# University of Southampton

Faculty of Arts and Humanities

Department of Philosophy

**Doing and Allowing Harm to Future Generations**

by

**Charlotte Franziska Unruh**

ORCID ID 0000-0003-3953-7617

Thesis for the degree of Doctor of Philosophy

November 2020

# University of Southampton

## <u>Abstract</u>

Faculty of Arts and Humanities

Department of Philosophy

Thesis for the degree of <u>Doctor of Philosophy</u>

Doing and Allowing Harm to Future Generations

by

Charlotte Franziska Unruh

In this thesis, I aim to clarify what moral accounts of harm can tell us about the strength of the reason against real-world long-term harm doing. To this end, I develop a novel account of harm, and defend the Doctrine of Doing and Allowing (DDA) as a principle that partly explains the strength of harm-based moral reasons. The DDA says that the reason against doing harm is stronger than the reason against merely allowing harm. I discuss three challenges for the DDA which, if successful, undermine the ability of the DDA to guide action in real-world cases. First, the DDA has difficulties accounting for cases of letting oneself do harm. Second, it is unclear how the DDA applies to cases under risk, and specifically, cases in which agents offset risks of harm. Third, the DDA has been argued to have deeply implausible implications in cases of long-term, indirect and unpredictable harm doing. I defend the DDA against these challenges and, in doing so, clarify how the DDA applies to real-world long-term decision making.

However, such clarification is only useful when it can be combined with a plausible and comprehensive account of harm. I develop a novel account of harm, which combines a *hybrid view* on the nature of harm with a *two-dimensional view* on harm-based moral reasons. The hybrid view on harm says that both ill-being and a loss of well-being constitute a harmed condition. The two-dimensional view on harming says that behaviour can qualify as harming either because it makes a difference to, or because it causally contributes to, a harmed condition. I argue that my account shares advantages with views discussed in the literature, while avoiding their major problems. I specifically defend the temporal comparative account of harm (as one component of the hybrid view) against seemingly fatal counterexamples and show that it is in fact more plausible than the much more prominent counterfactual version of the account. The upshot is that accounts of harm can inform decision-making in long-term cases. This supports the view that the notion of harm deserves its central role in moral theory, and the view that the DDA is well suited to provide real-world action guidance.

# Table of Contents

# Table of Tables and Figures

# Research Thesis: Declaration of Authorship

Print name: Charlotte Franziska Unruh

Title of thesis: Doing and Allowing Harm to Future Generations

I declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as: Material from chapter 4 has been accepted for publication at the *Journal of the American Philosophical Association*.

Signature:

Date:
18/11/20

# Acknowledgements

Over the last three years, I have been very lucky to be able to share and discuss my ideas with many smart and inspiring people. Without these conversations, this thesis would not have been written, and I owe a great debt of gratitude to all those people. I hope that I have included the most obvious candidates here, but I will undoubtedly have missed some. I ask those who I have missed to forgive me.

I am very grateful to the participants of the Higher Seminar in Practical Philosophy at the University of Uppsala for the invaluable discussion and detailed comments on a draft version of the first chapter. For helpful questions and comments, I would also like to thank audiences at the following conferences: Rocky Mountain Ethics Congress (Boulder 2019), Postgraduate Bioethics Conference (Swansea 2019), Understanding Value VIII (Sheffield 2019), REAPP Early Career Conference (Reading 2019), Value and Uncertainty Conference (Lisbon 2018), ISUS 2018 (Karlsruhe 2018), and the audience at the workshop 'Do deontological distinctions survive future generations?' (Southampton 2018) and Bennett Francis for his helpful response. I would also like to thank all participants of the Southampton Graduate Seminar 2017-2020, my respondents in the Graduate Seminar, and the examiners and audience of the milestone progression reviews at the University of Southampton.

For extremely helpful written comments on draft chapters, I would like to thank Friedemann Bieber, Garrett Cullitty, Anna Folland, Timotheus Riedel, Ziggy Schilpzand, and anonymous reviewers. Thanks to Atus Mariqueo-Russell and Carla Wiggs for their generous help with proofreading.

For their support and friendship, I thank my fellow PhD students in Southampton, in particular Ziggy Schilpzand, Ben Paget-Woods and Carla Wiggs, who have somehow managed to make three years of thinking about harm a lot of fun. For his love and support, I am forever grateful to Henry Tellwright.

I am very grateful to the AHRC SWW DTP for a full studentship and generous funding for conference travel, and to the Centre of Moral and Political Philosophy at the Hebrew University of Jerusalem for awarding me a PhD Summer Workshop Fellowship in May and June 2018, which has influenced my thinking on risk and uncertainty ever since.

I owe the greatest debt of gratitude, however, to my fantastic supervisors, Fiona Woollard and Brad Hooker. Their insight and thoughtfulness can hardly be overstated. I am very lucky to have had the opportunity to work with and learn from such great philosophers.

# Chapter 1 Introduction

Harm is at the centre of moral theorizing. However, the concept of harm is controversial. The main accounts of harm in the literature face significant problems. The jury is still out on whether (a revised version of) the proposed accounts of harm can overcome these problems, or whether, in Ben Bradley's words, we should '[l]et harm go the way of phlogiston' (2012, 411) and omit the concept from moral theorizing. Moreover, there is widespread disagreement about the moral status of different kinds of harmful behaviour. Many consequentialists think that there is no morally relevant difference between doing harm and merely allowing harm. In contrast, most non-consequentialists, in line with common sense moral intuitions, hold the Doctrine of Doing and Allowing Harm (DDA)[1], the view that the moral reason against doing harm is stronger than the moral reason against merely allowing harm. However, the recent literature has posed several challenges for defenders of the DDA. These include challenges arising from cases under risk, cases of "letting oneself do harm", and cases of long-term indirect harms. If successful, these challenges would severely limit the applicability of the DDA to real-world cases, in particular behaviour that leads to harmful consequences in the far future.

Today's actions have the potential to affect the lives of those who will live in the far future. It has been argued that this knowledge requires an extension, if not a revision, of moral theory (Jonas 1979, 28), as moral concepts and principles need to be adapted in order to accommodate long-term consequences. This includes the concept and morality of harming. Decisions regarding, for example, climate change mitigation and adaptation, transforming food systems, developing technology and securing the stability of financial and political systems might have harmful long-term consequences. But how do these future harms compare to harms suffered by those who presently exist? Are they the same kind of harm? And what can moral principles about doing harm tell us about the strength of harm-based reasons against actions with potential harmful long-term consequences? These questions have so far not been thoroughly examined in the literature on the concept and morality of harm (with the notable exception of difficulties that arise from the non-identity problem).

In this thesis, I defend the DDA as a principle that can explain the strength of moral reasons against harming. My aim is to clarify how the theoretical debates on the nature and strength of harm-based moral reasons can help us think about real-world cases, specifically cases of long-

[1] While contemporary discussion of the moral relevance of the distinction between doing and allowing harm started earlier (e.g. Foot (1967)), to my knowledge Quinn (1989a) was the first person to refer to the claim that doing harm is harder to justify than allowing harm as the "Doctrine of Doing and Allowing" (DDA).

term harm doing. If such clarification is successful, this can give us reason to think that the notion of harm deserves its central role in moral theory. However, such a clarification would not be useful if it turned out that the notion of harm itself is in too much trouble to be used in real-world moral decision making. I therefore develop a novel account of harm, which can overcome challenges that have been raised in the literature.

The chronological order in which I present these arguments will be as follows. In the first part of my thesis, I develop a novel account of harm. My account combines a *hybrid view* on the nature of harm with a *two-dimensional view* on harm-based moral reasons. I argue that these views can combine the advantages of some of the views in the literature, while avoiding their difficulties. Very roughly, the hybrid view says that there are two ways in which someone's welfare can be affected in a way that constitutes a harmed condition. The first way in which someone can be in a harmed condition is by suffering ill-being. The second way in which someone can be in a harmed condition is by losing well-being. The two-dimensional view says that behaviour can qualify as harming along two dimensions. The difference-making dimension asks whether the behaviour makes a difference to some harm. The causal contribution dimension asks whether the behaviour causally contributes to harm. Behaviour that counts as harming on either or both dimensions is subject to harm-based reasons against this behaviour.

After having addressed worries regarding the viability of the concept of harm in the first part of this thesis, in the second part, I turn to a discussion of the DDA. Discussing DDA is worthwhile, not just because it is a moral principle that is widely accepted in common-sense morality and virtually all non-consequentialist moral theories. It also links to the account of harm that I defend in the first part of the thesis. Recent defences of the DDA, most notably Fiona Woollard's (2015; 2012b; 2012c; see also Foot 1967), analyse the distinction between doing and merely allowing harm in terms of the agent's involvement in the causal sequence leading to harm. The DDA, thus understood, is a moral principle that can explain the difference in the strength of harm-based moral reasons against behaviours that differ along what I call the causal contribution dimension. (Doing harm and allowing harm can make the same difference to a harmful outcome and might therefore not differ along the difference-making dimension.)

I discuss three challenges that have been raised against DDA in the literature. All three challenges confront the defender of the DDA with problematic cases that the defender of the DDA seemingly cannot adequately account for. I suggest that all three challenges can be combined under the headline that if they are successful, they undermine the ability of the DDA to guide action in real-world cases. The first challenge is raised by cases of letting oneself do harm (Persson 2013; J. Hanna 2015a; 2015b). The second challenge regards cases under risk (Kagan 1989, chap. 3), and specifically, cases in which agents offset risks of harm that they themselves have imposed. The third challenge is about cases of long-term, indirect and unpredictable harms (MacAskill and Mogensen 2019). A failure to respond to these challenges would constitute a serious, if not fatal, blow to the DDA. All three challenges arise from cases that involve features which will be present in most real-world cases, and almost all cases of harming future generations. I defend the DDA against these challenges, and in doing so, clarify how the DDA applies to real-world long-term decision making.

My discussion of the three challenges serves two main purposes. The first purpose is to supplement the account of harm that I develop in the first part of the thesis, as a defence of the DDA is also a defence of the moral relevance of the causal contribution dimension of my two-dimensional account. It explains the difference in the strength of moral reasons against behaviours that are relevant to the causal sequence leading to harm in different ways.[2]

The second purpose of discussing the three challenges to the DDA is to clarify the implications of the DDA in cases of real-world and in particular, long-term harm doing. I argue that despite recent objections, the DDA does apply to such cases, and moral theories that contain the DDA are therefore well suited to provide real-world action guidance. This is an encouraging result for proponents of DDA and strengthens their case against critics of the DDA. The second part of the thesis should therefore be of interest to proponents and critics of the DDA alike.

What emerges is that the DDA is a strong competitor for moral principles that can guide real-world action and can be accommodated within an account of harm as a principle that expresses the comparative strength of moral reasons against harming.

[2] However, my account of harm does not depend on whether the DDA is ultimately defensible. First, the view that causally contributing to harm matters morally is plausible, independent of whether DDA is true. For example, it is widely accepted that moral responsibility for harm depends on causal responsibility for harm (Sartorio 2007). It seems possible at least in principle that the causal contribution dimension can be spelled out in ways that do not involve the DDA (for example, one might replace the view that there is a morally relevant difference between doing and allowing harm with the view that there is a morally relevant difference between actions that are proximate causes of harm and actions that are not, see section 6.3). Second, if causal contribution to harm is shown to have no moral relevance, then while this would mean that the strength of the reason against harming is purely determined by the difference-making dimension, classifying harming behaviours along the causal contribution dimension might still matter for other purposes, e.g. in order to pragmatically determine compensatory duties.

## 1.1 The Problem with Future Harm

I aim to show that we should keep the notion of harm as a central concept in moral theorizing. My account of harm, I believe, can adequately deal with cases of future harm. These are arguably some of the most difficult cases of harm to be dealt with in moral theory. Showing how these difficult cases can be dealt with brings abstract moral concepts and principles about doing harm a step closer to real-world issues facing humanity today, such as climate change.

Moreover, discussion of future harms as paradigm problem cases will also be useful to moral theorists working in the emerging field of global priorities research. Concern for the long-term future has gained traction in moral philosophy, with a debate emerging from the effective altruism community over the so-called “long-termism paradigm”. This debate addresses whether moral decisions should be made with regard to how they affect the long-term future, and whether moral theories should be judged by their ability to guide such decision making (see e.g. Beckstead 2019). A deontological perspective on long-termism is likely to focus on moral duties not to harm. How moral duties not to harm apply to cases of future harm can make a crucial difference for priority setting regarding trade-offs between short-term and long-term outcomes of actions.

Such a deontological perspective on long-termism should not only matter to committed deontologists. It should also matter to defenders of the Maximize Expected Moral Value Approach to moral uncertainty, who hold a non-zero credence in deontology. The Maximize Expected Moral Value Approach is the most prominent view on moral uncertainty, which holds what one ought to do depends on the moral value of possible outcomes of one’s actions according to all the theories that an agent has some credence in (and on the degree of credence) (Bykvist 2017, 4). This approach seems to imply that anyone with a non-zero credence in deontology should care about the implications of deontology for long-term obligations.

In sum, the availability of a concept of harm that is applicable to decisions involving future harm justifies the status quo of harm as a central moral concept, and moreover, provides new frameworks to think about moral obligations towards the long-term future, frameworks that are arguably needed given the global challenges with potentially harmful long-term effects that face decision makers today.

Before we proceed, I will explain in more detail why I see future harm as a paradigm problem case for accounts of harm, and specifically for accounts of harm that distinguish doing and allowing harm in terms of the causal sequences leading to harm. Roughly, many cases of doing harm to future people involve features that change the moral status of harmful actions. Not all cases of future harm involve all of these features, and some cases of harming present people might involve some of these features. Nonetheless, cases of future harm serve to illustrate the significance and difficulties of finding an account of harm that can guide action in real-world cases. While I will not restrict my discussion to cases of future harm, the features highlighted in cases of future harm will appear throughout the thesis, and many of my examples will draw upon these features.

The first feature is non-identity. Our relations to contemporaries are in significant ways different from our relations to future generations. The moral status of future people is difficult to determine. To start with, it is unclear how we can justify moral obligations towards non-existent people (for an overview of theories that nonetheless try to justify these obligations, see e.g. Campos (2018)).[3] Furthermore, some large-scale social policies, such as those that are driving climate change, plausibly change the identity of future people, such that these policies do not make anyone worse off (Parfit 1984). Some philosophers have argued that future people are not harmed in these 'non-identity' cases (e.g. Heyd 2009; see also Boonin 2014), many more have suggested an array of possible solutions (for an overview, see Roberts (2019)).

The second feature is that the timescale of future harm means that agents often have an opportunity to undo their own actions that will otherwise lead to future harm. In other words, they can intervene in causal chains leading to harm more than once. The moral status of such cases is unclear. Persson (2013) and J. Hanna (2015a) have argued that this is a problem for defenders of the DDA.

The third feature is uncertainty. Uncertainty is discussed in the literature on obligations towards the future, in particular, it has led theorists to suggest discounting of future utility when making long-term policy decisions (Davidson 2006; Purves 2016). There is an emerging literature on the possibility and shape of deontological risk ethics (the probably most comprehensive account for moderate deontology under risk thus far has been developed by Seth Lazar (2017a; 2017b; 2018), but see also Tenenbaum (2017) and Portmore (2017)). Kagan (1989) has argued that the defender of the DDA faces difficulties in cases under risk; however, there is to my knowledge no explicit discussion on risky cases in the context of the distinction between doing and allowing harm.

[3] I think that we can distinguish three main ways that have been explored to justify obligations to future generations. First, with regard to the interests or status of future individuals (Caney 2010; Steigleder 2016). Second, regarding the interests or status of present individuals (Howarth 1992; Vanderheiden 2006; Bos 2016; Düwell and Bos 2016; Gheaus 2016). Third, regarding the interests or status of future groups, or communities, of people (De-Shalit 2005; Page 2007; Baier 2010).

A fourth feature of cases in which our actions might have harmful long-term effects is cluelessness. The term has been coined by Lenman (2000), who discusses it as a challenge for consequentialists. Cluelessness denotes our inability to account morally for indirect and unpredictable effects of our behaviour. As MacAskill and Mogensen (2019) have recently argued, however, indirect and unpredictable effects of our behaviour also pose difficulties for deontologists who defend the DDA.

## 1.2 The Context of Discussion

Much of non-consequentialist moral thought in the last few decades has addressed the question under which circumstances it can be permissible to harm others. At the same time, moral and legal theorists have discussed the concept of harm, in order to correctly identify instances of harming.

The literature on the concept of harm is broadly divided into two main camps. Both camps try to capture an intuitive understanding of harm, which we might roughly state as saying that someone is harmed when their welfare is infringed upon (I will get a bit more precise in the next chapter).[4] Common examples for harm are instances of physical or psychological injury or illness.

In the first camp, some philosophers have defended a comparative view on the nature of harm (e.g. Klocksiem 2012; N. Hanna 2016). According to the comparative view, A harms B if and only if A makes B worse off than B otherwise would have been. In the second camp, others have defended a non-comparative view. According to the standard version, A harms B if and only if A causes B to be in a non-comparatively bad state (e.g. Shiffrin 1999; Harman 2009).

Both views face well-known difficulties, and proponents of both camps have proposed amended versions of their respective views, only for their opponents to reply with new criticism. The debate tends to focus on which of the two camps can provide an account of harm that can correctly identify instances of harm in line with ordinary intuitions about cases (or, failing that, provide an explanation for why our intuitions regarding certain cases are mistaken).

Both camps tend to answer two questions, which are seldom explicitly distinguished.[5] First, when is a victim in a harmed condition? Second, when does an agent's behaviour count as an act of harming? Distinguishing these questions helps to avoid confusion that arises when the answer to one question is evaluated by its ability to answer the other question. To illustrate the difference

[4] This well-being view is implicitly assumed in much of the literature (for an explicit defence, see Tadros (2014)).

[5] There are some exceptions, which I list in footnote 16.

between a ‘harmed condition’ and a ‘harming behaviour’, we can say that the harmed condition consists in what happens to the victim (for example, a shot wound). The harming behaviour is the agent’s act (for example, shooting a gun). Once we distinguish these questions, we can see that an account of harm needs two elements: on the one hand, an account of harmful outcomes, and on the other hand, an account of harming behaviour. I elaborate on this in the first chapter.

The focus on providing the correct account of harm leaves another question under-examined in comparison: namely, what makes some instances of harming morally worse than others? In other words, what are the factors that strengthen or weaken moral reasons against harming (Gardner 2017)? One might think that different accounts of harm do not fundamentally disagree, but rather pick on different factors that strengthen moral reasons against harming. Indeed, this is the line that I will take in the second chapter.

One candidate for a factor that influences the strength of the reason against harming is the way in which the agent’s behaviour relates to the causal sequence leading to harm. Compare

> (Push) Agent pushes a boulder which rolls over Victim, crushing her to death.
>
> (Non-Intervention) A boulder is rolling towards Victim. Agent could stop the boulder but does not do so. The boulder rolls over Victim, crushing her to death. (These cases are variants of cases discussed in (Woollard 2015; 2012b; Bennett 1993, 76–77).)

Intuitively, Agent *does* harm in (Push) and merely *allows* harm in (Non-Intervention). According to the DDA, this difference is morally relevant, in the sense that it can influence what it is permissible for an agent to do. For example, imagine that Agent needs to rush to the hospital to receive urgent treatment to save their own life (Woollard 2015, 9). In this situation, Agent might be justified to allow harm, but not to do harm. It might be permissible for Agent to fail to stop the boulder in (Non-Intervention) in order to get to the hospital in time, but it might be impermissible for Agent to push the boulder in (Push) if the boulder happens to block the road to the hospital.

Some philosophers have rejected the constraint despite its intuitive appeal, arguing that there is no clear way to distinguish doing from allowing harm, or if there is, that this distinction does not possess moral relevance. For example, some have advanced contrast cases that supposedly show that some cases of doing and allowing harm are intuitively morally equivalent (Rachels 1994; Tooley 2006).[6] Other philosophers have defended the distinction, offering accounts of distinguishing doing harm from merely allowing harm, and arguments for the moral relevance of the distinction (for an overview, see Woollard and Howard-Snyder 2016).

[6] See e.g. Frowe (2006) for the response that such cases are not intuitively morally equivalent.

Most recently, Fiona Woollard (2015; 2012b; 2012c), following Philippa Foot (1967), has offered an account of the doing/allowing distinction and its moral relevance. She distinguishes doing from allowing harm in terms of the causal relation between an agent's behaviour and a harm. According to Woollard, an agent's behaviour is *doing* harm if and only if a fact about the behaviour is part of a sequence leading to harm. An agent's behaviour is *merely allowing* harm if and only if a fact about the behaviour is relevant to, but not part of, a harmful sequence. Only *substantial* facts can be part of a sequence. Substantial facts have some feature that make them more than a mere background condition. An example for a substantial fact would be a fact about a voluntary bodily movement, such as Agent's pushing a boulder in (Push).

Next, Woollard argues that agents have a right of self-ownership, which entitles them to prima facie protection from both causal and normative imposition. The Doctrine of Doing and Allowing (DDA), the claim that doing harm is harder to justify than merely allowing harm, provides this protection in two ways. First, by including constraints against doing harm. This protects agents from *causal imposition* by other agents' harm doing. Second, by including permissions to allow harm. This protects agents from *normative imposition* by being put under obligations to put themselves at someone else's use.

In the following, I will mainly focus on Woollard's defence of DDA. This is because it is one of the most comprehensive accounts of the DDA to date, and because it fits in well with my account of harm which emphasizes the role that causal sequences play in defining and assigning responsibility for harmful acts. However, it is worth noting that other defences of the DDA have been put forward. Some of these explicitly provide an indirect defence of the DDA – i.e. they argue that even those who deny that the doing / allowing distinction carries an intrinsic moral relevance should acknowledge the moral relevance of the distinction in practical decision making. For example, McCarthy (2000) defends the DDA on the basis that accepting the DDA is necessary for those who accept moral options and agents' abilities to exercise those options. Woollard (2010) also presents an indirect defence of the DDA, arguing that it enables more responsible deliberation of potential harmful conduct. Scheffler (2004) seems to argue that our moral practice already presupposes (something like) the DDA. Relatedly, Haydar (2010) points out that rejecting the DDA might itself have implausible implications, namely that non-optimal harmdoing can sometimes be permissible. Finally, Kamm (2007) defends the DDA appealing to the intuitively plausible difference between positive and negative rights (see also Draper (2005)).[7]

[7] While the DDA is taken to be a key component of deontological moral theories, the DDA is likely also compatible with other moral theories. See Woollard (2015) for arguments that the DDA is compatible with Scanlon's contractualism and Hooker's rule-consequentialism, and Liu (2012; see also Haydar 2002) for an

The non-consequentialist debate about the permissibility of harming is very abstract. Much of the discussion centres on relatively artificial cases, such as (Push) and (Non-Intervention).[8] In these cases, individual agents face a one-off decision in which all the options and possible outcomes are known, the decision clearly and directly causes the outcome, and it affects someone for the worse in a way that is clearly harmful. This focus has been noted, and criticized, e.g. by Barbara Fried (2012a). The lack of attention for non-idealized cases leaves defenders of the DDA without a comprehensive account of its implications for real world action guidance, and, as I will argue, it also leaves them vulnerable to the charge of being unable to theoretically account for the moral status of real-world cases.

