle the unsettling implication of Setiya's argument. As far as I can see, consequentialisers would have two options here. Either they bite the bullet and accept the unsettling implications of *Equality* for at least the two types of cases discussed above; or they show that *Equality* does not actually have the implications I say it has. Perhaps, there is a way of arguing that the negative value of actual and prevented outcomes is only equal with intentional, direct killings. Accidental deaths or killings that occur as unintended side-effects, in contrast, are still worse if they actually occur than if they are prevented.

Either way, the question remains whether agent-neutral consequentialising leads to a plausible account of constraints. Setiya tells us that it is good—*agent-neutrally* good—if you do not kill to the extent that it is *better* if you do not kill than if you prevent many more killings. His agent-neutral constraintist view can avoid the value paradox as little as its agent-relative cousin. For how does a world in which many more killings occur deserve to be called *better* than a world in which exceedingly fewer killings occur? In

particular, it is just not obvious that there is a plausible sense of goodness according to which the world is a better place if it contains five actual killings, if compared to a world that contains an additional killing-*attempt* but in which five out of six killings have successfully been prevented. Just as agent-relative consequentialist constraintism, Setiya's agent-neutral account cannot show that constraints are not paradoxical.

It is not even obvious that agent-neutral consequentialism, once it aims to accommodate constraints, can avoid the rationality paradox, as it lacks the explanatory resources provided by agent-relativity. How can it be wrong to minimise killings if what we should do is produce better outcomes and if the fact that the agent herself would commit some of those killings must not feature in the explanation of why her killings are particularly bad? On a moral view that takes the good to be prior to the right and that, moreover, aims to define the good in agent-neutral terms, it must remain mysterious how it could ever be right not to kill if this would prevent more further killings.

5.6. Moderate and Absolute Constraintism

As we have seen, each constraintist view is either an *absolute* or a *moderate* view. Moderate constraintists believe that there are circumstances—for instance, if the numbers of rights violations one could prevent are particularly high—when the agent may permissibly commit a minimising violation. Absolute constraintists deny this. Consequentialisers must show that *both* absolute and moderate constraintism can be consequentialised. I begin with absolute constraintism.

5.6.1. Absolutism Consequentialised

Absolutists believe that it is always wrong to φ even to prevent more further φ-ings. To consequentialise absolute constraintism, consequentialisers would

have to claim that it is *better* not to *φ* than to prevent more further *φ*-ings, *for any number of φ-ings one could prevent with a single act of φ-ing*.

There are, roughly, two ways to get to this claim. The consequentialiser could either claim that infinite disvalue must be assigned to any single act of *φ*-ing that would prevent more *φ*-ings. This way, the disvalue of the agent's *φ*-ing would always outweigh the disvalue of *φ*-ings she could prevent, notwithstanding their number. Or the consequentialiser could evoke the idea of some trumping mechanism that renders the disvalue of the agent's *φ*-ing incomparable to the disvalue of *φ*-ings she could prevent. This way, the badness of her *φ*-ing would trump the combined badness of any number of other *φ*-ings to the effect that it would always be worse if she *φ*-s to prevent that many further *φ*-ings. In what follows I focus on the first strategy. But everything I have to say applies also to the second strategy.[152]

On the face of it, absolute constraintism does not fit well into a consequentialist framework. A non-consequentialist can argue that constraints *precede* any considerations about weighing goods. She can claim that however many killings I could prevent, I should not kill. The non-consequentialist's argument may go "directly to what I ought or ought not to do, without first examining the goodness of the alternatives" (Broome 1991: 9). The same line of thought, however, is not open to the consequentialist constraintist. As a consequentialist, she is still committed to the idea that the deontic status of any act is a function of the goodness of its outcome in relation to the goodness of alternative outcomes. As a constraintist, she is committed to the idea that constraints are a function of the relative goodness of the non-performance of certain acts in certain situations. Of course, the consequentialist

---

[152] One typical example of the second strategy is to introduce lexical priority for values or reasons. For instance, some moral theories hold that some practical reasons have lexical priority over others such that they trump other kinds of practical reasons. The distinction I'm making here is a subtle and not well-established one. It bears no further relevance for my argument. It is also worth questioning whether there is a clear distinction here in the first place, considering what choosing one or the other strategy would amount to. Sometimes, lexical priority is thus simply defined as giving some reasons infinitely greater weight than others (e.g., Huemer 2010).

constraintist could just claim that it is *always better* if I don't kill, no matter how many killings I could prevent. But this is a highly implausible claim. Here is why.