## 1.3 Terminology

I understand the DDA as the claim that agents have, everything else being equal, *stronger moral reasons against doing than against merely allowing (i.e. failing to prevent) harm*. I formulate the DDA in terms of reasons (as do MacAskill and Mogensen 2019, 6), rather than in terms of constraints or pro tanto duties (Kamm 2007, 14), or justification (Woollard 2015, 12–13; Quinn 1989a, 291). There are two reasons for this terminological choice. First, it connects straightforwardly with the framework for harm-based reasons that I give in the first part of the thesis. Second, the language of reasons lends itself to be used in real-world moral deliberation, where different morally relevant considerations must be weighed up against each other. Ultimately, however, I take these different terminologies to be compatible. They are different ways of expressing how the DDA influences the moral status of acts (such as the act's being permissible, impermissible, required, or supererogatory).

I assume in the following that the DDA can influence the moral status of acts in two ways: by putting a *constraint* on harm doings, and a *prerogative* on harm allowings.[9] (However, in this thesis, I focus almost exclusively on the constraint side of the DDA.[10])

argument that the DDA can fit in a consequentialist framework. In the more specific context of moral obligations towards the future, see Mulgan (2006, 153–54) for an argument that his rule-consequentialist account of obligations towards the future can accommodate the doing / allowing distinction.

[8] For more examples, consider the case of killing a man vs. letting a man die to help others in (Foot 1967, 4), Rescue I and II in (Foot 2002, 81), Rescue III and IV (Quinn 1989a, 298–99).

[9] I take this understanding from Fiona Woollard, who argues that the DDA protects from two types of imposition: 'agents are not permitted to causally impose on patients in a harmful way (doing harm is forbidden); patients cannot normatively impose upon agents in a harmful way (allowing harm is permissible)' (Woollard 2015, 102).

[10] Except for my argument in section 6.4.

As a constraint, the DDA says that doing harm is sometimes impermissible, even in instances where doing harm would minimize total harm. For example, many people think that it is impermissible to kill one person to prevent two others from being killed. We can use different terminology to explain why killing the one is impermissible. We might say that the act of preventing two others from being killed does not justify the act of killing the one, or that the pro tanto duty not to do harm the one is stronger than the pro tanto duty to aid the two,[11] or (the formulation that I will use in the following) that the reason against doing harm to the one is stronger than the reason against merely allowing harm to the two.[12] I will refer to this claim as the *Doing Thesis*.

As a prerogative, the DDA says that allowing (failing to prevent) harm can sometimes be permissible, even in instances where agents can prevent severe harm at minor cost to themselves. For example, many people think that it is permissible to spend money on a honeymoon rather than donating it to charity to relieve poverty. Again, we can use different terminology to explain why spending the money on a honeymoon is permissible. We might say that failing to donate the money is justifiable, that the pro tanto duty to aid does not amount to a moral requirement to donate the money, or that the reason to prevent harm is not strong enough to require the agent to aid in this case. I will refer to this claim as the *Allowing Thesis*.

To clarify, I do not think that these formulations are necessarily equivalent, only that the moral relevance of the DDA can be expressed in different terms. As I understand the DDA, it does not directly influence the moral status of behaviours. Rather, it states a difference in the moral relevance of doing harm and allowing harm. This difference can explain why cases of doing harm are sometimes impermissible, where otherwise equivalent cases of allowing harm are permissible.

On the understanding of the DDA that I have sketched, the DDA states a difference in the comparative strength of the pro tanto duty not to do harm (and corresponding reasons) and the

[11] Ross's notion of prima facie duties can be interpreted in terms of reasons, as Philip Stratton-Lake notes in his introduction to Ross's *The Right and the Good*: 'The Foundations of Ethics. So prima facie duties should be understood as features that give us genuine (not merely apparent) moral reason to do certain actions' (Ross 2002, xxxiv).

[12] What about cases where doing harm does not prevent further harm, but instead generates significant benefits? As pointed out by Kagan (1989) and MacAskill and Mogensen (2019), the constraint against doing harm needs to be able to not only compare harm doings and harm preventions, but also harm doings and provisions of benefits. I follow them in arguing that the constraint against doing harm is best understood as the conjunction between the DDA and a claim about the comparative reasons against harming and for doing good. However, my arguments will focus on cases of harming that do not involve benefits. I will therefore largely ignore discussion of benefits for the purpose of this thesis and understand "the constraint against doing harm", unless otherwise specified, to refer to the constraint side of the DDA that I call the Doing Thesis.

pro tanto duty to aid (or prevent harm) (and corresponding reasons), such that doing harm is harder to justify than failing to prevent harm. I assume that we are comparing clearly specified cases, in which agents have reasons or duties to either prevent or refrain from doing harm, and that are otherwise equivalent.

This understanding of the DDA is compatible with the view that duties not to do harm and duties to benefit are different in scope. Such a view might derive from Fiona Woollard's distinction between maximal defeasible duties from non-maximal defeasible duties:

> I suggest that when B is a general type of behaviour, we distinguish between maximal and non-maximal defeasible duties to B:
>
> An agent has a maximal defeasible duty to B if and only if, she has a defeasible duty to B to the greatest extent possible. Thus for any action, ϕ, if ϕ is an instance of B-ing, then she has a defeasible duty to ϕ or to perform some alternative action which involves B-ing to the same or greater extent.
>
> An agent has a non-maximal defeasible duty to B if and only if, she has a defeasible duty to B but does not have a defeasible duty to B to the greatest possible extent. (Woollard 2018, 139–40).

The difference between maximal and non-maximal defeasible duties is that 'a difference in the scope of liability to be called upon to justify one's actions' (Woollard 2018, 140).

It seems plausible that the duty not to harm tends more towards being maximally defeasible than the duty to aid (prevent harm).[13] It seems plausible, for example, that I have a general obligation not to kill others, *and* that I have a specific obligation not to kill in all situations when I could do so. (If I nonetheless kill someone, I need to justify my action, and in the absence of sufficient justification, my act is impermissible.) However, while many would agree that I have a general obligation to help those in need, many would think that I do *not* have a specific obligation to help in all situations when I could do so. (If I fail to help in such situations, I do not always need to justify my action, and my action need not be impermissible.)

Let me make some clarifications about the scope of my argument before I proceed. While the DDA has moral relevance, it does not have *direct* implications for the moral status of acts. Some acts of doing harm are permissible. Some acts of preventing harm are impermissible. This is

[13] I say "tend towards", because I do not think that the duty not to harm is maximally defeasible: we do not have to justify some of the minor harms we impose on others (such as annoying the neighbour by mowing the lawn at a perfectly acceptable time).

because many different principles and considerations, apart from the DDA, can weigh in and influence the moral status of these acts. Once these other influencing factors are assumed to be constant, the cases are equalized. Methodologically, such equalized cases can be used to test intuitions about the moral relevance of cases (Kamm 2000, 658).

Relatedly, harms are different from wrongs. Someone is wronged when their moral status is infringed upon in a way that is deemed impermissible, for example, instances in which someone's rights are violated, their autonomy is constrained, or they are otherwise not respected as full moral agents. Not all harms are inflicted wrongfully. Examples of non-wrongful harms might be proportionate self-defence, or lawful imprisonment. Not all wrongs are harmful. Examples of non-harmful wrongs might be paternalistic interventions, or unnoticed trespassing.

Finally, it is worth noting that since I am concerned with real-world decision making, by "moral reasons" I have in mind what has been called (e.g. by Smith (2010, 66)) *subjective* (i.e. belief- or evidence-based), rather than *objective* (i.e. fact-based), reasons. Agents do not always have access to objective reasons, since they seldom know for a fact how their behaviour will affect others.

## 1.4 Chapter Overview

The first part of the thesis develops an account of harm. This account of harm addresses two questions. The first question is about the nature of harm: When is someone harmed? The second question is about moral reasons against harming: When does a behaviour count as 'harming' in a morally relevant sense?

I address the first question about the nature of harm in Chapter 2. Comparative accounts say that being worse off constitutes harm. The temporal version of the comparative account is seldom taken seriously, due to apparently fatal counterexamples. I defend the temporal version against these counterexamples and show that it is in fact more plausible than the much more prominent counterfactual version of the account. Non-comparative accounts say that being badly off constitutes harm. I defend a simple version of the non-comparative account against a threshold version. I further argue that the temporal comparative and the simple non-comparative account specify different ways in which someone can suffer harm: by losing well-being, and by suffering ill-being, respectively. The upshot is that we should combine them in a novel hybrid account of harm. I argue that, unlike its competitors, the hybrid account is extensionally adequate and can be presented as a unified view on the nature of harm.

I address the second question about moral reasons against harming in Chapter 3. I develop a two-dimensional account of harming. Causal accounts of harming imply that we have stronger reasons against doing than against merely failing to prevent harm. Counterfactual comparative accounts of harming imply that we have stronger reasons against doing harm when our behaviour makes a difference to those who suffer harm, as opposed to when they will suffer regardless of what we do. I develop a novel account of harming, which I call the Two-Dimensional View. On this view, causally contributing and making a difference to harm both matter morally. I argue that both dimensions are necessary to correctly identify harm-based reasons. Roughly, only causal considerations explain the difference between doing and allowing harm, and only counterfactual comparative considerations explain the difference between allowing and being irrelevant to harm. Moreover, I show that recent arguments to the conclusion that either dimension grounds stronger reasons against harming than the other are unsuccessful.

The second part of the thesis focusses on the moral relevance of the distinction between doing and merely allowing harm. I discuss three challenges for the Doctrine of Doing and Allowing (DDA) and explore ways in which deontologists can respond to these challenges.

In Chapter 4, I discuss a challenge for the DDA which has recently been posed by Ingmar Persson (2013) and Jason Hanna (2015a): how can the DDA account for so-called cases of letting oneself do harm? I show that cases of letting oneself do harm are structurally similar to real-world cases such as climate change. I then explore different ways in which deontologists can solve this challenge and argue that the most promising way to conceive of cases of letting oneself do harm is as non-standard cases of allowing harm.

In Chapter 5, I explore implications of the DDA for cases in which agents offset their own risk impositions. In the literature on deontological constraints under risk, two views have been proposed. The time-relative view implies that agents have a reason against doing harm in the present choice situation. The time-neutral view implies that agents have a reason to minimize their own harm doings. I show that time-relativists and time-neutralists both have difficulties accounting for cases in which agents offset their own risk impositions. I then sketch an alternative view. According to the Relation-Centred View, for every possible victim V, and for every harm (in some respect and at some time) H, agents have a reason against doing something that might constitute doing H to V. This view gives the right result in offsetting cases, and it can explain intuitions that motivate its competitors.

In Chapter 6, I defend three claims which together clarify how the DDA applies to forward-looking, long-term decision making. First, the DDA applies only to harm that is sufficiently proximate to the agent's behaviour. Second, agents have limited prima facie permissions to impose risks of harm through everyday behaviour. Third, the DDA does not tell agents to refrain from behaviour that does not increase anyone's ex ante risk of suffering harm. Furthermore, I argue that this third claim enables defenders of the constraint to solve a challenge regarding long-term consequences of actions that has recently been put forward by MacAskill and Mogensen (2019). The upshot is that considerations of ex ante risks play a crucial role within an understanding of the constraint as an action-guiding principle in real-world cases.

Together, these chapters clarify how theoretical debates on the nature and strength of harm-based moral reasons can be used to evaluate real-world cases, specifically cases of long-term harm doing. I aim to show that the notion of harm deserves its central role in moral theory and to defend the DDA as a principle that is applicable to and informative in real-world moral decision making.

This thesis does not discuss several questions in the vicinity of doing harm and future generations that are doubtlessly important. Perhaps most importantly, I focus on the behaviour of individual moral agents, and mostly ignore questions of collective and institutional agency. These restrictions of my argument are partly motivated for reasons of time and space, but there is also a more principled reason for exploring the harm-based moral reasons for individual agents as a starting point for these other considerations. A better understanding of how moral principles and concepts apply to individual agents will likely help us understand how such principles and concepts can be applied to different moral agents.

# Chapter 2 A Hybrid Account of Harm

## 2.1 Introduction

Two questions are of central importance in the philosophy of harm. First, under which conditions can we say that someone suffers[14] a *harm,* in the sense of being in a harmed state? And second, under which conditions can we say that one person *harms* another, in the sense of bringing about the harmed state? To illustrate, imagine that Ann throws a stone at Bob, thereby breaking his nose. We can now ask whether, and if so, in virtue of which facts, Bob is in a harmed state.[15] We can also ask whether, and if so, in virtue of which facts, Ann has harmed Bob.

These questions are not always clearly distinguished in the literature.[16] As a result, it is easily overlooked that the notion of harm is distinct from the notion of harming and deserves separate attention. Perhaps this is not surprising, since the direct objects of moral and legal prescriptions are actions (e.g. whether Ann ought to throw a stone at Bob or whether Ann should be held liable for doing so) rather than states of affairs (Feinberg 1984, 31).

However, it would be a mistake to conclude that we should pay less attention to the notion of harm than the notion of harming. Indeed, it seems intuitive that any account of harming presupposes an account of harm, since a behaviour (such as Ann's throwing the stone) counts as harming only if it is related, in an appropriate way, to an outcome that counts as a harm (such as Bob's broken nose).[17] If this correct, then the question of what constitutes harm should be answered alongside the question when a behaviour counts as harming, and both need to be

[14] I use "suffering harm" in the sense of "being the subject of harm". I do not presuppose that harm consists in consciously experiencing suffering (or that it consists in any other mental state, for that matter).

[15] For the most part of this chapter, I assume in line with much of the literature that we should conceive of harms as states of affairs. However, I briefly discuss an argument by Hanser (2008) for understanding harms as events in section 2.4.1.

[16] There are exceptions to this, especially among critics of counterfactual comparative accounts of harming. Notably, Hanser (1990, 57) suggests that accounts of harming correctly describe 'the morally relevant relation that holds between an action, on the one hand, and the event of someone's suffering a harm, on the other, just in case the action is at least prima facie objectionable owing to its role in bringing harm upon that person'. The distinction between harmful outcome and harmful behaviour is also noted by Feinberg (1984, 31) and Bradley (2012, 403) in discussion of an account of harm offered by Hanser (2008, 2011). Shiffrin (2012, 360) notes that focusing on states of harm 'may help to disentangle what significance the condition of harm itself (or of suffering, experiencing, or enduring harm, if you prefer) contributes and what is attributable to human agency, causation, and responsibility'. See also Woollard's (2012a, 685) distinction between 'what it is for an agent to be responsible for a harm' and 'what is it for a patient to suffer harm'.

[17] This is nicely described by Feinberg, who states that: 'Acts of harming then are the direct objects of the criminal law, not simply states of harm as such. From the legislative point of view, however, states of harm are fundamental, for they determine in part which acts are to count as acts of harming, and to become thereby proper targets of prohibitory legislation' (1984, 31).

answered in order to correctly identify instances in which one person's behaviour subjects another person to an outcome that constitutes harm, in a way that constitutes harming. A complete theory of harm should include both an account of harming, and an account of harm.[18]

In this chapter, I focus on the notion of harm. (I investigate the notion of harming in the next chapter.) So, when does someone suffer a harm? Consider the following accounts of harm:

> (Comparative) Agent suffers a harm if and only if Agent is worse off.
>
> (Non-Comparative) Agent suffers a harm if and only if Agent is badly off.[19]
>
> (Hybrid) Agent suffers a harm if and only if Agent is (i) worse off, or (ii) badly off.[20]

The comparative account is most often defended in its *counterfactual* version. In its standard formulation, the counterfactual comparative account says that an event harms Agent if and only if Agent would have been better off in the absence of this event.[21] The distinction between harm and harming is not usually made by proponents of the counterfactual comparative account, but their account can be formulated as an account of harm as follows: Agent suffers a harm if and only if there is an event such that Agent is worse off than Agent would have been in the absence of the event.[22]

Another version of the comparative account that tends to be quickly dismissed in the literature[23] is the *temporal* comparative account, which says that Agent suffers a harm if and only if Agent is

[18] Similar points have been made by Feinberg (1984, 31), Hanser (1990, 57), and Shiffrin (2012, 360).
[19] Accounts along these lines have been defended by (Harman 2004; 2009; Shiffrin 1999; Rivera-lópez 2009).
[20] While it has been suggested in the literature that an account of harm might have to combine comparative and non-comparative elements (see e.g. McMahan (2013, 8) and Woollard (2012a, 688)), the possibilities of such an account have not been comprehensively explored. An exception is Meyer's (2016) disjunctive notion of harm, which I discuss in section 2.4.
[21] Accounts along these lines have been defended by (Feinberg 1984; Norcross 2005; Klocksiem 2012; Purshouse 2016).
[22] Thanks to Anna Folland for suggesting this formulation.
[23] I am aware of three exceptions in the literature. Perhaps most emphatically, Perry (2003) defends the temporal comparative account. However, interestingly, Perry takes the temporal comparative account to be a necessary, but not a sufficient, condition for harm which needs supplementation by considerations regarding responsibility and liability (2003, 1292). Foddy (2014) defends a version of the temporal comparative account, which, however, relies on a hedonistic account of well-being and is thus not axiologically neutral. Moreover, I believe that Foddy's version of the temporal comparative account faces difficulties in cases such as (Alleviate Suffering) discussed below (I explain why in footnote 33). Finally, Velleman (2008) defends a temporal comparative account of harm and benefit. He acknowledges that his account cannot explain why future generations are harmed in non-identity cases; however, he suggests that this shows that 'because we cannot harm or benefit future persons via their inheritance […] our moral relation to them should not be conceived in terms of harm and benefit in the first place' (Velleman 2008, 244). As will become clear, I agree with Velleman about the temporal comparative account in non-identity cases but draw the different conclusion that this should motivate us to accept a different notion of harm, rather than giving up on harm-based explanations of the wrongness of actions in non-identity cases altogether.

worse off than Agent was previously. The *simple* non-comparative account stated above can be qualified further by introducing thresholds. On a *threshold* non-comparative account, Agent suffers a harm if and only if Agent is badly enough off, as defined by some threshold.

In this chapter, I argue that the most plausible version of the comparative account of harm is the temporal account, and the most plausible version of the non-comparative account of harm is the simple non-comparative account. However, neither account, by itself, is able to recognise all cases that plausibly constitute harms as such. I argue that we should therefore combine them in a *hybrid account* of harm. Roughly, the hybrid account that I defend says that agents suffer harm when their welfare is adversely affected. In the case of temporal comparative harms, adverse effects consist in the loss of well-being. In the case of simple non-comparative harms, adverse effects consist in the presence of ill-being. I argue that unlike its competitors, the hybrid account correctly classifies harms. One of the main upshots of the chapter is that taking the distinction between harm and harming seriously enables us to respond to apparent counterexamples against the temporal comparative account and provides us with a reason to reject the much more prominent counterfactual comparative account of harm.

This chapter proceeds as follows. In section 2.2, I start by arguing that the counterfactual comparative account is ill-suited to answer the question what constitutes harm. Instead, I defend the temporal comparative account of harm. The temporal comparative account can correctly identify losses of well-being as harms, but unfortunately not instances of ill-being. In section 2.3, I therefore turn to the non-comparative account of harm. I distinguish between the simple and a threshold version of the non-comparative account and argue that the threshold version fails as an attempt to make the non-comparative account extensionally adequate. However, while the non-comparative account of harm fails in correctly identifying losses of well-being as harms, it can correctly identify instances of ill-being as harms. In section 2.4, I argue that we should adopt a hybrid account of harm, which combines the temporal comparative and the simple non-comparative account of harm. What emerges is a hybrid account that is extensionally adequate and presents a unified account of the nature of harm.

Before we start, let me clarify the desiderata for a theory of harm. I borrow these from Bradley (2012, 359), together with the provision that the desiderata should be understood as a rough and incomplete guideline to keep our investigation on the right path, rather than as strictly necessary or sufficient conditions for suitable accounts of harm. First, accounts of harm should be extensionally adequate. This means that they should get the scope of harm right: they should not identify states as harms that are clearly not instances of harm, but they should also not miss out any states that clearly are instances of harm. Second, accounts of harm should be axiologically neutral. They should be compatible with different accounts of welfare, and not make substantial assumptions about the nature of welfare. Third, accounts of harm should be unified, in the sense that they 'should explain what all harms have in common' (Bradley 2012, 395).[24]

I am interested here primarily in *pro tanto* harms (harms in some respect), rather than *overall* harms. A decayed tooth might be overall beneficial for the patient who falls in love with the dentist and lives with the dentist happily ever after, but having a decayed tooth still counts as a pro tanto harm.[25]

Lastly, some notes on terminology. I will largely use "A suffers a harm" and "A is in a harmed state" synonymously. This is because I am primarily interested in the question of which states count as harm, and not at which moment in time the harm occurs. However, for anyone interested in the timing of harm, I suggest that on a stricter reading, we should understand "A suffers a harm" as "A enters a harmed state", whereas "A is in a harmed state" can refer to A's state at any time during which the harmful condition obtains.

## 2.2 Comparative Harm

The comparative account says that agents suffer harm if and only if they are worse off. Stated like this, however, the account is incomplete. In order to find out whether an agent suffers harm in some state of affairs, we need to compare how well off the agent is in this state of affairs with how well off the agent is in some comparison state. In other words, the comparative account identifies harm by comparing it to a baseline. Perhaps the most influential account of harm uses a counterfactual baseline:

> (Counterfactual Comparative) Agent suffers a harm if and only if there is an event such that Agent would have been better off in the absence of the event.

Recall that the counterfactual comparative account of harm is usually presented as a unified account of both harm and harming. The counterfactual comparative account of *harming* has frequently been criticized on the basis that there are cases in which a person seems to be harmed by an action without being made worse off by the action that brought the harm about. Consider:

[24] Bradley (2012, 395) lists four additional desiderata that, however, will not feature much in my discussion: prudential and normative importance of harm, amorality, and ontological neutrality.