Inherent to the idea of a ranking of outcomes is the idea that the ranking may *change* according to the changing of relevant factors. Suppose that a murderer has left two strangers, *A* and *B*, to die on separate islands. You have a boat but can only save either *A* or *B*, not both. Thus, you could produce any of the following three outcomes:

O_1 Save *A.*

O_2 Save *B.*

O_3 Save no one.

Presumably, it is better to save someone than not to save anyone, so O_1 and O_2 should outrank O_3. There is no reason, however, to prefer either O_1 or O_2. Presumably, it is permissible to produce either O_1 or O_2 because both entail that you save one stranger and fail to save another stranger.

Now add a third stranger, *C*, who has been left to die on the same island as *B*. That means, you could now produce a fourth outcome:

O_4 Save *B* and *C.*

Each of the other three outcomes has changed because of what they now entail. For instance, O_1* now entails that you fail to save not just *B* but *B and C*. O_2* entails that you fail to save not only *A*, but *A* and *C*. And O_3* now entails that you fail to save not only two, but three strangers.

What we should expect is that by adding *C* the ranking of outcomes has changed. The worst outcome is still that you fail to save anyone. In fact, O_3* is even worse than O_3 because it entails the unprevented killing of an

extra person. O_1* and O_2* should now share second place. And the best possible outcome should be O_4, i.e., that you save *B* and *C*.

What this shows is that the number of killings you could prevent is a relevant factor when determining the correct ranking of outcomes. Still, absolute constraintists claim that you should not kill to prevent *any number* of further killings. As a version of consequentialism, absolute constraintism would thus hold that *although* the number of killings you could prevent is a relevant factor when determining the ranking of outcomes, in paradox cases it does not even have the *potential* to change the ranking. The ranking of outcomes would be static in the sense that changing a relevant factor (the number of killings) has no impact whatsoever on the ranking. But it seems that on any *plausible* consequentialist view that holds that rightness is a function of a ranking of outcomes, the changing of factors relevant to that ranking must at least have the potential to change the ranking of outcomes, and thus the ordering of acts that determines whether you should (still) obey a constraint on killing.

On an at least partially non-consequentialist account, the claim that we should not kill, notwithstanding the number of killings we could prevent, has some plausibility because non-consequentialists—to stay in the metaphor—may chose not to put killings on either side of the scale. They may refuse to even touch the scale. A consequentialist, by contrast, cannot make this choice. She must hold that the weighing of killings is fundamentally prior to the wrongness of killing.

5.6.2. Moderatism Consequentialised

Now to moderate constraintism. Moderatists are sensitive to the degree of harm you could prevent by committing a minimising violation such that there are circumstances when it is right for you to do so. They think that while it is sometimes (and maybe in most cases) wrong to commit a rights violations where this would minimise violations of the same right in everyone, there's

some degree of harm you could prevent that justifies committing a minimising violation.

Whereas it seems hard to consequentialise absolute views at all, moderatism lies more comfortably with a consequentialist framework. Suppose, for instance, that we have consequentialised agent-relative constraintism, as laid out earlier. Our consequentialised theory holds, among other things, that it is wrong to kill a single person even to prevent someone else from killing two others. As we have seen, the agent-neutral disvalue of killings that the agent could prevent remain on the record. The two killings are just outweighed or trumped by the disvalue attached to the single killing that, from the agent's perspective, would be worse-relative-to-her. (For the sake of completeness, let us also take record of the agent-neutral disvalue of the agent's potential act of killing.)

We can now start adding to the numbers on the prevention side of the scale: the agent could prevent three killings with a single killing, then four, then five, and so on. It is foreseeable that most consequentialists' hesitation to agree with the statement *It is wrong to kill even to prevent n further killings* will start to grow parallel with the growing number *n*. Maybe, it is wrong to prevent ten killings with a single killing. But is it still wrong to prevent twenty, fifty, or a hundred killings with a single killing?

It should come naturally to consequentialists to say that there is a point where the (agent-neutral) disvalue of a huge number of killings eventually outweighs the (combined agent-relative and agent-neutral) disvalue of a single killing, such that it becomes right to commit that killing. Once consequentialists and non-consequentialists agree that there are deontic constraints—once *this* debate has been settled—consequentialism seems to offer a plausible explanation for why these constraints should not be taken to be absolute.

However, moderate constraintism will still be a version either of agent-relative or of agent-neutral constraintism (of which I have only

exemplified agent-relative moderate constraintism). Thus, a consequentialist account of moderate constraintism must commit to the implausible value theories of either agent-relative or agent-neutral consequentialist constraintism.

Perhaps, consequentialised moderatism seems less puzzling than its absolute cousin, if we consider that it allows for at least some cases where the great disvalue of rights violations one could prevent may change the ranking on outcomes such that their prevention is permissible or even required. But this does not help to avoid the value paradox. For one thing, for the value paradox to arise the number may be as small as one against two. It is enough to assume that we should not commit a single killing even to prevent two further killings. For the value paradox simply says that where the harm the agent could do and the harm she could prevent are of the same type, it seems inappropriate to focus on anything other than the minimisation of that type of harm overall.

For another thing, consequentialised moderatism can handle higher degrees of harm only at the expense that it is vulnerable to two other problems. First, moderatism faces "the notorious problem of locating the threshold" (Betzler and Schroth 2019: 125). As a consequentialised theory, moderatism needs to tell us how much preventable harm is enough to justify a minimising violation. For any degree of harm that the moderatist can refer to here, one could rightfully ask why we should think that this is the point at which the ranking changes. To be sure, this might not be a grave problem. We would expect any threshold to some moral principle to be vague. The vagueness of the threshold would not prevent the constraintist from making clear statements about *some* numbers, for instance, that we may not kill one to prevent two other killings, but we may kill to prevent a hundred more.

The other problem, which I take to be the more worrisome one, is that it is not clear on a consequentialist account how the idea that minimising violations can be justified can be built into the concept of constraints in the

first place. A constraint on *φ*-ing is a constraint on letting the prevention of further *φ*-ings *matter in a certain way*. The prevention of *φ*-ings must not matter in the sense that it can change the ranking of outcomes. Once we claim that the prevention of *some* greater number of *φ*-ings matters in this way, how can we claim that the prevention of more *φ*-ings does not always matter in the way that makes minimising violations permissible?

5.7. Conclusion

To consequentialise various forms of constraintism, consequentialisers will have to make different claims about which considerations should feature in the description of the outcomes of acts. Then, they will have to make claims about the ways in which outcomes should be ranked in order to reach the same verdicts as the target constraintist views.

I have not argued that consequentialisers are not in a position to make these claims. Notwithstanding certain technical difficulties, we may grant that all versions of constraintism *can somehow* be consequentialised. Broadening the focus to the whole of the consequentialising project, I have not argued that consequentialisers are not in a position to make the claims that they need to make in order to show that *all* theories can be consequentialised. Instead, once we take the idea of consequentialising at face value, it turns out to be trivial that all theories can be consequentialised because, according to consequentialisers, there are no general restrictions on what may feature in the description of an act's outcome, nor on the way in which outcomes are to be ranked.

Consequentialisers sometimes say that all "remotely plausible" (Portmore 2007: 39) or "non-crazy" theories (Suikkanen 2009: 2) can be given a consequentialist reinterpretation. Of course, they might say this because there is no good reason *for wanting to consequentialise* crazy or not even

remotely plausible moral theories. But it is insightful that, in principle, even these theories can be consequentialised.

Imagine a moral theory *L* which holds that lying is wrong only if the agent sees a black cat in the moment just before telling a lie. To be fair, connecting the moral wrongness of lying to the previous perception of black cats does not make much sense. So, probably, *L* would not count as a remotely plausible or non-crazy theory. But why would this mean that *L* cannot (or cannot easily) be consequentialised? Admittedly, there might not be a good reason for wanting to consequentialise such an implausible theory about lying. But why would *L*'s lack of plausibility place any constraints on the technical possibility of consequentialising? Applying the universal recipe, the consequentialiser could simply claim that an outcome in which the agent lies is only bad if it is also an outcome in which she previously sees a black cat. This is enough to make it the case that there is a consequentialised version of *L* that has the same deontic properties as *L*.

To be sure, I am not saying that consequentialisers would reject the idea that *L* can be consequentialised. It is understandable why they want to focus on non-crazy theories. The point I am making is that once we grant that we can consequentialise *some* moral theory in the way consequentialisers propose, we have to grant that we can use the same recipe to consequentialise *any imaginable* theory. This is so because whatever some theory—crazy or not—holds to be morally right, nothing keeps us from arguing that it is *best* to act in that way. For any act that some non-consequentialist theory holds to be right, we can come up with a ranking of outcomes such that the act in question produces the best outcome.