[25] Pro tanto harms and overall harms are usually taken to be interdefinable: 'an event is overall harmful to someone iff its pro tanto harms to that person outweigh its pro tanto benefits to that person' (Bradley 2012, 393).

> (Pre-emption) If Agent does not break Victim's nose, Villain will do so. Agent breaks Victim's nose.[26]

> (Non-Identity) Agent chooses to conceive a child that will have a severe health condition, when Agent could instead have chosen to have a different child that will not have this condition.[27]

Intuitively, Agent in these cases acts wrongly, and the most natural explanation seems to be that Agent harms Victim in (Pre-emption), and her child in (Non-Identity). However, the counterfactual comparative account of harming implies that Agent does not harm Victim. Victim would still have ended up with a broken nose had Agent acted differently; and the child with the severe health condition would never have existed had Agent chosen to have a different child. Both Victim and the child are not worse off than they would have been, had Agent acted differently.

However, one might think that these problems only concern the counterfactual comparative account as a view on harming, but not as a view on harm.[28] For example, even if Agent's action does not harm Victim, there might be some other event such that without this event, Victim's nose would have remained intact. Victim might have decided to go to the pub, where Agent and Villain pick up a fight. Had Victim not decided to go to the pub, her nose would not have been broken. Since there is an event such that without this event, Victim's nose would have remained intact, the broken nose constitutes a harm, according to the counterfactual comparative account.

In response, it is unclear whether there is such an event for every pre-emption case. Imagine, for example, that a huge meteor destroys Earth. There are some events of which it is true to say that, had they not occurred, the inhabitants of Earth would not have suffered harm (e.g. some events leading to the meteor formation). However, for all these events, in the relevant possible world that does not contain this event, a different event would have occurred such that a different meteor would have destroyed Earth. Such a world is at least metaphysically possible, and perhaps empirically possible. And surely, we would want to say that the inhabitants of Earth suffer harm. The counterfactual comparative account of harm does not seem to support this judgement. It is even more difficult to see what the relevant event without which the child would have remained

[26] For responses to the pre-emption problem, see e.g. Feit (2015), Boonin (2014), J. Hanna (2016), Klocksiem (2012). For criticism of these responses, see Johansson and Risberg (2019).

[27] Non-Identity cases have famously been introduced in moral philosophy by Parfit (1984), Schwartz (1978), and Adams (1979); see also Kavka (1982). Since then, a huge and ongoing debate has emerged (for an overview, see Roberts (2010)).

[28] Thanks to Jens Johansson and Anna Folland for pressing me to clarify this.

unharmed would be in a non-identity case. After all, the point of this case is that there is no possible world in which *this* child would have existed in good health.

To put the point more generally, the problem in these cases is that the counterfactual comparative account of harm relies on comparisons with possible worlds in which events do not occur. However, it seems that it is at least possible that there are cases in which for every possible event, in the nearest world in which this event does not occur, a different event occurs that would leave Victim just as badly off. In these cases, counterfactual comparativists would be committed to the implausible conclusion that Victim has not been harmed.

To add to these difficulties, consider that it is sometimes undetermined what would have happened, had an event not occurred:

> (Indeterminacy) A breaks B’s nose. There is no fact of the matter what would have happened, had A not broken B’s nose. We can stipulate that it would have been 50% likely that C would have broken B’s nose, and 50% likely that nothing would have happened.[29]

Clearly, B suffers harm. However, we can stipulate that there is no event such that had this event not occurred, it would have been certain that nothing would have happened. (We can imagine that no event that would have prevented A from breaking B’s nose would also have made it impossible for C to break B’s nose.) The counterfactual comparative account of harm then seems to imply that B does not suffer harm. But this seems clearly wrong.

I suggest that the counterfactual comparative account is plausible as an account of harming, but much less plausible as an account of harm.

The counterfactual comparative account explains the plausible claim that agents who make a difference to a Victim’s harmful condition are morally responsible for this condition.[30] The counterfactual comparative account allows us, in most cases, to attribute responsibility for an outcome to agents who have made a difference to this outcome. Pre-emption and non-identity cases are cases in which the agents do not seem to make a difference to another person’s suffering: this makes it more difficult to determine the degree of their responsibility for the outcome.

I suggest that the connection between events and outcomes that the counterfactual comparative account uses to identify harms does not seem constitutive to harms. It seems less plausible to

[29] Gardner (2017, 77) gives a similar case as an objection to comparative counterfactual accounts.

[30] I will say more about this claim in the next chapter.

think that whether an agent suffers harm in an outcome depends on counterfactual dependence relations between events and the outcome than to think that it depends on properties of the outcome itself. To illustrate, imagine two possible scenarios in which Anne suffers harm. The difference is that in the first scenario, the outcome counterfactually depends upon Bert's action, whereas in the second scenario, the outcome does not counterfactually depend upon Bert's action. It is plausible to assume that there might be a difference between the two scenarios regarding whether Bert harmed Anne. But it is not plausible to assume that there is a difference between the two scenarios regarding whether Anne suffers harm. Whether Anne suffers harm is a question that needs to be settled independently of whether Anne's condition counterfactually depends upon Bert's action. The same might be said about anyone else's action, or other events.

The temporal comparative view is sometimes mentioned in the literature as an alternative to the counterfactual comparative view. It replaces[31] the counterfactual with a temporal baseline, which describes the agent's state before suffering harm:

> (Temporal Comparative) Agent suffers harm if and only if Agent is worse off than Agent was at some earlier time.[32]

The temporal account is rarely defended as a serious alternative to the counterfactual account of harm. It seems that it can easily be dismissed with reference to cases such as the following:

> (Alleviate Suffering) The condition of a terminally ill patient is rapidly deteriorating. Doctor can alleviate her suffering by administering a drug, but she is unable to stop or reverse the deterioration. Doctor administers the drug, and Patient's condition deteriorates at a slightly slower pace.[33]

[31] Feinberg (1986, 151) briefly considers the possibility of an account of harm that *combines* a counterfactual and a temporal comparative baseline, and shows that such a combined view should be rejected, because it cannot explain cases of harm such as his "Miss America" case: 'Suppose that A's gang abducts B, the beauty queen, on the eve of the Miss America contest […] but that if they had not done so, then Cs gang, in a quite independent conspiracy, would have abducted her only a few minutes later. In any case, B does not win the contest'. For an argument, pace Feinberg, that the beauty queen is harmed in a temporally comparative sense by being deprived of an opportunity to win the contest, see Perry (2003, 1293).

[32] Here, the question arises which earlier moment in time should serve as a comparison baseline. The simplest version of the view compares the agent's well-being right before and after a salient event that is deemed to be harmful. A perhaps more promising view determines the baseline with regard to the context of a case. More needs to be said here. Since my focus here is solely on determining whether the account can correctly identify instances of harm in a given context, however, I assume the broadest context possible: an agent can suffer harm in a temporal comparative sense if and only if they are worse off (in some respect) than they were (in this respect) at any earlier moment. Thanks to Olle Risberg for pressing me on this.

[33] I have adapted this case from Norcross (2005, 149). I think that (Alleviate Suffering) poses a significant problem for Foddy's version of the temporal account of harm. Here is Foddy's account: 'An event benefits me if: (1) it moves the world from a state of containing less pleasure to a state of containing more pleasure, or if it moves the world from a state of having more pain to a state of having less pain, and (2) the change in

The temporal account seems to imply that Doctor harms Patient, because Patient is worse off after receiving the drug than Patient was before. This is very implausible. Intuitively, it seems that on the contrary, Doctor has benefitted Patient.

I deny that the temporal account of harm has the implausible implication that Doctor harms Patient. The temporal account merely implies that Patient suffers harm (admittedly, less severe harm than Patient would otherwise have suffered). This seems like the right result. What is implausible is not that Patient suffers harm.[34] What is implausible is that Doctor's action counts as harming. But remember that the temporal account of harm does not say anything about whether Doctor's action stands in a 'harming' relation to Patient's state. It might be the case that Patient is in a harmed state, but Doctor has not harmed Patient.

In (Alleviate Suffering), the objection against the temporal comparative account is that it overgenerates harm, i.e. finds harm where there is none. Another objection against the temporal account is that it undergenerates harm, i.e. finds no harm where there is harm. Consider:

> (Delayed Recovery) Patient is about to recover when Doctor administers a drug that delays Patient's recovery.[35]

It seems that the temporal account implies that Patient does not suffer harm in this case. After all, Patient is just as badly off after Doctor administered the drug than before Doctor administered the drug. In response, the defender of the temporal account can point out that a worsening in recovery prospects is certainly likely to have a significant impact on Patient's mental and physical well-being, and thereby cause harm. However, this does not solve the problem: even under the assumption that there is no such additional harm (we might assume that Patient is unaware of what is happening), it still seems that Patient suffers harm in (Delayed Recovery).

Thomson (2011, 444–45) suggests another response on behalf of the temporal comparativist: Patient might be worse off in virtue of having worse recovery prospects than they had before

the world's state is brought about by a change in the amount of pleasure or pain experienced by me' (Foddy 2014, 158). However, in (Alleviate Suffering), Patient experiences neither more pleasure nor less pain after receiving the drug than before receiving the drug. So, Foddy's account cannot explain why Doctor benefits Patient in (Alleviate Suffering).

[34] There is an additional difficulty here, which I will only mention briefly and discuss in more detail in section 2.4. The temporal account correctly implies that Patient suffers harm. However, it can be objected that the temporal account does not correctly explain why Patient is in a harmed condition. It seems that Patient is in a harmed condition because she is in ill health, and not because her health is worse than it was before. I think that this is correct. Therefore, I suggest that the scope of the temporal account should be restricted to harms that consist in losses of well-being (with the presence of ill-being being a different kind of harm).

[35] I have adapted this case from Holtug (2002, 368). A similar case is given by Hanser (2008, 429).

being administered the drug.[36] However, Rabenberg (2014, 18) objects that this response will not do, for three reasons.

First, Rabenberg argues that prospects of future benefits aren't desirable for their own sake, rather, they are desirable for the sake of the future benefits, and thus cannot be components of well-being. However, this objection is not fatal for temporal comparativists. The defender of the temporal account can deny this. Good prospects can be desirable for reasons beyond their promise of future benefits, namely insofar as they enable long-term planning and encourage optimism towards the future.

Second, Rabenberg argues that a maximally well-off person has a 0 percent chance of becoming better off. But if better prospects are a component of well-being, then there would never be a maximally well-off person, since such a person would have to have a 100 percent chance of becoming better off. However, the defender of the temporal account might suggest that not 'better prospects' but 'good prospects' are a component of well-being. For the maximally well-off person, the prospects of remaining in her current state constitute good prospects. Not so for Patient.

Rabenberg's third objection is that Thomson's proposal commits the temporal comparativist to the view that a person suffers harm whenever she does not receive a benefit (since not receiving a benefit that she would otherwise have received lowers her prospects). Not receiving a birthday present, for example, would then constitute harm. This seems implausible.

However, this implausible implication only follows if temporal comparativists understand "prospect" in counterfactual terms. They should not do so. Understanding "prospects" counterfactually invites contradictions: According to the prospect approach, failing to receive a benefit would only count as harm if the person's prospects explicitly included that benefit (otherwise failing to receive a benefit would not 'worsen' their prospects). But if a person's prospects include a future benefit, then we should not call it 'benefit', for benefits should be things that *better* someone's prospects (or make good prospects more likely, or something like that).[37] The real force of the objection seems to be to point out that the prospect response needs to find *some* way of defining prospects, in a way that is not (standardly) counterfactual. But the

[36] Thomson's response here is similar to Velleman's temporal account of harm. According to Velleman's account, an agent is harmed when her interest are infringed, by having either her current quality of life or future prospects adversely affected (Velleman 2008, 243).

[37] Another reason for why prospects should not be understood counterfactually is given by Tadros (2014, 184), namely that this would disable the temporal comparative view to deal with pre-emption cases, and thereby give up on the advantage of the temporal comparative view over the counterfactual comparative view.

reply does not show that temporal comparativists could not, in principle, come up with such a definition. I conclude that Rabenberg's criticism, while pressing important points, is not necessarily fatal to the prospect response.

Let me give what I think is a more powerful reply on behalf of the temporal comparativist to cases like (Delayed Recovery). This reply is very straightforward. Of course, someone suffers harm when they are ill. Patient suffers harm in virtue of being worse off than Patient was before falling ill and needing treatment. As far as the temporal comparative account of harm is concerned, when Doctor prolongs Patient's suffering, Doctor causes Patient to remain in a harmful state for longer.[38] Because severity of harm arguably increases with the duration of suffering, what Doctor's action changes is that the harm that Agent suffers is more severe. But Doctor's action does not make a difference as to whether Agent suffers harm. (If it is true that Agent not only suffers harm, but Doctor also harms Patient, and if the temporal comparative account of harming cannot explain why this is the case, then this might show that the temporal comparative account is not as good an account of harming as it is of harm.)

One might object that this response raises the further difficulty of how to individuate harms. Intuitively, Agent in (Delayed Recovery) suffers harm twice: first, when falling ill, and second, when Doctor administers the drug.[39] But on the understanding of the temporal comparative view that I have advanced, Agent seems to suffer harm only once, namely when falling ill. Agent then remains in this harmful state unless and until Agent returns to the previous welfare level. I think that this is true. Recall that I understand "suffering harm" here in the sense of "entering or being in a harmful condition/a harmed state" rather than in the sense of "being harmed by an event" or "being subject to a harmful event". While I agree with the objector that the temporal comparative account says that Agent is already in a harmed state when Doctor administers the drug due to falling ill, the account can still acknowledge that Agent is then also subject to the harmful event "being given a drug that slows down recovery". This event is harmful in virtue of worsening the harmful state that Agent is in, by prolonging the duration and intensity of what makes the state of affairs harmful.

Unfortunately, there is a kind of case that is more deeply troubling for the temporal comparativist:

[38] A suffers harm at a certain time, but A might remain in the resulting harmed state for some amount of time. The severity of harm is arguably dependent on both the magnitude of harm (the difference between the agent's current and previous level of well-being) and the amount of time spent in the harmed condition (and perhaps other factors as well, such as the absolute badness of the agent's current condition).

[39] Thanks to Jessica Pepp for raising this point.

> (Bad Start in Life) Patient was born in a certain condition and has never been better off with respect to this condition.[40]

No matter how badly off Patient is, if Patient has always been this badly off, the temporal comparativist will have to say that Patient is not suffering any harm. But this is implausible.

Stephen Perry (2003) suggests a tentative reply on behalf of the temporal comparativist to these kinds of cases. The temporal comparativist might argue that if Patient had the 'natural potential' (Perry 2003, 1298) to be born without the condition, then an interference with the development of this potential is a harm. This seems similar to Thomson's suggestion about counting worse prospects as harm.

However, the notion of a 'natural potential' is problematic. For example, changes in genetic material might result in a foetus lacking the potential to develop some condition. But surely, unfavourably changing someone's genetic makeup can harm the resulting person. Perry's suggestion could only explain this by presupposing a notion of some intrinsic natural potential of human beings in the sense of a level of welfare that we can reasonably expect any human being to have. But such an assumption would import a non-comparative standard of harm into the theory. Assuming a natural potential of human beings would thereby turn Perry's theory into some version of a hybrid account.

Bennett Foddy (2014) bites the bullet with regard to cases such as (Bad Start in Life), but argues, against Thomson, that such cases are not fatal for the temporal comparative account. Foddy advances a version of the temporal account according to which a person's existence is overall beneficial for this person. His account implies that in cases like (Bad Start in Life), Patient does not suffer harm. Rather, Patient has received a benefit, even if the benefit would have been even greater had Patient been born without the condition.

However, Foddy himself admits that this account is hard to square with our intuitions about harms (2014, 163). Moreover, I think that Foddy's response seems even more implausible once we consider lives that are, overall, not worth living. Imagine someone who is badly off their whole life. Foddy's view implies that there is no harm in that person's life, merely the absence of benefits. But this seems incredibly implausible. (For another difficulty with Foddy's approach, see footnote 33.)

[40] Similar cases are brought up in the literature as supposedly fatal objections to the temporal comparative account. See e.g. Thomson (2011, 445), Holtug (2002, 369) and Petersen (2014, 204).

To sum up, the counterfactual comparative account is too narrow. It cannot identify instances of harm in which a harmful outcome does not counterfactually depend on any particular event.

More generally, it cannot accommodate the plausible idea that whether states of affairs are harmful should be determined at least partly through properties of the state of affairs, rather than exclusively through contributing events. The notion of harm is fundamental to the notion of harming. This is because it identifies harms with reference to, rather than independently of, the harming behaviour. The temporal comparative account avoids both problems. However, the temporal comparative account, too, is too narrow. It cannot identify instances of harm in which someone is born into a harmful state.

In the remainder of this section, I discuss whether another version of the comparative account, namely Gardner's existence account of harm, can solve these problems. Gardner defines harm as follows:

> Harm (def.): A state of affairs, T, is a harm for an individual, S, if and only if (i) There is an essential component of T that is a condition with respect to which S can be intrinsically better or worse off; and (ii) If S existed and T had not obtained, then S would be better off with respect to that condition (Gardner 2015, 434).[41]

According to this existence account of harm, (Non-Identity) and (Pre-emption) are instances of harm. This is because the victims in non-identity cases and pre-emption cases such as the ones that I have described are (i) intrinsically badly off in virtue of their adverse health conditions; and (ii) if they existed and did not have these adverse health conditions, then they would be in better health. So, Gardner's account, unlike the counterfactual comparative account, can correctly classify those cases.

However, Gardner's account has difficulties identifying harm in cases in which agents are made worse off, without being badly off. I would like to show this by discussing a case that Gardner uses to illustrate her account:

[41] The comparison baseline suggested by Gardner seems similar to a baseline suggested by Woodward: 'the sort of analysis I have been exploring explains the wrongfulness of the choice of the nuclear policy by focusing on the difference between the situation of the nuclear people under the choice of the nuclear policy (when they are killed, injured, etc.) and an (unattainable) baseline situation in which the nuclear people exist and these violations of their rights do not occur' (1986, 817).

> (Loss of Fortune) Jeeves was once a world-renowned physicist with extraordinary intellectual abilities. He then had a stroke and suffered brain damage. The brain damage left him with average intellectual abilities (Gardner 2015, 431).[42]

Gardner maintains that, unlike some competitors, 'the existence account of harming can justify the claim that having average intellectual abilities is a harm for Jeeves, for the following counterfactual is true: (J) If it were true that both Jeeves existed and he did not have average intellectual abilities, then Jeeves would be better off in some respect. J is true because if Jeeves did not have average intellectual abilities, then he would not have had the stroke, and he would have retained his extraordinary intellectual abilities' (Gardner 2015, 438).

However, consider the following case:

> (More Loss of Fortune) An evil scientist will give Jeeves a pill that will lower his IQ. However, Jeeves has a stroke before the scientist can administer the pill. The resulting brain damage leaves Jeeves with average intellectual abilities.

It seems that the stroke harms Jeeves. But now counterfactual J does not seem to be true any longer. If Jeeves did not have average intellectual abilities, then he would not have had the stroke, but he would have been given the intelligence-reducing pill, and he would not have retained his extraordinary intellectual abilities. So, it seems that Jeeves does not suffer harm from the stroke, according to the existence account.

But perhaps, contrary to what Gardner seems to suggest, we should not think of T as 'the stroke'. Perhaps we should think of T as 'having a certain intelligence level'. This seems to fit in better with Gardner's definition in the first place. After all, the lower IQ is not a component, but a consequence of the stroke. However, if we specify T with reference to the specific condition that the agent is in (thus rendering her first clause trivially true), then Gardner's account says that an agent is harmed if and only if the second clause is true. For example, Jeeves' having an average IQ is a harm if and only if it is true that if Jeeves didn't have an average IQ, then he would have a higher IQ. However, the existence account would then imply that the following cases are instances of harm:

> (Appendix) Dave has an average IQ to begin with. He has his appendix removed using a new surgical procedure. A possible side effect of this new procedure is that it can raise someone's IQ. Not so this time: Dave's IQ remains average.

[42] Similar cases have been raised as counterexamples to non-comparative accounts (e.g. in (Hanser 2008; Boonin 2014, 72). I will discuss them further in the next section.

> (Enhancement) Steve's parents take a drug before conceiving that will prevent an embryo with less than average IQ to come into existence. Steve will therefore have either an average or a high IQ. Steve has an average IQ.

The following statement about (Appendix) seems to be true: If Dave didn't have an average IQ (after surgery), he would have a high IQ. And the following statement about (Enhancement) seems to be true: If Steve didn't have an average IQ (after conception), he would have a high IQ. So, Gardner's account seems to imply that Dave and Steve suffer harm. This is implausible.

More broadly, it seems that every respect in which someone would be better off, were their lives different in that respect, qualifies as a harm on Gardner's account. But this seems wrong. These cases seem to be more appropriately classified as withheld benefits. On Gardner's account, it seems that agents count as suffering harm whenever their welfare is any lower than it could be in any possible scenario.[43] This is implausible.

Gardner would presumably be happy to bite the bullet here. She suggests that it might not be overly implausible, after all, that 'not having wings, not being able to live to the age of 200, not having the intellectual abilities of a telepathic super genius […] may, indeed, be harms for most, if not all, of us' (Gardner 2015, 440). In the same way, she might concede that Dave and Steve, in fact, suffer harm in virtue of having merely average intelligence.

However, there are further considerations which might worry even those who are otherwise prepared to bite the bullet. Compare (Appendix) to

> (Appendix') Dave has a high IQ to begin with. He has his appendix removed using a new surgical procedure. A possible side effect of this new procedure is that it can lower someone's IQ. As it happens, Dave ends up with an average IQ.

As mentioned above, Gardner's definition of harm seems to imply that Dave suffers harm in both (Appendix) and (Appendix'). However, it also seems to have the implication that the harm is equally severe in both cases. This is because in both cases, if Dave existed and did not have an average IQ, he would have a high IQ. It seems that Gardner's definition of harm cannot distinguish between the harm that Dave suffers in these cases. But this is implausible. It seems that even if Dave suffers harm in (Appendix), the harm that he suffers in (Appendix') is much more severe.

[43] At this point, one might think that what my argument shows is that Gardner's starting point, the assumption that Jeeves suffers harm in (Loss of Fortune), is wrong. In other words, there are no purely comparative harms. I consider this "biting the bullet" strategy in the next section.

Gardner might reply that her account can at least partly account for this intuition by explaining why the reason against removing Dave's appendix is stronger in (Appendix') than in (Appendix).[44] Removing the appendix in (Appendix) does not constitute harming, because it does not cause Dave to have an average IQ, whereas removing the appendix in (Appendix') does.