One might think that this would have to mean that consequentialism is an empty term (e.g., Brown 2011). But I am also not merely repeating *this* point. What it shows, I think, is that any interesting dispute about whether moral theories *can* be consequentialised will happen on the *semantic* level of the consequentialising project. The question is not whether it is technically

possible to consequentialise non-consequentialist theories, but whether the resulting consequentialised theories are *plausible* or *compelling* variants of the original views. And here, extension cannot be all that matters. The plausibility of moral theories also depends on their explanatory properties.

I have also not argued that consequentialising *never* generates plausible moral theories. My argumentative aim was a more restrictive one. What I hope to have shown is merely that it does not generate plausible accounts of *constraintism*, as such accounts cannot avoid the value paradox of constraints. As such, they do not offer a plausible alternative to the hyperinviolability account. The constraints of consequentialising—this is the general lesson—are constraints not on what can *technically* be consequentialised, but on what can *plausibly* be consequentialised.

General Conclusion

Deontological constraintism captures a central feature of common moral thought: that an act of φ-ing which is wrong in virtue of its essential properties cannot be rendered permissible by the fact that φ-ing would prevent more further φ-ings. According to constraintism, it cannot be right, for instance, to kill or torture someone only to prevent more further killings or tortures. The rational justification of deontological constraintism, however, has been haunted by two persistent ideas.

The first idea is that the default rational way to engage with a moral value is to promote it—to maximise or optimise its presence through action. The second idea which has been haunting deontological constraintism is that a moral theory which tells us to do anything other than promoting the presence of what we should value must be based on a conception of morality that cannot give shared moral aims to all agents. Together these ideas have nurtured the persistent objection that deontological constraintism is unavoidably paradoxical.

In this thesis, I have presented arguments against both these ideas—arguments of which I hope that they could convince not only those sympathetic to constraintist views but also opponents of constraintism of the thought that deontological constraintism is a coherent kind of moral view, a view free of paradox. Against the first idea I have argued that a moral theory can sometimes rationally prohibit the maximisation of that which it identifies

as valuable, namely when the relevant value is one which cannot be promoted through action. Against the second idea I have argued that we can make sense of constraints on value-maximisation as robustly agent-neutral principles that give all agents shared moral reasons to oppose certain kinds of actions.

Yet what might be most notable about a piece of philosophical work are its limitations, as making its limitations explicit also sets out what we may learn from it. Neither have I aimed to show that constraintism is true, nor was it my primary aim to make the reader accept a constraintist position. Instead, all I have aimed to show is how we could clear the air of paradox surrounding deontological constraintism. In this regard, I have emphasised a certain aspect of moral agency: that our actions not only have the power to *change* the world (for the better or for the worse), but also the power to *represent* certain alternatives about how the world *could be*.

There is a very familiar sense in which *to act* just means to make it the case that the world is in a certain state: *to ought to act* might then simply mean to ought to make it the case that the world is in a certain state. However, our actions also have other powers. Since it will always be undecided—to a certain extent—what kind of moral world we live in, our actions have the power to represent moral ideals simply by fitting these ideal worlds rather than aiming at changing the actual world. Our actions are not only *productive* of the good, but also *representative* of certain conceptions of the good.

This is by no means a new lesson for ethical theory; but it is one worth repeating. A rational and plausible moral theory can (and perhaps should) give a place to the representative aspect of agency. Such a theory does not simply indulge in a kind of wishful thinking. Instead, it emphasises a way in which ethical theory has a more direct impact upon our world—one that goes beyond the simple thought that we should make the world a better place.

References

Alexander, L. and Moore, M. (2020). Deontological Ethics. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Available at: <https://plato.stanford.edu/archives/sum2021/entries/ethics-deontological/>.

Alvarez, M. (2018). Reasons for Action, Acting for Reasons, and Rationality. *Synthese*, 195(8), 3293–3310.

Alvarez, M. (2010). *Kinds of Reasons: An Essay in the Philosophy of Action*. Oxford University Press.

Alm, D. (2009). Deontological Restrictions and the Good/Bad Asymmetry. *Journal of Moral Philosophy*, 6(4), 464–481.

Bader, R. (2016). Conditions, Modifiers, and Holism. In E. Lord and B. Maguire (Eds.), *Weighing Reasons*. Oxford University Press, 27–55.

Baumann, M. (2019). Consequentializing and Underdetermination. *Australasian Journal of Philosophy*, 97(3), 511–527.

Bennett, J. (1998). *The Act Itself*. Oxford University Press.

Betzler, M. and Schroth, J. (2019). The Good of Consequentialized Deontology. In C. Seidel (Ed.), *Consequentialism: New Directions, New Problems*. Oxford University Press, 115–135.

Brand-Ballard, J. (2004). Contractualism and Deontic Restrictions. *Ethics*, 114(2), 269–300.

Brink, D. O. (2006). Some Forms and Limits of Consequentialism. In D. Copp (Ed.), *The Oxford Handbook of Ethical Theory*. Oxford University Press, 380–423.

Brook, R. (1991). Agency and Morality. *Journal of Philosophy*, 88(4), 190–212.

Broome, J. (1991). *Weighing Goods. Equality, Uncertainty and Time*. Blackwell.

Brown, C. (2011). Consequentialize This. *Ethics*, 121(4), 749–771.

Burri, S. (2017). Personal Sovereignty and Our Moral Rights to Non-Interference. *Journal of Applied Philosophy*, 34(5), 621–634.

Chappell, T. (2011). Intuition, System, and the 'Paradox' of Deontology. In L. Jost and J. Wuerth (Eds.), *Perfecting Virtue. New Essays on Kantian Ethics and Virtue Ethics*. Cambridge University Press, 271–288.

Chappell, T. (2007). Integrity and Demandingness. *Ethical Theory and Moral Practice*, 10(3), 255–265.

Cummiskey, D. (1990). Kantian Consequentialism. *Ethics*, 100(3), 586–615.

Cushman, F. (2015). From Moral Concern to Moral Constraint. *Current Opinion in Behavioral Sciences*, 3(1), 58–62.

Dancy, J. (2004). *Ethics Without Principles*. Oxford University Press.

Dancy, J. (1993). *Moral Reasons*. Blackwell.

Darwall, S. L. (1986). Agent-Centered Restrictions From the Inside Out. *Philosophical Studies*, 50(3), 291–319.

Dean, R. (2013). Humanity as an Idea, as an Ideal, and as an End in Itself. *Kantian Review*, 18(2), 171–195.

Dougherty, T. (2013). Agent-Neutral Deontology. *Philosophical Studies*, 163(2), 527–537.

Draper, K. (2005). Rights and the Doctrine of Doing and Allowing. *Philosophy and Public Affairs*, 33(3), 253–280.

Dreier, J. (2011). In Defense of Consequentializing. In M. Timmons (Ed.), *Oxford Studies in Normative Ethics*, Vol. 1. Oxford University Press, 97–119.

Dreier, J. (1993). Structures of Normative Theories. *The Monist*, 76(1), 22–40.

Emet, S. F. (2010). Agent-Relative Restrictions and Agent-Relative Value. *Journal of Ethics and Social Philosophy*, 4(3), 1–13.

Enoch, D. (2009). Wouldn't It Be Nice If p, Therefore, p (for a Moral p). *Utilitas*, 21(2), 222–224.

Foot, P. (1983). Utilitarianism and the Virtues. *Proceedings and Address of the American Philosophical Association*, 57(2), 273–283.

Foot, P. (1967). The Problem of Abortion and the Doctrine of the Double Effect. *Oxford Review*, 5(1), 5–15.

Fried, C. (1978). *Right and Wrong*. Harvard University Press.

Gert, J. (2016). The Distinction between Justifying and Requiring: Nothing to Fear. In E. Lord and B. Maguire (Eds.), *Weighing Reasons*. Oxford University Press, 157–172.

Hammerton, M. (2020). Deontic Constraints are Maximizing Rules. *Journal of Value Inquiry*, 54(1), 571–588.

Hammerton, M. (2019). Distinguishing Agent-Relativity from Agent-Neutrality. *Australasian Journal of Philosophy*, 97(2), 239–250.

Hammerton, M. (2017). Is Agent-Neutral Deontology Possible? *Journal of Ethics and Social Philosophy*, 12(3), 319–324.

Hammerton, M. (2016). Patient-Relativity in Morality. *Ethics*, 127(1), 6–26.

Hare, C. (2013). *The Limits of Kindness*. Oxford University Press.

Hart, H. L. A. (1955). Are There Any Natural Rights? *The Philosophical Review*, 64(2), 175–191.

Hauser, M. (2006). *Moral Minds*. Harper Collins.

Heath, J. (2008). *Following the Rules: Practical Reasoning and Deontic Constraint*. Oxford University Press.

Heuer, U. (2011). The Paradox of Deontology, Revisited. In M. Timmons (Ed.), *Oxford Studies in Normative Ethics*, Vol. 1. Oxford University Press, 236–267.

Hooker, B. (2002). Intuitions and Moral Theorizing. In P. Stratton-Lake (Ed.), *Ethical Intuitionism: Re-Evaluations*. Oxford University Press, 76–161.