However, I am not sure whether Gardner can fully avoid counterintuitive implications by appealing to the strength of reasons against harming. To see why, consider

> (Enhancement') Like Enhancement, but it is stipulated that taking the drug causes Steve to come into existence (someone else would have been conceived instead had his parents not taken the drug).

Now, compare (Enhancement') and (Appendix'). Intuitively, Dave suffers harm in (Appendix'), and the procedure harms him. It is much less intuitive that Steve suffers harm in (Enhancement'), or that his parents harm him. However, Gardner's account seems to imply that the harm in both cases is equally severe. Moreover, in both cases, there is an action that causes the harm: Steve's parents taking the drug, and Dave's doctor performing the operation. It seems to follow that the moral reason against taking the drug is as strong as the moral reason against performing the operation. But this seems very implausible, given that the operation will result in an actual loss in brain capacity for Dave, whereas taking the drug will merely ensure that Steve comes into existence (rather than another person with a lower IQ).

Given these considerations, it seems that even when one is prepared to accept the implication of Gardner's account that any negatively described state of affairs can constitute harm, this does not solve implausible implications completely.

I conclude that the existence account does not save the comparativist. It is too broad: it cannot distinguish instances of harm from instances in which agents are not being benefitted.

## 2.3 Non-Comparative Harm

Recall that the main competitor of the comparative account is the non-comparative account of harm:

> (Non-Comparative) Agent suffers harm if and only if Agent is badly off (in some respect).

[44] Thanks to Anna Folland for pressing me on this point.

On the simple version of this account, a person is harmed if and only if she is made intrinsically badly off in some respect. The problem with non-comparative harm, however, is that there are cases in which it seems intuitively clear that a person suffers harm without being in an intrinsically bad state. Consider

> (IQ) A professor with a very high IQ takes a drug that reduces their IQ by a few points.[45]
>
> (Billionaire) A billionaire is robbed of £900m.[46]

It seems that the professor and the billionaire suffer harm. However, we have stipulated in both cases that they are not badly off: the intelligent person is still very intelligent, and the billionaire is still comfortably off.

A defender of the non-comparative account might try to argue that they can explain why the agents are harmed in these cases, by appealing to the psychological discomfort that the intelligent person and the billionaire in these cases likely experience. However, this does not give the non-comparativist everything they want. We can easily imagine the professor and the billionaire very upset about the state they are in. But we can also imagine them to not be all that bothered, or even happy with the circumstances. For example, imagine that the professor is oblivious about her condition, but struggles with advanced math equations that would have been easy for her before. It seems that we should say that the professor has been harmed.[47] The problem in the professor and billionaire cases is that both agents seem to be harmed, yet neither of them is non-comparatively badly off.

However, perhaps the non-comparativist is prepared to bite the bullet and accept that neither the professor nor the billionaire is suffering harm, as long as they are not intrinsically badly off. Shiffrin seems to take this line when she says that 'comparative accounts […] identify as harm cases in which one merely loses or fails to receive a tremendous benefit. A billionaire's accidental loss of a thousand dollars will be said to be a harm to him, assuming he has a stake in his stockpile, as many billionaires do' (Shiffrin 2012, 371).

Biting the bullet comes at the huge price of revising the ordinary use of the notion of harm, and harming. Someone who steals from a billionaire clearly seems to harm the billionaire (not only to

[45] This case is similar to Hanser's 'Nobel Prize Winner' case (2008, 432) and Gardner's 'Loss of Fortune', discussed above.

[46] (Billionaire) might not be the best case to test our intuitions, since we might think that a purely monetary loss in this case does not affect the billionaire's welfare in a significant way. For a case that might be more intuitively compelling, one might think about a world-class artist or athlete whose special abilities are lowered (e.g. through an injury or illness) but remain above average.

[47] Thomson (2011, 440) makes a similar point.

wrong her). But the non-comparativist who bites the bullet must insist that this is incorrect: the thief merely takes away a benefit.

Moreover, in everyday usage we distinguish between welfare losses and mere failures to receive benefits (e.g. when the billionaire's brother refuses to get her an expensive birthday gift). Welfare losses seem morally significant in a way in which mere failures to receive benefits are not. This significance is explained by the intuitively plausible (comparative) claim that someone who loses welfare suffers harm, but someone who merely fails to receive a benefit does not suffer harm. Non-comparativists who bite the bullet, however, cannot accept this claim.

What this shows is that non-comparativists must give an account of how lost benefits compare to harms, and to mere failures to receive benefits. They must also spell out the implications of this account, for example regarding compensation owed. The billionaire who is robbed of £900m is entitled to demand justification and compensation for the financial loss she has suffered. Harm is generally thought to give rise to such entitlements against those who are responsible for the harm. If the billionaire's loss is not a harm, but merely a lost benefit, then an entitlement to compensation must be defended on non-harm-based grounds.

Perhaps non-comparativists can respond to these cases by adopting a threshold view. Before I explain how this would work, let me specify what I mean by the notion of a 'threshold'.

I begin by noting a general challenge for non-comparativists. They must answer the question how to distinguish harms and benefits in the first place. This requires specifying a welfare threshold below which somebody is harmed. It might be obvious that a broken leg constitutes a harm – but to what degree does a leg have to 'function' in order to not be in a condition that constitutes harm? What about agents who have less-than-average leg strength, poor posture, or lost athletic ability? Perhaps more obviously, it is unclear at what level bad eyesight or a low IQ constitutes harm.

At this point, we need more terminology. An agent's overall *well-being* is the sum of elements that make the agent's life go well in some respect.[48] I will call these elements 'pro tanto goods'. An agent's overall *ill-being* is the sum of those elements that make an agent's life go badly in some respect. I will call these elements 'pro tanto bads'.[49] The term well-being is sometimes used more broadly, to refer to how an agent fares overall. I will use the term *welfare* to refer to this

[48] Recall that I remain neutral with regard to the nature of well-being (for an overview of theories on well-being, see (Crisp 2017, sec. 4).

[49] Similar distinctions are made by (Kagan 2014, 262).

broad sense. One might then think of well-being as ‘positive welfare’ and ill-being as ‘negative welfare’.

What does it mean to be non-comparatively badly off in some respect? “Badly off” is ambiguous: It could mean “suffering ill-being”, or it could mean “having low well-being”. I will refer to the view that accepts only the first interpretation as the *simple* non-comparative view, and to the view that accepts both interpretations as the *threshold* non-comparative view.[50] The difference between the two views is that the threshold view allows for a wider range of states to count as instances of harm. According to the simple view, someone suffers a harm if and only if they suffer a pro tanto bad (i.e. ill-being in some respect). According to the threshold view, there are instances where someone does not suffer a pro tanto bad, but nonetheless suffers harm, because their well-being in that respect is below some threshold. More formally: for any agent A with a welfare level W in some respect, and given a non-comparative welfare threshold T, the non-comparative view says that A suffers harm in some respect if and only if $W<T$. T can either refer to a welfare level of zero (simple view) or it can refer to some positive welfare level (threshold view).[51]

According to the simple view, then, bad eyesight, or a low IQ, constitutes harm if and only if it constitutes a pro-tanto bad (i.e. instances of ill-being in some respect) according to one’s theory of welfare. This solves the challenge of distinguishing harms and benefits, as far as the theory of harm is concerned (recall that a theory of harm should be axiologically neutral: it therefore is an advantage, rather than a problem, of the simple view that its application requires one to adopt some substantial theory of welfare).

Unfortunately, however, the simple view is prone to counterexamples such as (Billionaire) or (IQ). The view implies that as long as an agent’s welfare remains above 0 in a given respect, this agent can never count as suffering harm in that respect. This is true even of agents who have lost a significant amount of welfare above the threshold. This is implausible. Biting the bullet is not an attractive strategy for non-comparativists. Defenders of the simple view are forced to bite the bullet in every instance where someone loses well-being in some respect without thereby

[50] There is a third logical possibility which I do not discuss here. “Badly off” could mean “suffering high enough ill-being”. On such a view, minor ill-being would not count as harm.

[51] The simple view and the threshold view have, as far as I am aware, not been clearly distinguished in the literature. However, I believe that the distinction matches existing accounts. On the one hand, I read Harman’s account of harm as a version of the simple view, coupled with a list account of ill-being (Harman (2004, 107) emphasizes that her view only gives a sufficient condition of harm). On the other hand, Rivera-López (2009) seems to be defending a version of a threshold non-comparative view.

becoming badly off in that respect. So, the simple view, by itself, does not get us all the way to an extensionally adequate theory of harm.

The threshold view can avoid at least some of these implausible implications. According to the threshold view, if someone loses enough well-being to have a below-threshold welfare, they suffer harm. The idea is that someone can be badly off simply in virtue of not having enough pro tanto good, even without any pro tanto bad. Unfortunately, this brings up the question of how to determine a (non-arbitrary) threshold. What makes one level of well-being, but not another, qualify as harmful? The defender of the threshold view might respond that a theory of welfare could provide some sort of rationale for a certain threshold. For example, defenders of objective list theories of welfare could argue that some items on their list of pro tanto goods are more basic than others, such that not having these items constitutes harm, without constituting ill-being.

However, even if such a view is in principle possible, it is not very attractive as an account of harm. This is because justifying harm thresholds via appeal to specific theories of welfare goes against the desideratum of axiological neutrality. Ideally, an account of harm should be usable by those with different views on what constitutes well- and ill-being.[52]

In any case, the threshold view is unlikely to avoid counterintuitive results completely. This is because it is insensitive to welfare losses above the threshold. For example, imagine that the billionaire loses half of her fortune. This is a significant loss. Intuitively, she suffers harm. However, she is still wealthy, and therefore surely above a reasonable threshold for harm. We cannot solve this by setting the threshold higher, for if it is set too high, then this means that everyone who has lower well-being than the threshold suffers harm. Not everyone who has less than half a billion on their bank account is thereby in a harmed state. I conclude that the defender of the non-comparative account of harm should not adopt the threshold view: it helps very little with regard to apparent counterexamples such as (IQ) and (Billionaire), and it opens up the new difficulty of defining non-arbitrary thresholds.

[52] Axiological neutrality might thus present problems for non-comparative accounts such as Shiffrin's (1999; 2012). She presents an account of harm according to which 'harm involves a distinctive sort of frustration or impediment of the will or of the ability to exert and effect one's will' (2012, 383). Such a view seems committed to a particular view of welfare (namely, one in which well-being and ill-being are correlated with certain will frustrations or impediments).

## 2.4 The Hybrid Account

It seems like we have reached a stalemate. Cases like (Pre-emption) and (Non-Identity) show that someone can be harmed without being comparatively worse off. Cases like (IQ) and (Billionaire) show that someone can be harmed without being non-comparatively badly off.

I think that the conclusion to draw from this is that we should adopt a hybrid account, combining comparative and non-comparative accounts. In general, a hybrid account says that an agent suffers harm if and only if an agent's welfare in some respect is either lower than a (non-comparative) threshold (for the simple non-comparative view, this threshold is 0), or lower than a (comparative) comparison state (which might be defined e.g. counterfactually, or temporally). More formally, let A be an agent with welfare level W. Then, for a non-comparative threshold T, and a comparison state C, the basic idea is this:

> (Hybrid Formula) A suffers harm if and only if either ($W<T$), or ($W<C$).

Fulfilling either component of the hybrid formula is sufficient for harm and fulfilling at least one is necessary. In the following, I will defend a version of the hybrid account which, given the previous discussion, I take to be particularly promising. My hybrid account of harm combines a temporal comparative condition with a simple non-comparative condition of harm.

We have seen in the previous discussion that the temporal comparative account of harm fails because it cannot identify ill-beings that people are born into as harm, but it can identify loss of well-being as harm. We have also seen that the non-comparative account fails because it cannot identify loss of well-being as harm when, after the loss, the person's welfare is still positive. However, the non-comparative account can identify birth conditions that contain ill-being as harms. As I will argue, the combination of these accounts is extensionally adequate:

> (Hybrid) A suffers harm if and only if either (i) A suffers ill-being, or (ii) A's well-being is lower than it was before.[53]

The temporal comparative account here is slightly altered: rather than applying to welfare, it applies to well-being only. This is because it allows the hybrid account to be presented as a unified account, rather than an ensemble of unrelated disjuncts (I say more about this in subsection 2.4.2).

[53] This version of the Hybrid Account has, to my knowledge, not been proposed in the literature.

Before we proceed, let me briefly consider a different version of (and therefore potential competitor to) the hybrid account, Meyer's 'disjunctive notion of harm':

> (Disjunctive) An action (or inaction) at time t1 harms someone only if either […] the agent thereby causes (allows) this person to be in a sub-threshold state, and, if the agent cannot avoid causing harm in this sense, does not minimize the harm; or […] the agent causes this person to be worse off at some later time t2 than the person would have been at t2 had the agent not interacted with this person at all (Meyer 2016, sec. 3.4).

The disjunctive notion only claims to provide a necessary (rather than necessary and sufficient) condition for harming. The first disjunct is a causal account of harming, combined with a threshold state notion of harm and the qualifying condition that an agent only harms if they could have done less harm. The second disjunct is a counterfactual comparative account.

I think Meyer's account faces several worries. To begin with, his account seems to classify instances of allowing harm as instances of harming, just like instances of doing harm. This is problematic for those who want an account of harm to draw a principled difference between doing harm and merely allowing harm. One might think, for example, that while I allow harm to a person by not preventing a third party from injuring this person, I do not thereby harm this person; rather, it is the third person who does so (I will talk much more about the distinction between doing and allowing harm in the following chapters). Moreover, the disjunctive notion provides an analysis of harming with a built-in analysis of harm. This embedding move not only complicates the analysis of harming, but also seems to have the implausible implication that harmed conditions are necessarily brought about by agents (given especially that Meyer presents the account as a notion of 'harm', not 'harming' or 'causing harm').

Finally, Meyer's view faces counterexamples. First, consider the case of a driver who, despite taking all precautions, runs over a child. Assume further that had he not run over the child, the driver behind him would have done so. Surely, the child is harmed, and the driver has harmed the child (though the driver may not be blameworthy for doing so). Meyer's disjunctive account does not explain this judgement. The first disjunct does not explain it, because by stipulation, the agent could not avoid causing harm and did minimize the harm (as far as possible). The second disjunct does not explain it, because it is not true that had the driver not run over the child, the child would have been better off.

Second, consider the case of Poisoner, who puts poison in Victim's drink at t0 and gives Victim a small amount of an antidote at t1. At t2, Victim (partly recovered) is worse off than she would have been had Poisoner never interacted with her. However, it seems wrong to say that giving Victim the antidote at t1 harms her, even though it causes (at least, causally contributes to) Victim's state at t2. It seems that a more appropriate assessment of this case should say that Victim suffers harm when she is poisoned at t0, and that the harm is *lessened* by receiving the antidote at t1. Meyer's disjunctive account does not explain this judgement.

The hybrid account, in contrast, does not imply that the child and Victim remain unharmed. In case of the child, the hybrid account can explain why the child is in a harmed condition after the lorry incident, since the child suffers ill-being. In the Poisoner case, the hybrid account does not imply that Poisoner harms Victim by giving the antidote (being an account of harm rather than harming, the hybrid account remains silent about this), but it does explain why Victim is in a harmed condition at t2 (because she suffers ill-being – if the poison merely reduces well-being, then the hybrid account still classifies Victim's state at t2 as a harm as long as she is worse off in t2 than she was at t0).

### 2.4.1 Is the Hybrid Account Really Extensionally Adequate?

The hybrid account is extensionally adequate if and only if it recognizes all cases of harm as such and does not find harm where there is none. I argue that the hybrid account correctly identifies all cases of harm. The hybrid account's first clause can explain why Agent is harmed in (Pre-emption) and (Non-Identity). In these cases, the victims suffer harm because they are non-comparatively badly off: they are in a state of ill-being. The hybrid account's second clause can explain why Agent is harmed in (IQ) and (Millionaire). In these cases, the victims suffer harm because they are comparatively worse off: their welfare is lower than in an appropriate comparison state. So it seems that the hybrid account can identify harms in the cases where we intuit them.

However, this might be too quick. There is one kind of case that might also seem to involve harm, but the hybrid account does not classify it as such. This is a possible false negative case. (I will discuss whether the hybrid account faces false positives at the end of this section.)

> (Lottery Ticket) Ann persuades Bob to refrain from buying a lottery ticket. If Bob had bought the ticket, he would have won.

Bob is not badly off, and Bob is not worse off than he was before. According to the hybrid account, Bob does not suffer harm. But it is not clear that this is true. After all, if Ann hadn't persuaded Bob to refrain from buying a ticket, Bob would now be a rich man. So, it might seem that Ann has harmed Bob. But Ann cannot have harmed Bob if Bob is not in a harmed state to begin with. So, either Ann has harmed Bob (i.e. the hybrid account in the version that I am defending here is incorrect) or Ann has not harmed Bob (and intuitions to the contrary are mistaken).

I think the hybrid account is correct in saying that Bob does not suffer harm. Rather, Bob does not receive a benefit. Since Bob does not suffer harm, Ann has not harmed Bob. She has prevented him from receiving a benefit.

This does not mean that the hybrid account cannot account for preventive harms:

> (Ambulance) Ann stops the ambulance that is on its way to Bob, who has suffered a heart attack.

In this case, the hybrid account implies that Bob suffers harm. He is clearly badly off (in virtue of suffering ill-being). Since Bob suffers harm, it is possible (and in this case plausible) that Ann has harmed Bob.

One might object that there are other cases similar to preventive harms that are more problematic for my account: those in which someone is about to receive a benefit which is then prevented.[54] My account implies that this person does not suffer harm. The objection might be that this implication is counterintuitive. For example, it seems intuitive that Ann harms Bob by persuading Carl to not give Bob a surprise gift.[55] I am happy to bite the bullet here, however, and argue that we should reject this intuition. It seems to me that the intuition that Ann harms Bob in this case is not very strong and is plausibly due to the implicit assumption that Ann acts out of malice or disrespect towards Bob.

To illustrate, imagine that Ann convinces Carl to give the gift to Dave, who (we might imagine) is more needy or more deserving than Bob. If we specify Ann's intentions in this way, it seems no longer intuitive that Ann harms Bob, or that Bob suffers a harm by not receiving the gift. But nothing about this modified case changes the way in which Carl's decision affects Bob. These intuitive judgements can be explained, I think, with regard to some implicit assumptions we might make in this case. I suspect that we implicitly assume that Bob has some claim to the gift when Carl decides to give it to him, and that only Carl can permissibly cancel this claim. To illustrate, imagine that the gift is already on its way to the post office. Intuitively, Carl can change his mind and rush to the post office to take the gift back to keep it to himself without thereby harming

[54] Thanks to Olle Risberg and Anna Folland for raising this worry.
[55] This case is a version of Bradley's widely discussed Batman case (Bradley 2012, 397).

Bob. However, if Ann takes the gift from the post office, she thereby takes away something that is either still Carl's or already Bob's property, and she therefore harms either of them or both.[56]

My account does not imply, however, that actions that prevent benefits are permissible, or of less moral importance than actions that constitute harming. Even if Ann does not harm Bob when she persuades Carl to not give Bob a surprise gift, she might still wrong Bob, and there might be strong reasons against her actions, since she prevents Bob from receiving a benefit.

A different objection against the hybrid account regards the harm of death. It is frequently argued that the non-comparative account of harm cannot account for the harm of death (e.g. Bradley (2012, 401)). A dead person is not in a state of being badly off. In fact, a dead person is not in any state. As Hanser (2008, 437–40) notes, this problem extends to comparative accounts. A dead person is not worse off than in any comparison state. Dead people do not have a welfare level. So, if death is a harm to those who die, it is not because it makes them badly off, or worse off. What does the hybrid account make of these cases? Insofar as death is a problem case for both comparative and non-comparative accounts of harm, the hybrid account shares this problem. However, if one of the accounts can solve the problem of death, the hybrid account will also have solved it.[57] If accounting for the harm of death is a problem for the hybrid account, this is a problem that it shares with its rivals. But it has greater chances than each of them to overcome it. I conclude that the problem of death does not pose a *special* threat to the defender of the hybrid account.

This might seem too quick. Perhaps it is the case that neither comparative nor non-comparative accounts (and, by extension, the hybrid account) can solve the problem of death. Hanser (2008) gives an argument to this conclusion. He argues that both comparative and non-comparative accounts of harm wrongly conceive of harms as states of affairs. The problems that both accounts face (including the problem of death) can be solved by conceiving of harms as events, rather than states of affairs. Hanser's event-based account of harm, very roughly, equates harms with losses of basic goods. The badness of death, on Hanser's account, can be explained as follows: 'death is a harm not because its end-state is bad, but because it consists in the loss of a cluster of powers

[56] I will say a bit more about cases of preventing benefits in intergenerational contexts towards the end of the chapter.

[57] For an example of how the problem of death might be solved within a state-based account, see McMahan's Time-Relative Interest Account of the badness of harm. This view 'relativizes the evaluation of the death to the state of the victim at the time of death. The evaluation must be based on the effect that the death has on the victim as he is at the time of death rather than on the effect it has on his life as a whole. This is what the Time-Relative Interest Account of the badness of death does' (2002, 170).

which we may, for convenience, call the 'vital' powers, these powers (typically) constituting basic goods’ (2008, 443).

I am not sure that an event-based account of harm can solve the problem of death. According to Hanser (2008, 440), state-based accounts cannot explain why death is a harm, roughly because we cannot say of someone who does not exist anymore that they are now in a bad or worse state. However, if we cannot be said to *have lost* something once we stop existing, then how can we *lose* something while we cease to exist? To use a slightly crude analogy, imagine that I have a theory about what it means for a cup to lose its handle. I drop a cup on the floor, and it breaks into tiny little pieces. The event-based theorist of handles comments, while this is happening, ‘This cup is losing its handle’. The state-based theorist of handles comments, sadly looking at the broken pieces: ‘This cup is now without a handle’. Both ways of speaking sound decisively odd. Something that is not a cup can hardly be said to be ‘a cup without a handle’. But something that loses its cup-ness and its handle-ness at the same time can hardly be said to be a ‘cup that loses its handle’, for that way of speaking seems to imply that the cup exists at least for a moment without a handle.

Similarly, a defender of the event-based account of harm might say of a person at the moment she dies “This person is losing basic goods”. But this sounds just as odd as the statement of a defender of a stated-based account, who says about a body “This person has lost basic goods”. The point is that it is difficult to see how a person can experience a ‘loss’ if she vanishes simultaneously with the goods that are lost. It therefore seems that the event-based account does not fully solve the problem of death either. I conclude that while the problem of death poses a challenge for the hybrid account, it does so to a similar or even greater degree to its competitors.