Huemer, M. (2010). Lexical Priority and the Problem of Risk. *Pacific Philosophical Quarterly*, 91(3), 332–351.

Hurka, T. (2003). Moore in the Middle. *Ethics*, 113(3), 599–628.

Hurley, P. (2013a). Paradox of Deontology. In LaFollette (Ed.), *The International Encyclopedia of Ethics*. Blackwell, 3790–3794.

Hurley, P. (2013b). Consequentializing and Deontologizing. In M. Timmons (Ed.), *Oxford Studies in Normative Ethics*, Vol. 3. Oxford University Press, 123–153.

Jaworska, A. and Tannenbaum, J. (2021). The Grounds of Moral Status. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Available at: <https://plato.stanford.edu/archives/ spr2021/entries/grounds-moral-status/>.

Jeske, D. (1998). Families, Friends, and Special Obligations. *Canadian Journal of Philosophy*, 28(4), 527–555.

Johnson, C. M. (2019). The Intrapersonal Paradox of Deontology. *Journal of Moral Philosophy*, 16(3), 279–301.

Kagan, S. (1991). Replies to My Critics. *Philosophy and Phenomenological Research*, 51(4), 919–928.

Kagan, S. (1989). *The Limits of Morality*. Clarendon Press.

Kamm, F.M. (2007). *Intricate Ethics: Rights, Responsibilities, and Permissible Harm*. Oxford University Press.

Kamm, F.M. (2001). *Morality, Mortality. Volume II: Rights, Duties, and Status*. Oxford University Press.

Kamm, F.M. (1992). Non-Consequentialism, the Person as an End-in-Itself, and the Significance of Status. *Philosophy and Public Affairs*, 21(4), 354–389.

Kamm, F.M. (1989). Harming Some to Save Others. *Philosophical Studies*, 57(3), 227–260.

Kant, I. (1998). *Groundwork of the Metaphysics of Morals*. Edited by M. Gregor. Cambridge University Press.

Kearns, S. (2016). Bearing the Weight of Reasons. In E. Lord and B. Maguire (Eds.), *Weighing Reasons*. Oxford University Press, 173–191.

Kleingeld, P. (2020). How to Use Someone 'Merely as a Means'. *Kantian Review*, 25(3), 389–414.

Korsgaard, C. M. (1993). The Reasons We Can Share: An Attack on the Distinction Between Agent-Relative and Agent-Neutral Values. *Social Philosophy and Policy*, 10(1), 24–51.

Lang, G. and Lawlor, R. (2013). In Defence of Batman: Reply to Bradley. *Journal of Ethics and Social Philosophy*, 3(1), 1–6.

Lang, G. (2005). Fairness in Life and Death Cases. *Erkenntnis*, 62(3), 321–351.

Lazar, S. (2019). Moral Status and Agent-Centred Options. *Utilitas*, 31(1), 83–105.

Leibowitz, U. D. (2011). Scientific Explanation and Moral Explanation. *Noûs*, 45(3), 472–503.

Lippert-Rasmussen, K. (2009). Kamm on Inviolability and Agent-Relative Restrictions. *Res Publica*, 15(2), 165–178.

Lippert-Rasmussen, K. (1999). In What Way are Constraints Paradoxical? *Utilitas*, 11(1), 49–70.

Lippert-Rasmussen, K. (1996). Moral Status and the Impermissibility of Minimizing Violations. *Philosophy and Public Affairs*, 25(4), 333–351.

Lord, E. and Maguire, B. (2016). Introduction. In E. Lord and B. Maguire (Eds.), *Weighing Reasons*. Oxford University Press, 3–26.

Louise, J. (2004). Relativity of Value and the Consequentialist Umbrella, *Philosophical Quarterly*, 54(1), 518–536.

Mack, E. (1998). Deontic Restrictions Are Not Agent-Relative Restrictions. *Social Philosophy and Policy*, 15(2), 61–83.

McNaughton, D., and Rawling, P. (2006). Deontology. In D. Copp (Ed.), *The Oxford Handbook of Ethical Theory*. Oxford University Press, 424–458.

McNaughton, D., and Rawling, P. (1998). On Defending Deontology. *Ratio*, 11(1), 37–54.

McNaughton, D., and Rawling, P. (1995a). Agent-Relativity and Terminological Inexactitudes. *Utilitas*, 7(2), 319–325.

McNaughton, D., and Rawling, P. (1995b). Value and Agent-Relative Reasons. *Utilitas*, 7(1), 31–47.