Finally, let me briefly discuss whether the hybrid account gives rise to false positives. We need to ask two questions.

First, are there losses in well-being that the temporal comparative account classifies as harm, but that do not constitute harm? Many of us lose well-being in some respect when we get older: our athletic ability might decrease, or we might need glasses. We do not ordinarily think about such changes, within certain boundaries, as harms. But the hybrid account does classify them as such. However, I think that this classification is correct. Lost athletic ability, insofar as it reduces someone’s well-being, can constitute a harm, irrespective of its cause or whether the loss was to be expected (though this might matter regarding whether actions count as *harming*).

Second, are there instances of ill-being that the simple non-comparative account wrongly classifies as harm. One might think that instances of very minor ill-being constitute such cases, such as the itch from a mosquito bite. However, those who think that a mosquito bite does not constitute harm should also think that it does not constitute ill-being. Such worries seem to point to one's underlying intuitions about theories of ill-being and are better solved on this level.

### 2.4.2 Is the Hybrid Account Unified Enough?

The hybrid account might seem dubiously ad hoc. Even if it gets the cases right, it seems to tell two different stories about the nature of harm, rather than only one story about what the unified core of harm is. In other words, an objection to the hybrid account is that it cannot explain what makes cases such as (Millionaire) and (Bad Start in Life) two instances of the same phenomenon, rather than two distinct kinds of cases.

It might be tempting for the defender of the hybrid account to argue that if only a hybrid account fulfils the desideratum of extensional adequacy, then perhaps giving up the desideratum of unity is the price to pay for this.

However, I do not think defenders of the hybrid account have to give up unity. They can give a straightforward explanation of what comparative and non-comparative harm have in common. Agents who suffer harm, whether comparative or non-comparative, have one thing in common: Their welfare is adversely affected. Someone who enters a state of being comparatively worse off than they were before suffers an adverse effect on their welfare. Someone who occupies a state of being non-comparatively badly off similarly suffers an adverse effect on their welfare.

In contrast, the welfare of someone who merely fails to receive a benefit is not adversely affected: such a person does not occupy a state of welfare that is negative either absolutely (i.e. ill-being) or relative to their previous welfare (i.e. loss of well-being).

At this point, we can see why it is useful to restrict the scope of the comparative account to well-being, and the scope of the non-comparative account to ill-being. It allows us to clearly see the difference between the two ways in which an agent's welfare can be adversely affected: by losing some degree of well-being, and by gaining (or continuing to occupy) ill-being. Agents can be prevented from gaining in well-being and can be prevented from reducing ill-being. If they fail to gain in well-being, they are still in an unharmed state. If they fail to lose ill-being, they are still in a harmed state. This is why Bob is in a harmed state in (Ambulance), but not in (Lottery Ticket).

### 2.4.3 The Hybrid Account and the Severity of Harm

The opponent of the hybrid account, however, might press the objection further. In real-world decision making, we do not only deal with cases in which one option clearly involves bringing about harm and the other option clearly doesn't. The more complicated, and more problematic, cases involve weighing options that contain different sets of harms and benefits. If an account of harm should inform such decision making, it needs to provide a way to measure the severity of harm. This is because the severity of harm influences the strength of the reason against harming. Other things being equal, it is harder to justify causing more severe harms than it is to justify causing less severe harms. It is therefore a desideratum for a theory of harm that it enables us to measure harm.

Now, the opponents of the hybrid account might say that even if the hybrid account can correctly identify instances of harm, it will inevitably encounter problems when used to establish the magnitude of harm. The following claim seems plausible:

> (Severity) The severity of harm is directly proportional to the extent to which Agent's welfare is adversely affected.

The non-comparative component of the hybrid account seems to imply that harm is more severe the more (additional) ill-being the agent suffers. In contrast, the comparative component of the hybrid account seems to imply that harm is more severe the more well-being the agent has lost. How should one compare an increase in ill-being with a decrease in well-being? The following claim seems intuitively plausible:

> (Equivalence) Agent's welfare is adversely affected to the same degree when Agent loses n units of well-being and when Agent gains n units of ill-being.

Together, (Severity) and (Equivalence) imply that the agents in the following scenarios suffer equal harm:

> (A) A's ill-being is -20. A never had less ill-being.
>
> (B) B's well-being is 1000. B's well-being used to be 1020.

This result might seem implausible. B loses only a small part of her well-being. A suffers significant ill-being. However, it would be too quick to conclude that we should reject the Hybrid Account. What the objection shows is that more work needs to be done to work out the relationship between well-being and ill-being, and how this relationship affects the severity of harm. Future research will enable us to scrutinize, revise, and perhaps replace (Severity) and (Equivalence), in order to get a better understanding of the implications of the Hybrid Account of harm.[58]

### 2.4.4 The Hybrid Account on Future Harm

The hybrid account has advantages over its competitors, specifically in cases in which future people suffer harm. To begin with, unlike the counterfactual comparative account, the hybrid account of harm can explain why people can be harmed in non-identity cases. Future people can be harmed by present actions, even when they owe their very existence to these actions. In standard non-identity cases, future people suffer ill-being of some kind. For example, a policy can lead to a future catastrophe, and a drug can lead to a child being born with a significant health condition. In these cases, the simple non-comparative component of the hybrid account explains why those affected suffer harm. Moreover, unlike the non-comparative account, the hybrid account of harm can explain why future people can be harmed when they lose well-being. In these cases, the temporal comparative component of the hybrid account explains why future people can be harmed if their standard of living drops during their lifetime, even when they are not badly off.

According to the hybrid account, future ill-being and future losses in well-being both constitute harm and can therefore in principle ground harm-based reasons against actions that bring about these harms. Since counterfactual comparative accounts and non-comparative accounts can each only account for one of these types of harm, the hybrid account of harm has the potential to ground stronger harm-based reasons against harming future people than either of these accounts. This might help to explain widespread views that we have strong duties to avoid harming future people, e.g. by causing global warming and its predicted significant impacts on humanity.

However, the hybrid account also implies that losing potential future improvements does not constitute harm to the future people affected.[59] Unlike people who exist at present, future people

[58] The relationship between these two types in which welfare can be adversely affected might help account for something which, Hanser claims, the standard temporal comparative account fails to explain: why we 'have a strong reason to come to the aid of people in danger of suffering injury or losing their property, but we have at best a weak reason to come to the aid of those simply in danger of undergoing declines in momentary well-being' (Hanser 2019, 865). The defender of the Hybrid View could say here that suffering injury is an instance of ill-being, whereas such a simple decline in momentary well-being is not.

[59] Such an implication is plausible, even outside of harm-based reasoning. Mulgan (2006) develops a moderate consequentialist account of moral and political obligations to future generations. He argues that

do not have a clear default level of well-being. Present people always are at some level of well-being, and it is usually clear when they lose well-being, compared to a previous state. In contrast, someone does not “lose” well-being by being born in a state of less well-being than they otherwise might have been born into. So future people are not harmed if they are born in a state that contains less well-being than it could have contained – or indeed, contains no well-being at all. [60]

This implication has potentially far-reaching consequences for reasons not to harm future people. It shows a difference between the moral status of actions that lower the well-being of presently existing people, and the moral status of actions that lower future people’s ‘starting conditions’. This is because only the outcome of the action affecting present people constitutes harm, whereas the outcome of the action affecting future people do not.

Harm-based reasons arguably do not exhaust moral reasons that govern long-term decision making. If what I have said above is correct, then constraints on doing harm cannot justify obligations to ensure that future people are born into a state of positive welfare. However, common standards of intergenerational justice are demanding. (For example, egalitarian standards demand that future people are at least as well off as we are ourselves (Gosseries and Meyer 2009, chaps 4–5)). If we have moral obligations to fulfil such demanding standards in intergenerational justice, then my argument so far suggests that these obligations are grounded in duties of beneficence or justice, not in the duty to avoid harming others.[61]

## 2.5 Conclusion

In this chapter, I have investigated the notion of harm. I argued that the best account of harm combines both temporal comparative and non-comparative elements, and then proposed that we should adopt a hybrid account of harm. I have argued that a version of the hybrid account that

present people privilege their own lexical level when evaluating future scenarios, such that they would feel required to prevent a decline in future well-being, but not to raise the well-being of future generations (Mulgan 2006, 228). This might motivate agents to be pessimistic and risk-averse when evaluating and choosing future scenarios (Mulgan 2006, 244).

[60] Hanser (1990) develops a solution to the non-identity problem which has similar implications: policies that lower future well-being cannot be said to harm future people. Hanser remarks about these cases: ‘Being made "worse off" in this way lacks the moral significance of true harming […] although there may be something to be said against choosing a policy that sacrifices future welfare for immediate gain, the objection cannot, I think, be that the choice harms future people, or puts them into a state that it is bad for them to be’ (1990, 66).

[61] See Gardner (2016) for a characterization of harm-based and impersonal welfare solutions to the non-identity problem, and an overview of their respective difficulties.

includes a temporal comparative well-being baseline and a non-comparative ill-being baseline is extensionally adequate and provides the resources to be presented as a unified account of harm. I have noted that the hybrid account of harm can acknowledge that there are other kinds of moral reasons that are broader in scope than the reasons deriving from either the counterfactual comparative account or the non-comparative account. Further research is needed to spell out the implications of the hybrid account for measuring the severity of harm.

# Chapter 3 Two Dimensions of Harming

## 3.1 Introduction

In this chapter, I discuss the question how the relation between a behaviour and a harm can give rise to moral reasons against the behaviour. I call these reasons *harm-based reasons*. I will argue that there are two kinds of behaviour-harm relations that give rise to harm-based reasons against the behaviour.[62]

Many people believe that we have strong obligations not to do harm to others. In particular, these obligations are typically thought to be stronger than obligations to prevent harm. This view is central to both common sense morality and deontological moral theories:

> (Doing Thesis) The reason against doing harm is stronger than the reason against merely allowing, i.e. failing to prevent, harm.[63]

Many people also believe that we have strong obligations to prevent harm if we can do so at no great cost to ourselves. When our behaviour can make a difference to people's lives, we have a reason to not make them worse off:[64]

> (Difference Thesis) The reason against a behaviour is stronger if there is someone who will be worse off if the behaviour is performed.[65]

Both principles are very plausible. However, it is not clear how they relate to each other. In particular, it can be puzzling to work out which of these principles is stronger. To illustrate, as Derek Parfit (1984; 2010) has famously argued, large-scale policies change people's behaviour, and thus indirectly change who will be born in the future. Therefore, future victims of today's unsustainable policies would not have existed had we acted differently. These policies, then, do not make anyone worse off than they would have been. The same is not true for failing to aid present people. These people already exist – their identity is not dependent on our behaviour.

[62] I prefer the term 'harm-based reasons' over 'reason against harming' (which is used e.g. by Gardner (2017)), because the term can be used more broadly to encompass reasons against behaviours that relate to harm in different ways, not all of which might qualify as harm doing or harming.

[63] As I have explained in the introduction, I think that the Doctrine of Doing and Allowing (DDA) implies the Doing Thesis. However, the Doing Thesis can also be accepted by those who do not accept the DDA (for example, because they reject the prerogative element of DDA).

[64] Similarly, when our behaviour does not make anyone's life worse, it seems plausible that we have less reason to refrain from the behaviour.

[65] Fiona Woollard (2012a) argues for a version of the Difference Thesis and points out its inherent tension with harm-based solutions to the non-identity problem.

Under these assumptions, the Difference Thesis implies that the reason against failing to aid present people is *stronger* than the reason against causing bad future outcomes. However, failing to aid present people constitutes merely allowing harm, whereas causing bad future outcomes through the implementation of unsustainable policies seems to constitute doing harm. If the harm in both cases is comparable, the Doing Thesis implies that the reason against failing to aid present people is *weaker* than the reason against causing bad future outcomes.

At the start of the previous chapter, I have distinguished between two important questions in the philosophy of harm. First, when does some state of affairs count as a (pro tanto) harm to someone? I have discussed this question in the last chapter. I have rejected the counterfactual comparative account of harm, and instead argued for a novel hybrid account of harm. Second, when does a behaviour count as harming?[66] This is the question that I will address in this chapter. I will answer it by developing a novel two-dimensional account of harming.

It is important to see that the hybrid account of harm and the two-dimensional account of harming address different questions, and therefore, my defence of the hybrid account does not commit me to any specific view on harming. In the last chapter, I pointed out that whether an account is a good (bad) account of harm does not necessarily have any implications regarding whether it is a good (bad) account of *harming*. I suggested that the standard counterfactual comparative account might be more plausible as an account of harming than as an account of harm. In this chapter, I will argue, however, that we should reject the standard counterfactual comparative account of harming.

In the following, I develop a novel account of harming. On the two-dimensional account, the Doing Thesis and the Difference Thesis express two dimensions of harming which are both necessary to identify harm-based reasons. The two-dimensional view combines the intuitive appeal of the main account of harming in the literature, the counterfactual comparative account, and its main rival, the causal account. I first introduce the two-dimensional view and then defend its two components separately. Finally, I discuss whether the reason against counterfactual comparative harming is stronger than the reason against causal harming or vice versa. I reject recent arguments to the conclusion that the reason against harming differs in strength across dimension and conclude that there is prima facie reason to assume that these reasons are equally strong.

[66] I focus on pro tanto harming in this chapter, for reasons given in section 3.4.

## 3.2 The Two-Dimensional View

Call the relation that obtains between a harming behaviour and a harm (i.e. a harmful outcome) the harming relation. In the literature, we see two main views about how this relation should be specified.

> (Counterfactual Comparative) A's behaviour harms B if and only if B is subject to harm and B would not have been subject to harm had A not performed the behaviour.[67]
>
> (Causal) A’s behaviour harms B only if A’s behaviour causes B to be subject to harm.[68]

I suggest that both views appeal to different dimensions of what makes harming a morally relevant relation.[69] On the one hand, harming seems to be about making a difference. More precisely, it is about making someone worse off – making it the case that someone is subject to harm when they could have lived without it. Being worse off than they could have been is bad for, and against the interests of, the person affected. The counterfactual comparative account of harming is plausible because it appeals to this *difference-making dimension of harming*.

On the other hand, harming seems to be about *doing* harm. More precisely, it is about making a causal contribution to a bad outcome, which constitutes an intrusion into the proper sphere[70] of the person affected. Such an intrusion is bad for, and against the interests of, the person affected. The causal account of harming is plausible because it appeals to this *contribution dimension of harming*.

Let us assume that the reason against a behaviour that constitutes harming is stronger than the reason against a behaviour that does not constitute harming, everything else being equal. Since

[67] The counterfactual comparative account of harm as defended in the literature (e.g. by Klocksiem (2012, 285), Boonin (2014, 53), Norcross (2005, 150) and Feinberg (1984, 34)) combines an account of harm with an account of harming, saying that A harms B if and only if B is worse off than B would have been, had A acted differently. I use ‘subject to harm’ here, which admittedly makes the formulation of the account somewhat idiosyncratic, to remain neutral with regard to what constitutes harm (whether it is being worse off than one would have been, or worse off than one was previously, or simply badly off, for example).

[68] Proponents include Harman (2004; 2009), Shiffrin (1999), and Gardner (2015).

[69] The idea of distinguishing causing from difference making is not new, even though its implications for the notion of ‘harming’ have not been comprehensively explored. Cullity (2019) has recently pointed out that both causing and difference making are morally relevant, though his discussion of climate harms largely focusses on the difference-making dimension. The distinction features more prominently in the recent literature on the relation between causation and moral responsibility. Kaiserman (2018), following Bernstein (2017), distinguishes the ‘production’ and ‘dependence’ dimension of causation and points out that they can conflict. Bronner (2018) similarly distinguishes ‘causing as difference making’ and ‘causing as doing’ and argues that the killing/letting die distinction rests on the ‘doing’ understanding of causation; however, criticisms of the distinction have predominantly focused on the ‘difference making’ understanding.

[70] This understanding of ‘doing’ harm has been advanced by Woollard (2015; 2012c). She builds on this characterisation to defend the principle that doing harm is harder to justify than merely allowing harm.

harming is a morally relevant phenomenon, this assumption seems plausible. However, on this assumption, the two accounts of harming have different implications. The assumption coupled with the counterfactual comparative account implies the Difference Thesis, but not the Doing Thesis; the assumption coupled with causal account supports the Doing Thesis, but not the Difference Thesis.

It is clear that the counterfactual comparative account implies the Difference Thesis. It does not imply the Doing Thesis, because both doing harm and allowing harm can make a difference to those affected. Imagine one case in which Anne eats Mary's cookies (doing) and another case in which Anne fails to stop Joe from eating Mary's cookies (allowing). Both behaviours make a difference to the content of Mary's cookie jar.

The causal account supports the Doing Thesis. To see why, it is important to distinguish two interpretations of the causal account. On the first interpretation, doing harm just *is* causing harm. On the second interpretation, the distinction between doing and allowing harm does not neatly line up with the distinction between causing and not causing harm to occur. For example, there might be cases of causing harm that qualify as allowing, but not as doing, harm. The first interpretation clearly implies the Doing Thesis. The second interpretation implies the Doing Thesis under the additional assumption that not only whether, but also how, the behaviour is causally related to the outcome (e.g. in a doing way or in an allowing way) matters morally.

The causal account does not imply the Difference Thesis, because whether a behaviour makes a causal contribution to an outcome is independent from whether it makes a difference to the outcome. For example, it might be true that if Anne had not eaten Mary's cookies, then Joe would have eaten them. While it is Anne, and not Joe, who has caused Mary's cookie jar to be empty, it is not true that the jar would still be full had it not been for Anne's behaviour.

The difference-making dimension and the contribution dimension do not in principle conflict. A behaviour can make a difference to the outcome, and also causally contribute to the outcome in a 'doing' way. In fact, most standard examples of harming seem to tick both boxes. (Consider, for example: "Anne has harmed Joe, by punching him and thereby breaking his nose", or "Joe has harmed Anne by breaking her favourite toy car in two pieces.")

It is therefore both possible and plausible to combine both dimensions in one account:

> (Two-dimensional) A has a harm-based reason against a behaviour if and only if the behaviour constitutes (i) counterfactual comparative harming or (ii) causal harming or both.

My aim in this section is to show that we should accept something like the two-dimensional view. However, none of my arguments rely on this specific version of the view. Any views, including revised versions of the counterfactual comparative or causal view, qualify as a two-dimensional view for my purposes, as long as they acknowledge the moral relevance of both dimensions.

Table 1

| | Doing | Not doing |
|---|---|---|
| Making a difference | 1 | 2 |
| Not making a difference | 4 | 3 |

According to the two-dimensional view, we can classify behaviours as illustrated in table 1. A standard case of harming qualifies as harming on both dimensions. In these cases, someone is doing harm, making the affected person worse off than they would otherwise have been (case 1). Then there is behaviour that does not qualify as harming on either dimension (case 3). This includes cases in which the behaviour has no causal effect or counterfactual relation to the harm, such as Mary's painting the cookie jar blue.[71]

As one might suspect, the difficult cases are those in which the two dimensions come apart. There are two types of such cases. First, cases in which an agent does harm without making a difference to anyone (case 4), e.g. pre-emption and non-identity cases. Second, cases in which an agent does not do harm, but nonetheless makes a difference (case 2), e.g. cases of merely allowing harm.

Not all behaviours that give rise to a harm-based reason according to the two-dimensional view are behaviours that we would ordinarily call 'harming'. This fact is why I use the slightly cumbersome notion of 'behaviour that is subject to harm-based reasons against it'. It seems generally appropriate to talk about a behaviour as 'harming' if it qualifies as both counterfactual comparative and causal harming (e.g. Anne's punching of Joe), and it seems generally inappropriate to talk about a behaviour as 'harming' that qualifies as neither counterfactual comparative nor causal harming (e.g. Mary's painting the cookie jar blue).

[71] It might also include behaviour that has very remote relations to the harm, relations that are too weak to qualify as harming in either a counterfactual comparative or causal sense.

However, for behaviour that qualifies as harming on only one dimension, it seems sometimes appropriate and sometimes inappropriate to refer to it as 'harming'.[72] Let me therefore clarify my terminology: In the following, unless stated otherwise, I will reserve the unqualified term 'harming' for behaviour that qualifies as harming on both dimensions (type 1 cases, as illustrated by table 1). I will use the terms 'counterfactual comparative harming' or 'difference-making' to refer to (1 and 2), and 'causal harming' or 'doing' to refer to (1 and 4).

According to the Doing Thesis, the reason against the agent's behaviour in type 4 cases is stronger than the reason against the agent's behaviour in type 2 cases. This is not surprising, since the Doing Thesis is supported by causal considerations. The causal account of harming does not distinguish between cases 1 and 4, which both qualify as doing harm, and it does not distinguish between cases 2 and 3, neither of which qualifies as doing harm.

According to the Difference Thesis, the reason against the agent's behaviour is stronger in type 2 cases than in type 4 cases. As I have argued, the Difference Thesis is supported by counterfactual comparative considerations. The counterfactual comparative account of harming does not distinguish between cases 1 and 2, and it does not distinguish between cases 3 and 4. In the following, I argue that a plausible account of harming should account for the moral relevance of both dimensions of harming.

## 3.3 Why We Need the Causal Dimension

Recall that according to the comparative counterfactual view a behaviour qualifies as harming if and only if the harmful outcome counterfactually depends on the behaviour. This view does not classify cases as harming that, intuitively, are clear cases of harming. Two types of these cases

[72] I think that whether we would find it appropriate to talk about 'harming' in these cases depends largely on contextual factors, such as pre-existing rights and duties, the nature of the harm, the associated cost, and expectations that might differ with cultural and social settings. I suspect that people's intuitions will differ about these cases, which makes it hard to find good examples. However, I think that many people would be inclined to count the following behaviours as instances of Anne harming Joe: (1) Joe's mum Anne does not make sure that Joe's seatbelt is fastened before starting the car. (2) Anne eats all of Joe's cookies (if she had not finished them, Mary would have done so). I also think that many people would hesitate to count the following behaviours as instances of Anne harming Joe: (3) Anne watches Mary eat Joe's cookies. (She could have easily prevented Mary from eating them.) (4) If Anne had not chosen to implant an embryo with a gene that causes a minor impairment, Joe would never have been born.

have been extensively discussed in the literature.[73] For an example of a so-called pre-emption case, consider

> (Shooting Match) Victor has made two terrible enemies, Adam and Barney. Barney is just about to shoot and kill Victor. Barney is protected by a bullet-proof, sound-proof shield so that Adam can neither stop him forcibly nor dissuade him. Adam knows this, but Victor's death by another's hand will not satisfy his thirst for vengeance. Adam shoots Victor (Woollard 2012, 684).

It seems intuitively clear that Adam (impermissibly) harms Victor. However, Adam does not make Victor worse off than he would have been had Adam refrained from shooting, since in this case Barney would have shot Victor. For another example, consider a non-identity case:[74]

> (Risky Policy) Suppose that, as a community, we have a choice between two energy policies. Both would be completely safe for at least two centuries, but one would have certain risks for the further future. If we choose the Risky Policy, the standard of living would be somewhat higher over the next two centuries. We do choose this policy. As a result, there is a similar catastrophe two centuries later, which kills and injures thousands of people. […] But if we had chosen the alternative Safe Policy, these particular people would never have existed. Different people would have existed in their place (Parfit 2010, 112–13).

It seems intuitively clear that the Risky Policy (impermissibly) harms future people. However, since the identity of the future people depends on which policy we choose, we do not actually make anyone worse off than they would have been, had we chosen differently. The counterfactual comparative account cannot capture the difference between behaviour that does not cause harm, and behaviour that does, as in (Shooting Match) or (Risky Policy). In contrast, the causal account explains the intuitive judgements that the agents' behaviour in these cases constitutes harming, and that they have a moral reason to refrain from their behaviour. Defenders of the causal account see this as a key advantage of the causal account (e.g. Woollard (2012a) and Gardner (2017)).

[73] Carlson (2019) raises even more problems for the comparative counterfactual view, arguing that the view is incompatible with the prudential and moral relevance of harm and benefit. See Klocksiem (2019) for a response and Carlson (2020) for a response to the response. Moreover, Carlson, Johansson and Risberg (forthcoming) present further problems for all views that define harms by comparing the actual consequences of events with possible hypothetical consequences of that event.

[74] In non-identity cases, the same behaviour that causes the harm also determines the identity of the victim of harm. The term 'non-identity problem', which refers to the difficulty of spelling out why such behaviour is wrong, was coined by Derek Parfit (1984).

I argue that there is more to say in support of the causal account. The defender of the counterfactual comparative account needs to appeal to the causal account in order to individuate (harmful) choices. To see why, consider first that we could turn any policy affecting future people into a non-identity case:

> (Shuffle Policy) Any policy that affects future people will automatically be supplemented with a policy that ensures that those who will be affected by this policy would not have existed otherwise (without doing any further harm). For example, it might include changing the dates of national holidays, changing timings of popular TV broadcastings, changing college admissions and civil service application procedures, introducing new taxes or lowering existing ones, and so on. The assumption is that these measures will change people's behaviours in ways that will eventually influence when and with whom they procreate.

By adopting (Shuffle Policy), and following the counterfactual comparative approach, policy makers would not have any reasons having to do with harm to future people against their decisions. This is highly implausible. If environmental pollution harms future people, then surely environmental pollution, followed by Shuffle Policy, still harms future people.

This seems even more obvious in cases of *allowing* harm to future generations. Imagine that we could prevent a future catastrophe but fail to do so. We clearly seem to allow harm. However, imagine that rather than preventing the future catastrophe, we adopt the less costly Shuffle Policy, claiming that we have prevented the harm in this way. But I argue that this cannot be right. We do not prevent harm merely by changing the identity of those who will be affected.[75]

My argument works even without the need to introduce an explicit policy. Suppose that all large-scale policies (and perhaps smaller-scale policies in conjunction with other policies) change the identities of future people. Governments implement policies (and decide and change them) all the time. Assume that the government of a country adopts the Risky Policy. The policies that this

[75] This implication of the counterfactual account of harming in cases of harm preventions in non-identity cases has to my knowledge not been discussed. However, the topic deserves further attention. It constitutes not only a problem for defenders of counterfactual comparative accounts. Using a similar line of argument, one might suggest that it broadens the scope of behaviours that are subject to the non-identity problem. The non-identity problem not just affects cases of doing harm, but also cases of preventing harm or providing benefits for future generations. (Regarding benefitting, one might ask, why are behaviours that benefit future people and influence their identity at the same time are morally good, if they do not make these people better off than they otherwise would have been? If bringing about impersonal good is enough reason to motivate long-term beneficence, then why is there so much resistance to the idea that the same should hold for harm? This question is different from the question whether we can benefit people by bringing them into existence (for an affirmative answer to this second question, see e.g. Harman (2004, 108); for the view that existence harms people more than it benefits them, see Benatar (Benatar 2008; for criticism, see e.g. Weinberg 2012). )

government implements in the next few years, taken together, presumably change the identity of the victims of the Risky Policy (and the government's policies are even more likely to have these effects when seen in conjunction with other agents' actions, e.g. foreign governments or supranational entities). If that is true, then *any* policy that affects future people is a non-identity case.

To avoid this implication, the defender of the counterfactual comparative view can argue that instances of counterfactual comparative harming should be individuated not only with regard to the outcome they bring about, but also with regard to how they do so. Outcomes should include the causal pathway leading to the outcome. The defender of the counterfactual comparative view could then claim that in (Risky Policy), the causal pathway that leads to the existence of the future people is different from the causal pathway that leads to the future catastrophe, and that these therefore constitute two different choices (this applies in the same way to the case of allowing harm).

However, this is precisely the idea underlying the causal dimension: It matters *how* the agent brings about an outcome. This is what enables the causal account to distinguish doing from merely allowing harm in the first place. The agent's role in the causal chain leading to harm matters morally. So, the amended counterfactual comparative view draws attention to precisely those features that matter according to the causal account. It acknowledges the moral relevance of both counterfactual comparative and causal considerations, and therefore is a two-dimensional view in my sense.

## 3.4 Why We Need the Counterfactual Comparative Dimension

These considerations shed light on the wider question of whether the doing/allowing distinction applies to behaviour that influences the identity of future people. In non-identity cases, it is assumed that the doing behaviour makes it the case that different people will be born, and thus be affected. However, it is not clear whether this 'makes it the case' should be understood as a causal doing, or as a counterfactual dependence relation. I think that we should not understand it as a causal doing relation.

This is because otherwise we could not make sense of non-identity cases in which an allowing behaviour makes it the case that different people will be born, and thus be affected. For example, it might be true that if I press a button, the Risky Policy will not be implemented; however, I fail to press the button. While it might be true that different people would have been born had I pressed the button, it is not true that I have done anything that caused a certain set of people to exist (after all, I have merely failed to act). If we need to appeal to counterfactual comparative considerations to make sense of such cases, this gives us a reason to abandon a purely causal account.

The causal view faces even more problems. This becomes clear once we consider cases of merely allowing harm. I suggest that agents only allow harm when they fail to prevent some harm that they could have prevented. Allowing harm, therefore, makes those suffering the harm worse off than they would have been, had the Agent prevented the harm. If that is true, then counterfactual considerations are necessary to correctly classify cases of allowing harm. Consider

> (Allowing Shooting) Adam is about to shoot and kill Victor. Claire can prevent Adam from shooting but does not do so. Adam shoots Victor and Victor dies from the bullet wound.

What does the causal view say about (Allowing Shooting)? Again, this depends on the position of the proponent of the causal view on the question whether omissions are causes. If the defender of the causal view thinks that omissions can be causes, then the causal view implies that Claire has a harm-based reason to prevent the shooting.[76] Defenders of the causal view then face the further problem of specifying how to distinguish causes that are doings from causes that are mere allowings.[77]

If defenders of the causal view think that mere omissions cannot be causes, then it is more complicated. Since her failure to prevent the shooting is merely an omission, the causal view would then imply that Claire does not have a harm-based reason against her behaviour. But this leaves the defender of the causal view with a problem. They cannot explain what distinguishes Claire's behaviour from a behaviour that is totally unrelated to harm. This is problematic. Claire is related to Victor's death in a morally relevant way. Moreover, this relation seems to be at least

[76] It seems intuitive to identify doing harm with causing harm (see e.g. Callahan (2012, 118–19): In (Push), Agent causes Victim's death, whereas in (Non-Intervention), Agent does not cause Victor's death. However, on common notions of causality (perhaps most prominently, David Lewis' (1974) counterfactual account of causation), many instances of allowing harm count as causing harm. For an argument that cases of allowing harm involve causing harm, see Bennett (1995, 128–30), Kagan (1989, 92–94), Quinn (1989a, 293–94). The question whether omissions are causes is also controversial in philosophy of science. Process theories deny that omissions are causes. Phil Dowe (2004), for example, calls causation by omission 'quasi-causation' and maintains that it is importantly distinct from standard causation. However, see Jonathan Schaffer (2004) for an argument that causation by omission is genuine causation.

[77] One straightforward option might be viewing causing by action as doing harm and causing by omission as allowing harm. However, counterexamples from the literature suggest that it might not be this easy. For example, an actor can spoil a performance by failing to turn up (Foot 1967, 7), and a benefactor might let a child die by cancelling a direct debit to charity (Bennett 1995, 91).

partly explicable in causal terms. Claire could have prevented Adam from shooting, and Adam's shooting is the cause of Victor's death.

Fortunately, there are ways in which the defender of the causal account can spell out the way in which Claire's behaviour is relevant to the causal sequence leading to harm. For example, borrowing a move from Woollard (2015, chap. 3), they might say that a negative fact about Claire's behaviour – the fact that she did not prevent the shooting – forms part of this causal sequence, whereas totally unrelated actions or events do not form part of the sequence. However, consider

> (Pre-emptive Allowing Shooting) Both Adam and Barney are about to shoot Victor. Either shot would be sufficient to kill Victor. Claire can now prevent Adam, but not Barney, from shooting. Claire does not prevent Adam from shooting. Adam shoots Victor and Victor dies from the bullet wound.

Recall that the causal account classifies Adam's behaviour in (Shooting Match) as doing harm, irrespective of whether he makes a difference to the outcome. Similarly, the causal account classifies Claire's behaviour in (Pre-emptive Allowing Shooting) as allowing harm, irrespective of whether she can make a difference to the outcome. This follows regardless of whether omissions are causes. If they are, then Claire's failure to prevent Adam's shooting causes Victor's death in both cases. If omissions are merely causally relevant to harm, then Claire's failure to prevent Adam's shooting is causally relevant to Victor's death in both cases.[78]

This implication of the causal account is implausible. Consider what would have happened had Claire acted differently: Claire would have prevented Adam from shooting Victor. However, she would not have prevented Victor from being shot. Claire would have prevented the outcome 'being shot by Adam'. However, she would not have prevented the outcome 'being shot'. From Victor's perspective, surely, Claire has not prevented the harm at all. The harm consists in being killed, and Claire could not have prevented that. In fact, there never was a possibility for Claire to prevent what constitutes the harm.

[78] One might respond that omissions sometimes are causes (e.g. in Allowing Shooting) and sometimes not (e.g. in Pre-emptive Allowing Shooting). A causal version of this asymmetry between actions and omissions has been put forward by Carolina Sartorio (2005). If Sartorio is right, however, then this supports my conclusion that the causal view needs to appeal to a counterfactual element to give plausible results, since at the heart of the distinction between omissions that are causes and omissions that are not causes is a counterfactual test. Sartorio defends what she calls the 'New Asymmetry (Causal)': 'An action can cause an outcome even if the outcome would still have occurred in the absence of the action. By contrast, an omission cannot cause an outcome if the outcome would still have occurred in the absence of the omission' (2005, 470).

The two-dimensional view avoids the implausible implication: since Claire's action is neither making a difference to the harm, nor causing harm in a 'doing' way, the two-dimensional view does not commit us to saying that Claire has a harm-based reason to prevent harm.[79]

Moreover, we can construct cases in which Claire similarly allows harm, in the sense of influencing the causal process that will result in harm, without thereby changing anything as crucial as who the perpetrator of harm is. Imagine that Victor is about to die from poisonous gas in the air. Claire can now put him in a vacuum, where Victor would suffocate. Claire does not put Victor in a vacuum. Has she allowed harm to Victor? Perhaps even more clearly: Victor is about to die from poisonous gas in the air. Claire can shield Victor from one side, such that the gas atoms that cause Victor's death are different from the gas atoms that would otherwise cause Victor's death.[80] It seems clear that if Claire does not shield Victor, thus ensuring that different gas atoms cause Victor's death, she has not thereby harmed Victor.

Return to (Pre-emptive Allowing Shooting). Imagine that if Claire calls out Adam's name, Adam will be distracted for a split second, and Barney's bullet will hit Victor before Adam's bullet does. If Claire does not call out Adam's name, Adam will shoot, and his bullet will hit Victor before Barney's bullet does. The causal account suggests that if Claire stays silent, she thereby allows harm to Victor. But this seems wrong: Claire might allow Adam to do harm, but she does not allow harm.[81] Her behaviour is relevant to the causal chain related to the harm doing (since she fails to prevent this instance of harm doing), but only indirectly related to the harm, in a way that does not qualify as failing to prevent harm.

This case points to a further complication. When Claire calls out Adam's name, she thereby allows Barney's (rather than Adam's) bullet to hit and kill Victor, from which it would follow that Claire allows harm to Victor, whether she stays silent or not. This, too, seems wrong.

[79] To clarify, my point is not that Claire is not blameworthy for allowing harm. Rather, I claim that it is implausible to categorize her behaviour as allowing harm in the first place, given that she could not have done anything to prevent the harm.

[80] Admittedly, it is difficult to form intuitions about these cases, since it is tempting to think that Adam's and Claire's action at least potentially influence the likelihood of harm ('if Adam does not shoot, perhaps it becomes a little less likely that Victor dies'). It is important to remind ourselves of the initial stipulation of certainty.

[81] Perhaps the intervening agents are muddling intuitions here. Consider this analogous case:
(Automatic Weapon) Several loaded guns are pointing towards Victor. Every single shot will be fatal for Victor. The automatic system always keeps at least two of the guns activated and will fire all active guns at time t1. Currently, guns 1 and 2 are active. Colin can access the system at t0; however, all he can do is to deactivate gun 1 (in which case the system will activate gun 3) or do nothing. In both cases, the system will fire two bullets at t1, and Victor will die. Colin does nothing.

The point, of course, is that the mere capacity to change the causal pathway in which someone suffers a harm does not seem sufficient to constitute allowing harm. 'Changing the causal pathway' means to influence not the harm, but merely which of the existing threats of harm will materialize. The mere ability to change the causal pathway does not constitute allowing harm. I have argued that this seems intuitively plausible (as I have suggested in the poisonous gas cases), and moreover, that it helps us avoid the counterintuitive implication that people can be in a situation in which everything they do would constitute allowing the same harm, without them having a possibility to avert the threat. The mere capacity to change the causal pathway, then, is not a good candidate for pre-emptively allowing harm. An agent can only allow harm if there was a real opportunity for the agent to prevent harm.[82]

The upshot is that the defender of the causal account lacks the resources to distinguish (Pre-emptive Allowing Shooting) from cases such as (Allowing Shooting).[83] Claire's behaviour is relevant to the causal chain leading to Victor's death in both cases. Moreover, her behaviour constitutes a mere allowing in both cases (it is not a doing behaviour).

However, if I am right that agents only allow harm if the harm would not have occurred had they acted differently, then agents only have a reason against allowing harm when they had the capacity to make a difference to the outcome. But this capacity to make a difference is precisely what the counterfactual comparative account of harming measures. Agents can only have a reason against allowing harm if their behaviour constitutes counterfactual comparative harming.

If defenders of the causal account accept that counterfactual comparative considerations strengthen harm-based moral reasons against allowing harm, then they should also accept that counterfactual comparative considerations can strengthen harm-based moral reasons against doing harm.[84] But this is just to say that whether a behaviour makes a difference to harmful outcomes influences the strength of harm-based moral reasons: in other words, it is acknowledging the moral relevance of both dimensions of harming.

One might object that there are counterexamples to my claim that allowing harm involves a counterfactual comparative element. Some harms have good *overall* long-term consequences for

[82] This is a necessary condition for allowing harm. It does not seem to be sufficient. Whether a behaviour counts as allowing harm might depend on other factors that need to be met in addition, such as whether the harm is reasonably foreseeable.

[83] As far as I know, this objection has not been discussed by defenders of the causal account. If I am right that the causal account only escapes the objection by appeal to counterfactual comparative considerations, then the objection seems fatal to the (one-dimensional) causal account.

[84] Alternatively, they could provide an explanation for why the relevance of counterfactual comparative considerations is restricted to allowing harm. However, I cannot see what a rationale for such a view would be. In any case, such a view would have to reject the intuitively plausible Difference Thesis.

those affected, and by preventing such harms, one would also prevent an outcome that would have been better for the victim overall. In an example given by Woodward (1986, 810–11), an airline refuses to sell a flight ticket to Smith, who is black, because of racial discrimination. The plane later crashes. Smith suffers pro tanto harm, but not overall harm. Now, imagine that a member of the flight crew was present and could have intervened, thereby enabling Smith to board the plane, but didn't do so. The crew member allows pro tanto harm, but not overall harm. Her behaviour makes Smith worse off in a respect, but not all things considered.

I agree that in such a case, the crew member might not have allowed harm to Smith in an overall sense. But surely, she has allowed harm to him in a pro tanto sense. Since I am interested in pro tanto harms, I do not take this case to be a counterexample to my claim.

The reasons I focus on pro tanto harms here are the following: First, whether a behaviour harms someone in an overall sense arguably depends on the balance of the pro tanto good and bad effects that this behaviour has on someone. If this is true, then an account of pro tanto harming is more fundamental than, and can ultimately generate, an account of overall harming. Second, it seems to me that an account of overall harming is significantly more difficult to apply to many real-world cases, in particular in cases in which agents merely allow harm. This is because it requires agents to state what would have happened had they intervened – not merely whether they would have been able to prevent some pro tanto harm, but whether their intervention would have been overall beneficial for the person affected. This will be difficult to predict or evaluate in hindsight, and might often be indeterminate, and therefore, impossible to evaluate and compare. Finally, even in cases in which agents' behaviours do not constitute overall harming, intuitively, it still matters morally whether they constitute pro tanto harming. Pro tanto harmful effects are morally important, even if bringing them about is justified by pro tanto beneficial effects. All this is not to say that the notion of overall harming has no meaning or use in moral deliberation about (potentially) harmful behaviour, but to explain the focus on pro tanto harmful behaviour in this analysis.

## 3.5 The Strength of Harm-Based Reasons

So far, I have argued that harm-based reasons are strongest if the behaviour in question constitutes harming on both dimensions.[85] The question remains, though, how we should compare harm-based reasons, when the behaviour in question constitutes harming on only one dimension. In other words, does the Difference Thesis or the Doing Thesis ground a stronger harm-based reason, or are both equally strong?[86]

In this section, I start by discussing arguments by Fiona Woollard and Molly Gardner regarding the strength of the moral reason against harming and argue that these arguments do not conclusively show that one of the dimensions grounds stronger moral reasons than the other. I then argue, however, that this should not lead us to conclude that both dimensions are always equally weighty in determining harm-based reasons. Rather, the relative weight of both dimensions can be influenced by contextual factors, such as the cost to the agent of avoiding harming.[87]

Fiona Woollard offers an argument to support the view that the Difference Thesis grounds stronger harm-based reasons. She suggests that intuitive judgements about cases support the view that '[h]arming is significantly easier to justify if the harmed person is not made worse off' (2012a, 689). (For example, it seems impermissible to kill A as a side effect of saving B's life, but permissible to kill A as a side effect of saving B's life if A would have been killed anyway (2012a, 686).) The causal account cannot explain why the reason against harming is stronger when the agent makes the victim worse off than they otherwise would have been. Woollard does not elaborate in detail, but the thought seems to be something like this: The causal account fails to recognize a difference between standard cases of doing harm (in which the agents are made worse off than they otherwise would have been) and doing harm in pre-emption cases (and non-identity cases). But the reason against harming is not very strong in pre-emption cases. So, if the reason against harming is strong in standard cases of harming, the causal account cannot explain why. However, the counterfactual comparative account does recognize a difference between these cases and gives a plausible explanation for why the reason against harming is stronger in standard cases, namely because the victim is made worse off. It looks as if the counterfactual comparative account grounds stronger reasons against harming than the causal account.

[85] I am not committed to the view that there are type 3 behaviours that are subject to harm-based reasons. However, there might plausibly be such cases, e.g. cases in which an agent is complicit in a harming behaviour without counting as either causing harm or making a difference to the outcome.

[86] A fourth possibility, one which I ignore here, is that reasons stemming from different dimensions of harming are incomparable.

[87] This does not commit me to moral particularism; it is merely to say that the relative weight of the dimensions is not fixed across situations.

The defender of the causal account might object that it is not the job of a notion of harming to explain variations in the strength of harm-based reasons. Rather, this should be explained by appeal to other morally relevant considerations.[88] However, even if other considerations play a role here, this does not undermine the point. Generally, concepts that have more explanatory power are more useful than concepts that have less explanatory power. So, if a concept of harming can explain variations in the strength of harm-based reasons, this counts in favour of the concept.

The defender of the causal account might further object that the causal account can, in fact, explain such variations. For example, Gardner (2017) agrees with Woollard's judgement about cases and also agrees with the claim that notions of harming should be able to explain the difference between intuitive judgement about cases. However, she argues that a version of the causal account, supplemented with principles that set out conditions that can weaken the strength of harm-based reasons, can explain variations in strength without having to refer to counterfactual comparative considerations. She appeals to two principles that do the work here. The Redundancy Principle says that '[o]ther things being equal, the reason against redundantly harming an individual is weaker than the reason against non-redundantly harming an individual' (2017, 86). The Inevitability Principle says that '[o]ther things being equal, the reason against harming an individual is stronger, the less inevitable it is that the individual suffers the harm' (2017, 85).

According to Gardner, an actual state of affairs T 'is less inevitable (more avoidable), the less the world would be different if T did not obtain' (2017, 84). The difference to comparative counterfactual reasoning is that how inevitable an outcome is does not depend on what would have happened had the *harming behaviour* not been performed. Rather, it depends on what would have happened had whatever constitutes the *harm* not occurred.

I believe that Gardner's view leads to problems with measuring harm. After all, it seems unclear what we are supposed to imagine in a world in which the harm doesn't obtain. It cannot be the nearest possible world, for then overdetermined harms would not constitute harms. Presumably, we should imagine a hypothetical world that is just like the actual world but for the harm. However, such a hypothetical world seems difficult to construct in cases where the harmed condition partly constitutes the victim's identity (what would the hypothetical world look like for

[88] Such as *impersonal* considerations, based on total or average utilitarian views, or values beyond individual well-being (M. Roberts 2010). However, impersonal views face further difficulties, such as the repugnant conclusion (Parfit 1984, 381–90).

a Paralympics gold medallist who is a person with a congenital disease in the real world but not the hypothetical one?).