McNaughton, D., and Rawling, P. (1993). Deontology and Agency. *The Monist*, 76(1), 81–100.

McNaughton, D., and Rawling, P. (1992). Honoring and Promoting Values. *Ethics*, 102(4), 835–843.

McNaughton, D., and Rawling, P. (1991). Agent-Relativity and the Doing-Happening Distinction. *Philosophical Studies*, 63(2), 167–185.

Mill, J. S. (2002). *Utilitarianism*. Edited by G. Sher. Hackett.

Moore, M. S. (2019). The Rationality of Threshold Deontology. In H. Hund (Ed.), *Moral Puzzles and Legal Perspectives*. Cambridge University Press, 371–387.

Moore, M. S. (2008). Patrolling the Borders of Consequentialist Justification: The Scope of Agent-Relative Restrictions. *Law and Philosophy*, 27(3), 35–96.

Moore, M. S. (1997). *Placing Blame: A General Theory of the Criminal Law*. Oxford University Press.

Muñoz, D. (2021). The Rejection of Consequentializing. *Journal of Philosophy*, 118(2), 79–96.

Nagel, T. (2008). The Value of Inviolability. In P. Bloomfield (Ed.), *Morality and Self-Interest*. Oxford University Press, 102–113.

Nagel, T. (2002). Personal Rights and Public Space. In T. Nagel, *Concealment and Exposure. And Other Essays*. Oxford University Press, 31–52.

Nagel, T. (1986). *The View from Nowhere*. Oxford University Press.

Nagel, T. (1972). War and Massacre. *Philosophy and Public Affairs*, 1(2), 123–144.

Nagel, T. (1970). *The Possibility of Altruism*. Princeton University Press.

Nick, C. (2019). Can Our Hands Stay Clean? *Ethical Theory and Moral Practice*, 22(1), 925–940.

Nozick, R. (1974). *Anarchy, State, and Utopia*. Blackwell.

Oberdiek, J. (2008). Culpability and the Definition of Deontological Constraints. *Law and Philosophy*, 27(2), 105–122.

Oberdiek, J. (2004). Lost in Moral Space: On the Infringing/Violating Distinction and Its Place in the Theory of Rights. *Law and Philosophy*, 23(4), 325–346.

Otsuka, M. (2011). Are Deontological Constraints Irrational? In R. Bader and J. Meadowcroft (Eds.), *Cambridge Companion to Nozick's Anarchy, State, and Utopia*. Cambridge University Press, 38–58.

Otsuka, M. (2000). Scanlon and the Claims of the Many versus the One. *Analysis*, 60(3), 288–293.

Otsuka, M. (1997). Kamm on the Morality of Killing. *Ethics*, 108(1), 197–207.

Parfit, D. (1984). *Reasons and Persons*. Oxford University Press.

Parfit, D. (1979). Is Common-Sense Morality Self-Defeating? *Journal of Philosophy*, 76(10), 533–545.

Peterson, M. (2010). A Royal Road to Consequentialism? *Ethical Theory and Moral Practice*, 13(2), 153–169.

Pettit, P. (2000). Non-Consequentialism and Universalizability. *Philosophical Quarterly*, 50(199), 175–190.

Pettit, P. (1989). Consequentialism and Respect for Persons. *Ethics*, 100(1), 116–126.

Pettit, P. (1987). Universalizability Without Utilitarianism. *Mind*, 96(381), 74–82.

Pleitz, M. (2018). *Logic, Language, and the Liar Paradox*. Mentis.

Portmore, D.W. (2019). *Opting for the Best: Oughts and Options*. Oxford University Press.

Portmore, D.W. (2014). *Commonsense Consequentialism: Wherein Morality Meets Rationality*. Second Edition. Oxford University Press.

Portmore, D.W. (2013a). Agent-Centered Restrictions. In H. LaFollette (Ed.), *The International Encyclopedia of Ethics*. Blackwell, 158–162.

Portmore, D.W. (2013b). Agent-Relative vs Agent-Neutral. In H. LaFollette (Ed.), *The International Encyclopedia of Ethics*. Blackwell, 162–171.

Portmore, D.W. (2009). Consequentializing. *Philosophy Compass*, 4(2), 329–347.

Portmore, D.W. (2007). Consequentializing Moral Theories. *Pacific Philosophical Quarterly*, 88(1), 39–73.

Preston-Roedder, R. (2014). A Better World. *Philosophical Studies*, 168(3), 629–644.