One might think that it is not a problem that we cannot figure out what the hypothetical world would look like exactly, as long as we can say whether this hypothetical world would be better or worse than the actual world (and thereby whether the state of affairs is a harm or benefit). However, this response will not do, for this procedure would not allow us to measure harm. In order to measure harm, i.e. in order to tell how severe a harm is, we need a more precise comparison baseline that allows us to assess not just whether the agent is better off in one world than in the other, but also, how much better off.

However, Gardner's view faces worries. First, the principles seem ad hoc: why should one accept a version of the causal account that contains exactly these principles?

Second, Gardner's principles seem to reflect what I have claimed above: that the difference between pre-emption or non-identity cases and standard cases of doing harm is clearly captured by counterfactual comparative considerations. Together, Gardner's principles imply that the reason against doing harm is weaker when the behaviour does not make a difference to the outcome. This is because cases in which the behaviour does not make a difference to the outcome seem to be either cases in which the outcome would have occurred anyway (cases that are captured by the Redundancy Principle) or cases in which the harm is constitutive to someone's identity, and thus more inevitable (cases that are captured by the Inevitability Principle). The revised causal account, then, seems to give us the same results as the two-dimensional account, by at least indirectly appealing to factors about the agent's behaviour that make a difference to the outcome such as whether it determined the victim's identity. If this is true, why should we accept the revised causal account, rather than a two-dimensional account that directly appeals to counterfactual comparative considerations? Pragmatic reasons seem to suggest the latter, since the counterfactual comparative account is already well developed and offers resources to draw upon.

Third, and perhaps most importantly, Gardner's causal account does not help with the problem raised in the last section. Gardner's causal account might be able to explain why the reason against the agent's behaviour in (Pre-emptive Allowing Shooting) is weaker than in (Allowing Shooting), but it must count cases such as (Pre-emptive Allowing Shooting) as cases of harming. If we believe that counting those cases as harming is implausible, then this gives us a reason to reject Gardner's account.

Moreover, I think that there is a third, more promising, way in which the defender of the causal account could respond to the objection that the causal account can only ground a weak reason against harming. Woollard's point is that the causal account cannot explain the difference in the strength of the reason against doing harm in standard cases (stronger) and the strength of the reason against doing harm in pre-emption cases (weaker). However, pace Gardner[89], conceding this point does not require one to conclude that the causal account can only ground a weak reason against doing harm in standard cases.

There are two ways to think about this difference in the strength of the reason against doing harm. Since the causal view does not distinguish between standard cases and pre-emption cases, it either understates the strength of the reason against doing harm in standard cases, or it overstates the strength of the reason against doing harm in pre-emption cases. On the first way of thinking, which is perhaps Gardner's reasoning, the causal account cannot explain why the reasons against doing harm in standard cases are as comparatively strong as they are. On the second way of thinking, however, the causal account cannot explain why the reasons against doing harm in pre-emption cases are as comparatively weak as they are. The causal account grounds a strong reason against doing harm in standard cases, but also an implausibly strong reason against doing harm in pre-emption cases.

One might think that the second way of thinking is not a particularly attractive alternative. "Whether it diminishes or exaggerates the strength of the reason against harming", the objector might say, "the causal account has implausible implications for one set of cases either way. If the causal account is so problematic, perhaps we should simply rely on the counterfactual comparative account to measure the strength of the reason against harming!"

However, the counterfactual comparative view faces a similar problem. After all, the counterfactual comparative account fails to recognize a difference between cases of doing harm and cases of merely allowing harm.[90] But the reason against harming is intuitively much stronger in cases of doing harm than in cases of merely allowing harm. It seems that the counterfactual comparative view faces a very similar dilemma here: either it overstates the strength of the reason against merely allowing harm, or it understates the strength of the reason against doing

[89] Gardner states that she sets out to consider 'Fiona Woollard's argument for the claim that, at most, an effect-relative account of harming could ground only a weak reason against harming' (2017, 74). Here, she seems to ascribe to Woollard the view that the reason against harming arising from causal considerations is necessarily weak.

[90] However, see Purves (2019) for a version of the counterfactual comparative account that incorporates the doing / allowing distinction.

harm. So, if the reason against harming is strong in standard cases of doing harm, the counterfactual comparative account cannot explain why.

The causal account recognizes the difference between cases of doing and allowing harm. It gives a plausible explanation for why the reason against harming is stronger in standard cases, namely because the agent's behaviour is causally related to the harm in a different way. “The counterfactual comparative view has equally significant problems as the causal view”, a reply to the objector might go. “Both accounts seem to be bad at explaining certain variations in the strength of the reason against harming, and better at explaining others.”

I conclude that while the two-dimensional view can explain why the reason against doing harm in standard cases is stronger than the reason against allowing harm in standard cases, and why the reason against doing harm in standard cases is stronger than the reason against doing harm in e.g. pre-emption cases, we have no principled reason to think that one dimension gives rise to stronger reasons against harming than the other dimension.

If this is the correct conclusion to draw, then it has wide-ranging implications for moral theory and practical moral deliberation. Most deontologists accept that the reason against doing harm is stronger than the reason against merely allowing harm. But I have argued that the reason against ‘doing harm without making worse off’ is not necessarily stronger than the reason against ‘making worse off by allowing harm’. If this is correct, then the scope of the constraint against doing harm is narrower than has explicitly been acknowledged so far. This is because many real-world cases of doing harm do not make anyone worse off. This is true in particular of large-scale political decisions affecting not just the circumstances, but also the identity of future generations.

Recall the climate change case in the introduction, where we need to decide between not doing harm to future generations (e.g. climate change mitigation) and preventing harm to our contemporaries (e.g. poverty relief). (Sadly, real-world climate change is no longer a mere distant future scenario. For the sake of the argument, I pretend here that it is.) We have now seen why we should be cautious to justify such obligations by reference to the strength of the reason of ‘doing’ harm. My discussion so far suggests that duties to aid present people can be weightier than duties not to harm future people, when the harming behaviour would be identity-affecting. (Brian Berkey hints at this when he says that ‘it is far from clear, given the Non-Identity Problem, that the moral reasons against harming [future people] […] are as weighty as the reasons against harming present people’ (2014, 170).)

Another case in which the considerations discussed in this chapter might become practically relevant are cases of overdetermined harms. Regarding duties to aid the global poor, Barry and

Overland (2016), for example, argue that responsibilities to not contribute to harms are less stringent when contributing does not make a difference to whether harm occurs. However, in response to worries that this might significantly weaken constraints against doing harm, Barry appeals to intuitions about cases where agents contribute only slightly to an overdetermined harm, to show that constraints against contributing to harm can still 'require that we be willing to bear significant inconveniences to avoid acting against them' (Barry 2019, 84).

## 3.6 Conclusion

I have argued that we should adopt a two-dimensional account of harming. This account gets the scope of cases of doing harm and allowing harm right and is compatible with the Difference Thesis and the Doing Thesis. I have furthermore rejected some arguments for the view that one dimension is in principle weightier than the other.

How we think of harm-based reasons has practically relevant implications with regard to actions with long-term harmful consequences. Policy makers often face decisions to either use limited resources to prevent harm to present people or use these resources to stop activities that have harmful long-term effects. If what I have argued in this chapter is correct, then the reason against doing harm might be weakened in non-identity cases. This casts doubt on the moral weight of the Doing Thesis in intergenerational ethics, which seems like a bad result for defenders of the Doing Thesis who want to use the notion of harm to defend strong moral obligations to avoid doing harm in the long-term.

However, in response to this worry, there are at least two reasons to think that policies with long-term harmful effects are subject to harm-based moral reasons on both the causal dimension and the difference-making dimension. While I do not have the space to develop them here, they should not go unmentioned.

First, the choice of the comparison scenario is important.[91] In philosophical thought experiments, such as Parfit's Risky Policy, it is assumed that the choice is limited to two scenarios, a good scenario and a bad scenario, and that the future people in these scenarios are not identical. The

[91] This opens up the question how to construct the relevant comparison baseline for the counterfactual comparative account. On its standard version, the baseline is what would (most likely) have happened, had the agent acted differently. However, several versions of the counterfactual comparative account propose different baselines. The contrastive account (Schaffer 2010; see also Bontly 2016) relies on a baseline scenario in which the agent acts lawfully or morally. For an argument against Bontly's version of the contrastive account, see (Carlson and Johansson 2019).

nearest possible world to the good scenario is then automatically the bad scenario, and vice versa. However, this is not how policy making works in the real world. Legislative and executive processes present myriad ways to put forward, formulate, amend, and implement policies. It does not seem very likely that the policy that ends up being chosen influences the identity of *all* future people that end up being affected by this policy, simply because it is often more likely that a similar policy would have been chosen instead of the actual policy, than an entirely different policy. However, this is an empirical claim, and to which degree it is true might partly depend on the political system in question.

A second reason to think that long-term policies can be subject to harm-based reasons on the different-making dimension can be seen when we consider the involvement of different actors in putting political decisions in practice. For example, consider a government decision about whether to adopt an extremely strict policy which regulates carbon emissions in industry. It seems plausible to assume that such a policy, over time, would change the identity of all future people. If the government decides not to adopt this policy, then this decision might have adverse effects for future people. However, the government's decision does not make a difference to whether *these* future people suffer harm (the policy does not make them worse off than alternative policies, since they owe their existence to this policy). So, it seems, the government has no harm-based reason in the difference-making dimension against the high emission policy.

Now, consider a complication. Assume that the adverse effects for future people come about in the following way. The government's failure to adopt a strict policy leaves individual companies more freedom to decide whether to increase, maintain, or reduce their greenhouse gas emissions. Incentivized by the government's policy, many companies decide to emit high levels of greenhouse gases, rather than maintaining or reducing their emissions, thereby arguably causally contributing to climate harms. Every individual company's decision is unlikely to affect the identity of *all* future people who suffer climate harms (even though it might affect the identity of *some* of them). If this is true, then it seems that every individual company does have some harm-based reason in the difference-making dimension against the high emission option.

But of course, these decisions are not unrelated. The policy makers knowingly (and, presumably, intentionally) incentivize companies to act in ways that constitute doing harm on both dimensions. This might make the policy makers at least partly responsible for the difference that the companies might make to individuals who would have existed regardless of the particular companies' actions.[92] And this, in turn, might give the policy makers a harm-based reason on the difference-making dimension against their decision.

The considerations presented in the last few paragraphs are sketchy and speculative at this point. Much more needs to be said to flesh out the relevance of my discussion for real-world cases further. Importantly, the moral principles about harm-based reasons that I have discussed in this chapter have assumed actions performed by individual moral agents. They might apply in different ways to actions performed by collective or institutional agents, such as governments or nation states. However, I hope that these last considerations serve to illustrate the prima facie plausibility of the following reasoning. While it might seem that some large-scale political decisions with long-term effects do not make future people worse off (because these decisions affect everyone's identities), these decisions might foreseeably bring it about that *other* agents make future people worse off (because these agents' decisions do not affect everyone's identities and make some people worse off than these very same people would have been otherwise). In such cases, political decision makers are likely to bear at least some moral responsibility for the effects of present actions on future people.

Let me sum up the account of harm and harming that I have presented. In the previous chapter, I have defended a hybrid view on the nature of harm. In this chapter, I have defended a two-dimensional view of harming. Together, the two views give us an account of harm that answers the questions that I have distinguished at the beginning of the previous chapter. To sum up, I have defended the following answer to the question "When does one suffer harm?": One suffers harm when one either loses well-being or experiences ill-being. I have defended the following answer to the question "When does an agent have harm-based reasons against a behaviour?": An agent has a harm-based reason against behaviour that either makes a difference to harm or causally contributes to harm. In concluding, I furthermore suggested that the discussion in this chapter can provide a starting point for further research on the practical application of these principles, in particular regarding cases of institutional or collective agency.

[92] I say a bit more about proximity conditions for harm-based reasons in section 6.3. However, my discussion there will be limited to individual moral agency and might therefore not be straightforwardly applicable to the cases discussed in this paragraph.

# Chapter 4 Letting Oneself Do Harm

## 4.1 Introduction

The Doctrine of Doing and Allowing (DDA) says that there is a morally relevant difference between doing and allowing harm.[93] Recall that the DDA consists of two claims. The first claim is a constraining principle. It limits what it is permissible for agents to do (some harm doings are impermissible). I call this claim the Doing Thesis. The Doing Thesis says that the reason against doing harm is stronger than the reason against merely allowing equivalent harm. It implies that at least in some cases, it is not permissible to do harm in order to prevent more harm. The Doing Thesis can explain, for example, why killing one person is harder to justify than merely letting two others die. The second claim is a prerogative principle: it widens what it is permissible for agents to do (some harm allowings are permissible). I call this claim the Allowing Thesis. The Allowing Thesis says that the duty to prevent harm is non-maximal. The Allowing Thesis implies that at least in some cases, it is permissible to allow harm, without thereby securing an equivalent benefit. It can explain, for example, why it is permissible to go on holiday rather than giving money to charity.

Defenders of the DDA need to show two things. First, they need to show that we can draw a reasonably clear distinction between doing and allowing harm. Second, they need to show that this distinction is morally relevant in a way that supports both the Doing Thesis and the Allowing Thesis. Problems emerge for the defender of the DDA when we start considering the moral status of cases that intuitively seem to be the kind of cases that the DDA should apply to, but it is unclear how the DDA applies to these cases.

In the remainder of this thesis, I discuss three types of cases that pose difficulties for the defender of the DDA. These cases do not seem to straightforwardly fit in either of the two boxes: they are neither clear cases of doing harm nor clear cases of merely allowing harm.[94]

[93] J. Hanna (2015a), whose work I focus on in the following, uses this characterization of the DDA following Quinn (1989a).

[94] A fourth type of such cases that has received much attention in the literature are so-called “Safety Net” cases, in which agents remove a barrier to a harmful sequence (Woollard and Howard-Snyder 2016). Some philosophers classify these cases as doings (e.g. Bennett (1995)), others classify them as allowings (e.g. Rickless (1997)). McMahan (1993) argues that some safety net cases are doings and others allowings. Hanser (1999) and Hall (2008) suggest that these cases form a distinct third category.

## 4.2 Letting Oneself Do Harm

The challenge for defenders of the DDA that I discuss in this chapter has been posed in recent work by Ingmar Persson (2013) and Jason Hanna (2015b; 2015a). The challenge is to specify the moral status of so-called cases of letting oneself do harm. Consider:

> (Poisoner) Earlier this morning, Agent deposited a dose of lethal poison into a teapot from which Victim drinks tea at the same time each afternoon. Unless Agent warns Victim now, Victim will drink the tea and die (Woollard and Howard-Snyder 2016; the original version of this case is found in J. Hanna 2015a, 678).

Failing to warn Victim would not be killing Victim. Rather, it would be failing to prevent Agent's previous behaviour (putting poison in the teapot) from constituting a killing. By failing to warn Victim, Agent would *let herself do harm* (J. Hanna 2015a, 679). What is the moral status of cases of letting oneself do harm?

Consider a variant on (Poisoner). Assume that Agent can either warn Victim or save the lives of two innocent strangers, who are drowning in a shallow pond. Agent does not have the time to do both. Intuitively, Agent ought to save Victim's life (J. Hanna 2015a, 678). It seems to matter morally that Agent herself has created the threat that Victim faces. Let us call this intuition the 'self-other divide' (Persson 2013, 103).

Deontologists[95] could explain the self-other divide by arguing that letting oneself do harm has the same moral status as doing harm. Just as it is impermissible to kill one person to save two others, it is then impermissible to let oneself kill Victim to save two strangers.

However, consider a second variant on (Poisoner). Assume that *two* victims will drink tea out of the poisoned tea pot, unless Agent warns them in time. However, Agent can only reach the victims in time if she runs over (and thereby kills) an innocent stranger trapped on the road. Intuitively, Agent ought not to kill the stranger (J. Hanna 2015a, 681). It seems to matter morally whether the harming behaviour occurs now or has been performed in the past. Let us call this intuition the 'present-self focus' (Persson 2013, 103).

[95] In the following, by "deontologists", I refer to those deontologists whose moral theory contains the DDA (it should be noted, however, that much of my discussion is also relevant for defenders of other moral theories that support or are compatible with DDA). I briefly address other deontological distinctions in sections 4.2 and 4.5.2.

Deontologists could explain the present-self focus by arguing that letting oneself do harm has the same moral status as merely allowing harm. Just as it is impermissible to kill one person to save two others, it is then impermissible to kill one person to let oneself do harm to two victims.

These explanations seem to conflict. This illustrates that the intuitions behind the self-other divide and the present self-focus point to two different ways of thinking about the moral status of cases of letting oneself do harm. These cases seem to occupy a curious status in between doing harm and merely allowing harm. The challenge for deontologists that arises from these cases is to provide a coherent account of cases of letting oneself do harm that reconciles these intuitions.

If deontologists fail to explain intuitive verdicts about cases such as (Poisoner), this is bad, insofar as it is usually claimed to be an advantage of deontological moral theories that they can explain common-sense moral intuitions. This is in contrast to views such as act-utilitarianism, which counter-intuitively imply that Agent ought to save the two drowning strangers instead of warning Victim, and that Agent ought to run over the stranger on the road to save the two Victims.

In this chapter, I explore J. Hanna's and Persson's challenge for deontologists. I argue for two claims. The first claim is that the challenge presented by cases of letting yourself do harm is more complicated and more urgent than J. Hanna and Persson have argued. The second claim is that deontologists nonetheless can respond to this challenge. Deontologists should analyse cases of letting yourself do harm as non-standard cases of allowing harm.

The chapter proceeds as follows. In section 4.3, I consider cases of letting yourself do harm and their more complicated relatives. For example, J. Hanna (2015a, 689–95) points out that difficulties arise from cases in which agents can undo a previous decision to not let themselves do harm. I argue that we can complicate cases even further, such that simple cases of letting oneself do harm like (Poisoner) are only a subset of cases that involve decisions about whether and in what ways to let oneself do harm. Some choices are about whether to let oneself do harm. Other choices are about whether to let oneself do harm to one group, rather than let oneself do harm to another group. Yet other choices are about whether to let oneself face a choice about whether to let oneself do harm to one group, rather than let oneself do harm to another group. I argue that there are many possibilities for further complication.

In section 4.4, I argue that cases of letting yourself do harm are not just a theoretical curiosity. I suggest that they are widespread in real life, and indeed, central to intergenerational ethics. Insofar as cases of letting oneself do harm challenge deontologists, they challenge the ability of deontologists to provide moral guidance in real-world cases such as climate change.

In section 4.5, I discuss possible analyses of cases of letting oneself do harm. Deontologists can analyse cases of letting oneself do harm as (i) non-standard cases of doing harm, or (ii) non-standard cases of allowing harm.[96] I argue against (i), on the basis that this classification would require deontologists to accept the implausible claim that the temporal order of harming behaviours matters morally. I then argue for (ii). Deontologists should accept the claim that agency over time matters morally. I argue that this claim can be defended by appeal to special obligations towards those one has wronged in the past or will wrong in the future.

## 4.3 Even More Complicated Cases

In this section, I argue that cases of letting oneself do harm are only one type of a whole family of cases with a complicated causal structure, whose moral status seems similarly unclear. The challenge presented by cases such as the letting oneself do harm is therefore much broader in scope than has been acknowledged.

Let us start by looking at the causal structure of cases of letting oneself do harm more closely. It will help to contrast them with standard cases of doing and allowing harm. A natural way to understand the distinction between doing and allowing harm is in terms of the causal relation between an agent's behaviour and a harmful outcome.

> (Push) Agent pushes a boulder which rolls over Victim, crushing her to death.
>
> (Non-Diversion) A boulder is rolling towards Victim. Agent could stop the boulder but does not do so. The boulder rolls over Victim, crushing her to death. (These cases are variants of cases discussed in (Woollard 2015; 2012b; Bennett 1993).)

(Push) is a standard case of doing harm. Agent pushes the boulder (harming behaviour) and Victim dies (harmful outcome). The harmful outcome follows directly from the harming behaviour. (Non-Diversion) is a standard case of merely allowing harm. Agent stands still when the boulder rolls towards Victim (allowing behaviour) and Victim dies (harmful outcome). If Agent had behaved differently (i.e. diverted the boulder), the harmful outcome would not have occurred. The DDA says that pushing the boulder in (Push) is harder to justify than failing to stop the boulder in (Non-Diversion), everything else being equal (Foot 1967; Kamm 2007; Quinn 1989a; Woollard 2015).

[96] I will also briefly discuss the option that cases of letting oneself do harm form a third category of in-between 'dallowing' cases.

In the real world, causal chains leading to a harmful outcome are much more complicated than in (Push) or (Non-Diversion). In reality, causal chains are extended in time. They might involve additional or intervening actions or events. Whether the harmful outcome occurs or not might, in part, depend on these additional or intervening factors.

It will be useful for our discussion to think about cases of letting oneself do harm, such as (Poisoner), in a more abstract way as one type of such complicated cases. We can think of them as cases in which an agent is relevant to a harmful outcome through more than one behaviour. I illustrate this in figure 1.

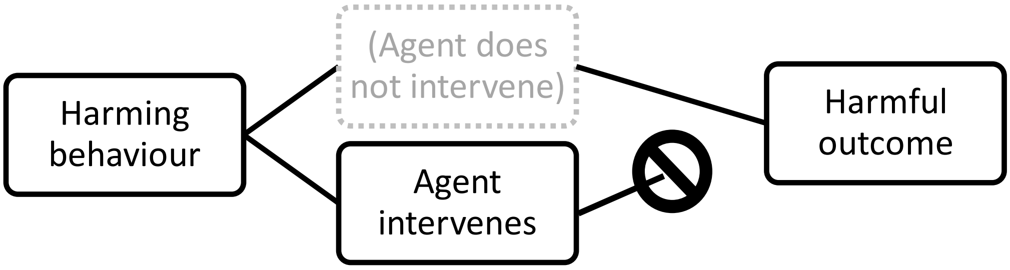

Figure 1

At the time of the harming behaviour (for example, when putting poison in the teapot), an agent creates a threat now that she has the opportunity to avert later, or that she could have averted earlier. At the time of the possible intervention, an agent either intervene or not. If the agent intervenes, she thereby stops the causal sequences leading to the harmful outcome. If the agent does not intervene, the threat that she herself has initiated leads to the harmful outcome. In (Poisoner), the intervention would be warning Victim. Failing to warn Victim seems harder to justify than merely allowing harm (this intuition is captured by the self-other divide), but easier to justify than doing harm (this intuition is captured by the present self-focus).