Quinn, W. (1993). Actions, Intentions and Consequences: The Doctrine of Doing and Allowing. In W. Quinn, *Morality and Action*. Cambridge University Press, 149–174.

Quong, J. (2016). Agent-Relative Prerogatives to Do Harm. *Criminal Law and Philosophy*, 10(4), 815-829.

Rawls, J. (1999). *A Theory of Justice*. Revised Edition. Harvard University Press.

Raz, J. (2003). *The Practice of Value*. Edited by R. J. Wallace. Oxford University Press.

Regan, T. (2004). *The Case for Animal Rights*. University of California Press.

Ridge, M. (2017). Reasons for Action: Agent-Neutral vs. Agent-Relative. E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Available at: <https://plato.stanford.edu/archives/ fall2017/entries/reasons-agent/>.

Rønnow-Rasmussen, T. (2009). Normative Reasons and the Agent-Neutral/Relative Dichotomy. *Philosophia*, 37(2), 227–243.

Ross, A. P. (2016). Inviolability and Interpersonal Morality. *Journal of Value Inquiry*, 50(1), 69–82.

Scanlon, T. M. (1998). *What We Owe to Each Other*. Second Edition. Harvard University Press.

Scheffler, S. (1994). *The Rejection of Consequentialism*. Clarendon Press.

Scheffler, S. (1985). Agent-Centred Restrictions, Rationality, and the Virtues. *Mind*, 94(375), 409–419.

Schroeder, M. (2011). Ought, Agents, and Actions. *Philosophical Review*, 120(1), 1–41.

Schroeder, M. (2007). Teleology, Agent-Relative Value, and 'Good'. *Ethics*, 117(2), 265–295.

Schroeder, S. A. (2017). Consequentializing and Its Consequences. *Philosophical Studies*, 174(6), 1475–1497.

Sen, A. (1982). Rights and Agency. *Philosophy and Public Affairs*, 11(1), 3–39.

Setiya, K. (2018). Must Consequentialists Kill? *Journal of Philosophy*, 18(1), 92–105.

Singer, M. G. (1965). Negative and Positive Duties. *The Philosophical Quarterly*, 15(59), 97–103.

Slote, M. (2020). *Common-Sense Morality and Consequentialism*. Revised Edition. Routledge.

Snedegar, J. (2017). *Contrastive Reasons*. Oxford University Press.

Steinhoff, U. (2016). Self-Defense as Claim, Right, Liberty, and Act-Specific Agent-Relative Prerogative. *Law and Philosophy*, 35(2), 193–209.

Suikkanen, J. (2009). Consequentialism, Constraints and the Good-Relative-To: A Reply to Mark Schroeder. *Journal of Ethics and Social Philosophy*, 3(1), 1–8.

Swinburne, R. (2008). God and Morality. *Think*, 20(1), 7–15.

Taurek, J. M. (1977). Should the Numbers Count? *Philosophy and Public Affairs*, 6(4), 293–316.

Tenenbaum, S. (2014). The Perils of Earnest Consequentializing. *Philosophy and Phenomenological Research*, 88(1), 233–240.

Timmermann, J. (2004). The Individualist Lottery: How People Count, But Not Their Numbers. *Analysis*, 64(2), 106–112.

Thomson, J. J. (2001). *Advice and Goodness*. Edited by A. Gutmann. Princeton University Press.

Thomson, J. J. (1990). *The Realm of Rights*. Harvard University Press.

Thomson, J. J. (1985). The Trolley Problem. *The Yale Law Journal*, 94(6), 1395–1415.

Väyrynen, P. (2021). Normative Explanation and Justification. *Noûs*, 55(1), 3-22.

Väyrynen, P. (2013). Grounding and Normative Explanation. *Aristotelian Society Supplementary Volume*, 87(1), 155-178.

Wenar, L. (2021). Rights. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Available at: <https://plato.stanford.edu/archives/spr2021/entries/rights/>.

Williams, B. (1973). A Critique of Utilitarianism. In J. J. C. Smart and B. Williams (Eds.), *Utilitarianism: For and Against*. Cambridge University Press, 77–151.

Woodard, C. 2013. The Common Structure of Kantianism and Act-Utilitarianism, *Utilitas*, 25(2), 246–265.

Woollard, F. (2008). Intricate Ethics and Inviolability: Frances Kamm's Nonconsequentialism. *Ratio*, 21(2), 231–238.

Zong, D. (2000). Agent-Neutrality is the Exclusive Feature of Consequentialism. *Southern Journal of Philosophy*, 38(4), 676–693.