J. Hanna brings up a further complication. Consider

> (Two Buttons) Last week, Agent initiated a threat that is now about to kill five people. Five minutes ago, Agent realized that she could save the five by pressing button 1, but that pressing button 1 would also kill one. She has pressed button 1. Now, she has the opportunity to press button 2. Pressing button 2 will interfere with the reactions triggered by pressing button 1 and will prevent both the saving of the five and the killing of the one. (This is a shortened version of a case given by J. Hanna (2015a, 689–90).)

The present self-focus implies that Agent's pressing button 1, thereby initiating the reaction, is morally wrong. But what should deontologists think about pressing button 2?

According to J. Hanna, they 'would presumably claim that Agent is morally required to press Button 2 and terminate the reaction. This is because it is very counterintuitive to hold both (a) that Agent is morally forbidden from initiating the reaction and (b) that, if he does initiate the reaction, he is permitted to let it continue when he could easily stop it' (2015a, 690).

Perhaps J. Hanna is right. The assumption that we should revert our wrong actions, if we can, seems intuitively plausible, and indeed implies that Agent should press button 2.

However, *pace* J. Hanna, one can argue that pressing button 1 has changed the decision situation that the agent faces when she considers whether to press button 2. After pressing button 1, the five are no longer under a threat, and it is impermissible to recreate this threat. On this view, Agent should not press button 2.

Be that as it may, it is not clear how the DDA would lend support to either view. Agent has initiated a threat to the five, and a threat to the one. She would let herself do harm in pressing button 2, and she would let herself do harm in not pressing button 2. Insofar as the moral status of letting oneself do harm is unclear, the moral status of cases such as (Two Buttons) is similarly unclear.

However, deontologists might object to this claim. They might argue that it is not the business of the DDA to give verdicts in cases like (Two Buttons) and (Poisoner). Rather, the moral status of these cases is determined by considerations outside the scope of the doing/allowing distinction. I defend the claim that at least some accounts of the DDA would be incomplete without an account of cases of letting oneself do harm in section 4.5. In the meantime, let me offer some reasons to think that neither an appeal to numbers nor appeal to the moral relevance of intention can fully explain the moral status of cases such as (Two Buttons).

Can deontologists just appeal to the numbers? Surely, letting oneself do harm to five people is worse than letting oneself do harm to only one person, everything else being equal. If this is true, then Agent should not press button 2. However, putting aside any principled reservations that deontologists might have about deciding by the numbers,[97] appeal to numbers does not get deontologists all the way.

Here is why. If J. Hanna is right and we have stringent requirements to revert our wrong actions, then deontologists should presumably press button 2, even though pressing button 2 saves less people overall. If, however, prohibitions against recreating threats are stronger than

[97] See e.g. Taurek (1977) for the view that in cases in which we need to trade off harms and benefits, the numbers do not have intrinsic moral significance.

requirements to revert our wrong actions, then deontologists should presumably not press button 2. Importantly, if prohibitions against recreating threats are this strong, it seems that they would prohibit pressing button 2 even in a different case in which pressing button 2 would save less people than not pressing button 2 (imagine that Agent has triggered a reaction that will kill five, in order to save one person, and she can now revert this reaction by pressing button 2). An appeal to numbers, therefore, seems in conflict with both the view that recreating threats is more objectionable than failing to revert one's wrong actions, and the opposite view that failing to revert one's wrong actions is more objectionable than recreating threats.

To be clear, I am not suggesting that the numbers do not or cannot matter on deontological views. I make the different claim that two plausible lines of deontological reasoning about cases such as (Two Buttons) seem to recommend saving the smaller number in at least some cases. If this is correct, then deontologists cannot simply appeal to the numbers in cases like (Two Buttons).

Alternatively, one might wonder whether deontologists can appeal to the distinction between intending and merely foreseeing harm.[98] The line of thought here might be the following. Poisoner might have put poison in the tea by accident, without intending to kill the five. However, when she drives over the one to save the five, she does not merely foresee the death of the one. Rather, she intends the death of the one. Driving over the one is impermissible, because killing someone intentionally is harder to justify than killing as a foreseen effect. However, appealing to the role of intention is unlikely to solve the problems raised by the cases discussed in this section.[99]

Whether Agent had the intention to kill does not by itself seem to change intuitive verdicts about cases such as (Poisoner). Whether Agent has put the poison in the teapot intending to kill Victim, or being merely negligent, or finding the only safe spot to hide it from the Mad Poisoner who is about to visit does not seem to change the intuition that Agent ought to save her potential poison victim rather than five strangers but should not drive over one other person to save two such potential victims.

More generally, we can easily imagine complicated cases in which Agent has, in the past, started harmful sequences with the intention of doing harm. Agent can now intervene and avert some of

[98] The claim that there is a morally relevant difference between intending and merely foreseeing harm is known as the Doctrine of Double Effect (DDE) (McIntyre 2019). The DDE has been defended e.g. by Quinn (1989b), but see e.g. (Kagan 1989, 128–82) for criticism.
[99] I am here in agreement with J. Hanna (2015a, 693).

these threats – but only by intentionally doing harm. Whatever Agent's choice, she will have intentionally done harm.

These considerations suggest that neither appeal to numbers nor to the relevance of intention can clearly provide an account of the moral status of complicated cases. I conclude that my previous claim still stands: insofar as the DDA does not provide a clear account of the moral status of cases of letting oneself do harm, the same can be said about cases with a more complicated causal structure, such as (Two Buttons).

I will now argue that (Two Buttons) is just one case in a family of cases in which agents can intervene in the causal chain leading to harm more than once. There are many possible cases that share a similar structure. An agent might have created, or know that she will create, multiple threats. She might now be able to avert some, but not all, of them. Or she has previously allowed harm that she now has a second chance to avert, but only at the cost of creating a further threat.

It will be helpful to consider the causal structure of such cases such as (Two Buttons). In these cases, an agent is relevant to more than one possible outcome through a single intervention behaviour. Figure 2 illustrates the Agent's options when she decides whether to press button 2 in (Two Buttons). She has previously imposed a threat on the five (first harming behaviour) and the one (second harming behaviour). Agent can now intervene with her second harming behaviour by pressing Button 2. If Agent presses button 2, the causal sequence leading to the death of the one (harmful outcome 2) will be stopped, but the causal sequence leading to the death of the five (harmful outcome 1) will be re-opened. If Agent does not press button 2, the causal sequence leading to the death of the five remains stopped, and the causal sequence leading to the death of the one remains open. In sum, the Agent's present choice is about which of her earlier harmful behaviours will lead to harm.

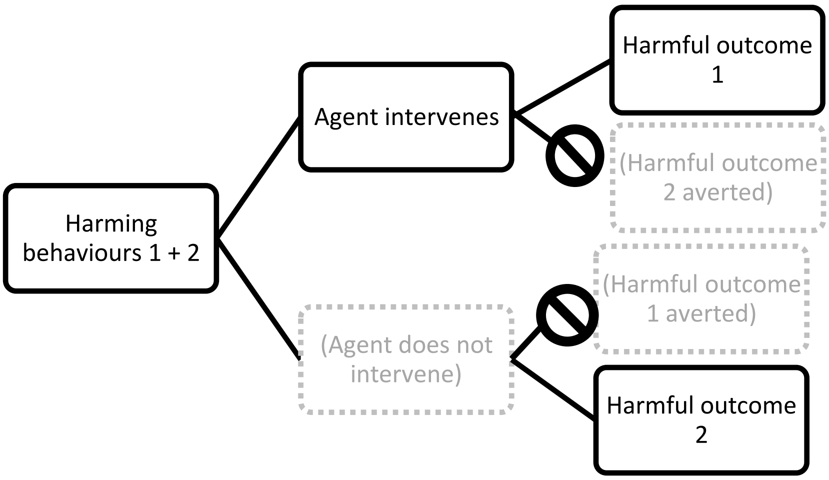
Harming behaviours 1 + 2
Agent intervenes
Harmful outcome 1
(Harmful outcome 2 averted)
(Agent does not intervene)
(Harmful outcome 1 averted)
Harmful outcome 2

Figure 2

To make matters worse, cases like (Poisoner) and (Two Buttons) do not exhaust the possibilities of complicating causal chains. Consider

> (Third Button) Agent anticipates well ahead of time that she will be in the situation described in (Two Buttons). She is uncertain about what she is supposed to do in that situation. She can now press button 3. Pressing button 3 disables button 2, so that once Agent has pressed button 1, no further intervention will be possible (and she will be spared the resulting moral quandaries!).

(Third Button) is an example for a case in which an agent acts in a way that influences whether, and if so how, she will be able to intervene in the causal chain leading to a harmful outcome in the future, as illustrated in figure 3.

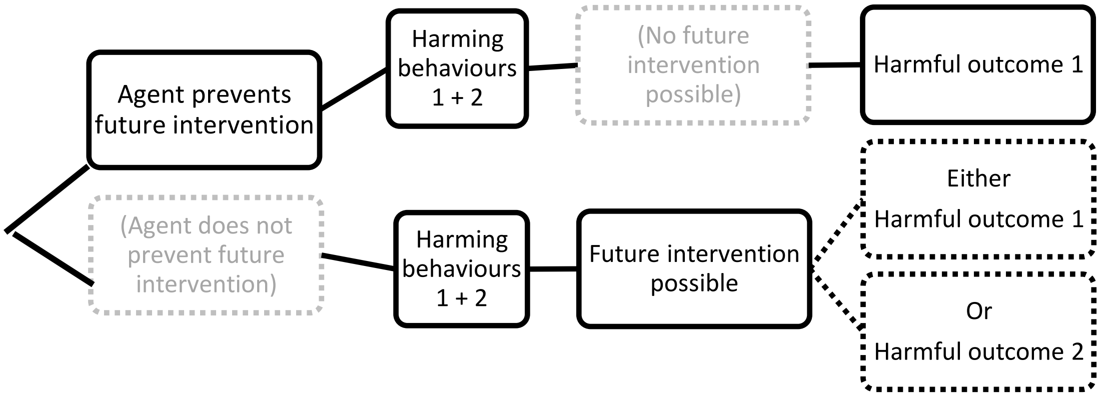
Agent prevents future intervention
Harming behaviours 1 + 2
(No future intervention possible)
Harmful outcome 1
(Agent does not prevent future intervention)
Harming behaviours 1 + 2
Future intervention possible
Either Harmful outcome 1
Or Harmful outcome 2

Figure 3

What should deontologists think about cases such as (Third Button)? If we should revert a wrong decision when we can do so, then it seems equally plausible that one should not prevent oneself from being able to revert a wrong decision in the future. Therefore, if one believes that Agent ought to press button 2, it seems that one should also believe that Agent ought not to press button 3.

However, if one believes that Agent ought not to press button 2, it seems that one should also believe that Agent is at least permitted to press button 3, since doing so would only make a morally wrong future option (pressing button 2) unavailable.

Again, the DDA does not seem to support either verdict. I have said above that Agent lets herself do harm both by pressing button 2 and by not pressing button 2. Therefore, when pressing button 3, Agent prevents herself from letting herself do harm later. When refraining from pressing button 3, Agent lets herself let do harm later. Since the moral status of pressing button 2 was unclear to begin with, the moral status of pressing button 3 seems even less clear.

There are several possible cases with a similar structure as (Third Button). For example, an agent can now initiate a threat while making sure that she will have an opportunity to avert the harmful outcome later. Or she can do something now to prevent a harmful outcome, while making sure that she can later undo the prevention, in which case the harmful outcome would nonetheless occur. (Third Button) is therefore just one of a family of complicated cases which share causal features with standard cases of doing and allowing harm. The moral status of these cases is unclear.

We can conclude that the range of cases in which the DDA does not deliver a clear verdict goes far beyond cases like (Poisoner). The possibilities for further complication are endless.

## 4.4 Complicated Cases in the Real World

These cases are admittedly confusing, and my anecdotal evidence suggests that few people have clear intuitions about them. All examples discussed above, from (Poisoner) to (Third Button), are very complicated and highly artificial. Given this, one might think, perhaps it is not so bad after all that we cannot come up with a theoretical account of such cases.

Many people can live very comfortably in the presence of unsolved quibbles in the periphery of our preferred moral theories. And even those who would prefer their moral theory to apply even to the most fantastic hypothetical cases might think that cases of letting oneself do harm do not make the priority list of problems to which defenders of the DDA should devote time and energy. After all, in real life, we do not often seem to find ourselves in scenarios such as (Poisoner). Everyday harm doing seems to be much more straightforward: We do something, and the harm is done, usually before we had another opportunity to intervene.

However, in the following, I suggest that there is a set of real-life cases that is structurally similar to the complicated cases discussed above. These real-life cases frequently arise in intergenerational ethics, specifically in cases involving doing harm to future generations.[100]

In such cases, by definition, we deal with the long-term consequences of our present behaviour. Present actions seldom have direct and inevitable effects that will obtain in a few centuries. Usually, there is something that we can do, now or later, to intervene in the causal chain leading

[100] It is usually seen as an intuitively plausible claim that future generations can be harmed. However, this claim has not gone unchallenged in intergenerational ethics. For an overview of criticisms and responses, see e.g. Campos (2018, 3–4).

to harmful outcomes in the (far) future. If we refrain from doing so, we let ourselves do harm. For example, think about climate change, and all the ideas that scientists have come up with to prevent global warming: from reducing carbon emissions in the first place, over technology to lower the level of carbon already in the atmosphere, to mitigating the adverse effects caused by global warming.

Of course, it is not straightforward who the ‘we’ in such cases refers to. Possible candidates include: ‘We’ = the developed nations that continuously emit carbon dioxide, thereby causing climate change. ‘We’ = nation so-and-so, which has in the past have exceeded the limits on carbon dioxide emissions. ‘We’ = city so-and-so, which will not enforce new regulations to ban plastic straws in restaurants. ‘We’= the Joneses, who still eat meat and drive cars. There are countless complexities arising here, regarding collective and shared agency, the notion of causation, and collective and institutional agency over time.

However, it seems plausible that at least in some such cases, the moral agent in question exists over time in a relevant sense. Consider the case of a government that provides extensive funding for coal-fired power stations. It seems intuitively plausible that the government is doing harm in contributing to climate change. It also seems plausible that, if the government 50 years from now is presented with an opportunity to combat climate change, they would let themselves do harm if they did not avert climate change. Even if the people who constitute the government will have changed, in a sense the government can be ascribed institutional agency over time. It is in an important sense the same government that has funded the power stations and later failed to combat global warming.

More generally, the following assumptions seem plausible: At least some collective and institutional agents such as governments, supranational organisations, companies, the inhabitants of village X, and perhaps even groups such as the ‘Millennials’, or the global affluent, possess agency over time in a relevant sense. They can therefore let themselves do harm to future generations. Because of the large time gap between harming behaviour and harmful outcome, these agents will often have numerous opportunities to intervene with their own actions.

To illustrate, consider a climate change example that is structurally similar to (Poisoner):

> (Climate Change I) In the past, we have emitted carbon dioxide in the atmosphere. If we do not take action now to mitigate climate change, future generations will suffer from, amongst other effects, a rise in infectious diseases.

Similarly, the following case is structurally similar to (Two Buttons):

> (Climate Change II) Climate Change will lead to a future rise in infectious diseases. We initiate mitigation through increased use of green technology, even though this will inevitably cause comparable harm – perhaps a future rise in infectious diseases due to industrial activities. However, shortly before these industrial activities start, we begin to doubt the morality of our action and consider stopping green technology.

The following case is structurally similar to (Third Button):

> (Climate Change III) Climate Change will lead to a future rise in infectious diseases. We can now decide to adopt regulations that prevent us from stopping green technology, once we have decided to go for it. This would mean that after we have made the decision to mitigate by increased use of green technology, we would not be able to revert this decision and industrial activities would inevitably cause a huge rise in infectious diseases.

It is beyond the scope of this chapter to consider the many fascinating issues that arise from these cases. One particularly interesting question is who qualifies as an agent that can let themselves do harm (and bear responsibility for this behaviour). While I cannot hope to conclusively answer this question here, I offer some considerations which underline my claim that cases of letting oneself do harm are more common than one might think.

An agent who can let themselves do harm needs to be the kind of agent who (i) is able to 'do' or 'allow' harm in the first place, and (ii) persists over time in a way that makes it possible to ascribe them responsibility for earlier actions. Regarding the first criterion, it seems plausible that in order to be able to do or allow harm, an agent must possess some causal power in the sense that their behaviour potentially makes a difference to the outcome. For example, one individual presumably cannot, on their own, make a perceptible difference to the overall level of carbon dioxide emissions. However, the signatories of the Kyoto Protocol can.

Regarding the second criterion, to persist over time in the relevant sense, an agent must be clearly identifiable as the agent performing both the allowing and the doing behaviour. For example, a city council might count as letting themselves do harm when they refrain from replacing the faulty train switch boards that they bought cheaply years ago, leading to runaway trolleys on the rails (and pedestrians being pressured into putting their moral views in action!). The interest group that has the opportunity to put pressure on the city council to replace the switch boards, but fails to do so, might count as allowing harm, but they do not let themselves do harm (assuming that they had not been involved in the previous decision to buy the cheap switch boards).

These considerations should lead us to resist the temptation to argue that, since past generations have caused climate change, present generations fulfil their obligations not to harm by keeping present emissions low, thus "not adding to the burden". Insofar as the agent who caused climate change in the past is best described as an agent existing over time (e.g. 'developed nations'), they still count as letting themselves do harm if they refrain from mitigating, even if they did not produce additional emissions.

One of the central aims of moral theorizing is to arrive at action-guiding conclusions. An essential part of that is to figure out how moral concepts and theories apply to real-world cases. If that is correct, then finding out how to apply the DDA to real-world cases in intergenerational ethics – including complicated cases such as those described above – is an urgent task. Indeed, I think the urgency of that task is underappreciated in contemporary moral philosophy.[101]

To illustrate the urgency of an account of the moral status of complicated cases, let me point out worries that arise in the absence of such an account. I suggested that complicated cases are much more common in future generation cases than in present generation cases, and the moral status of many future generation cases is therefore unclear. If this is right, then it seems natural to conclude that moral obligations towards future generations often differ from moral obligations towards present people. As long as the moral status of complicated cases remains unclear, the moral status of future generation cases remains unclear.

For example, when facing a decision between an option that involves a risk of standard harm doing and an option that involves a complicated case, it would be unclear whether, why, and how much one of these would be harder to justify than the other. In other words, not having an account of such complicated cases might give us a reason to reject either the arguably widespread belief that our obligations towards future generations are relevantly like obligations to those presently alive, or the distinction between doing and allowing harm itself – at least regarding its usability for actual decision-making.

However, we need not accept these implications just yet. What I have said so far might be less of a challenge than a pointer towards an area that is currently underexplored. This is what deontologists who accept the DDA might say: "Clearly, what you call the 'complicated cases' are not clear-cut cases of either doing or allowing harm. These cases are what one might call non-standard cases. They are subsets of harm doings, or perhaps harm allowings, and they are

[101] The point that cases of letting oneself do harm are common in real-world moral decision-making has, as far as I am aware, not been made in the literature. If this point is correct, it highlights the importance of the moral problem posed by these cases, and the practical relevance of moral theorizing about these cases.

different from those standard cases in morally relevant ways. If we can specify these differences in terms of causal relations to harm, this is great – spelling this out in more detail would tell us more about the dynamics that make a certain causal relation to harm easier, or harder, to justify."

If this is right, then my argument in the last section does not point to a flaw in deontological moral theory. Rather, it draws attention to an area where the otherwise abundant literature on the DDA remains surprisingly silent and points out the urgency to explore this area further. In the remainder of the chapter, I will start this exploration by looking at some possible directions that the defender of the DDA could take.

## 4.5 The Moral Status of Letting Oneself Do Harm

I start by motivating the claim that cases of letting oneself do harm are a challenge for defenders of the DDA. I will do this by providing a reason to think that the DDA would be incomplete without a causal analysis of complicated cases, and an account of the moral status of such cases.

Is there a moral difference between doing and allowing harm? Recent defences of the DDA provide a two-step strategy for answering this question in the affirmative. First, they analyse the distinction between doing and allowing harm. Second, they argue that the underlying nature of this distinction makes it morally relevant (Woollard and Howard-Snyder 2016). Most recently, Fiona Woollard (2013; 2015), following Philippa Foot (1967), distinguishes doing from allowing harm in terms of the causal relation between an agent's behaviour and a harm. Woollard's first-step analysis of the doing/allowing distinction relies on facts about how the agent's behaviour is relevant to the harm in question, where facts about the agent's behaviour form part of a causal sequence leading to harm.

Recall that according to Woollard (2015), an agent's behaviour is *doing* harm if and only if a fact about the behaviour is part of a sequence leading to harm. An agent's behaviour is *merely allowing* harm if and only if a fact about the behaviour is relevant to, but not part of, a harmful sequence. Only *substantial* facts - roughly, facts that are either positive, or against normal presuppositions – can be part of a sequence. Having thus analysed the distinction between doing and allowing harm, Woollard then goes on to argue that the distinction is morally relevant. Roughly, on her view, the DDA protects moral agents from undue imposition. It provides this protection by requiring agents to refrain from doing harm and by permitting them to allow harm to others.

However, these are exactly the kind of facts that seem to distinguish letting oneself do harm from merely allowing harm. If defenders of the DDA think that these facts make a moral difference, then they should be able to say whether, and if so, how exactly these facts make a moral difference in the case of letting oneself do harm (and, indeed, more complicated cases that I will consider in the following).

Letting oneself do harm as in (Poisoner), prioritizing saving some of one's own victims over others as in (Two Buttons) and barring oneself from possible future interventions as in (Third Button) are behaviours that are clearly relevant to the harm that occurs. However, it is less clear *how* the agent's behaviour is relevant to the harm in these cases: Is it in a doing way, in an allowing way, or in a yet-to-be-specified third way? In the remainder of this section, I will explore these possibilities in turn. I will focus on "simple" cases of letting oneself do harm, such as (Poisoner). An account of the moral status of these cases would constitute a first step towards a more comprehensive account and is likely to shed light on more complicated cases.

### 4.5.1 Letting Oneself Do Harm as Doing

If Agent does not warn Victim in (Poisoner), she will have done harm. Victim will die as a result of Agent's poisoning the tea. Agent will have killed Victim! So perhaps one might think that letting oneself do harm is best seen as an instance of doing harm.

One prima facie plausible way to think about this is to say that one should understand Agent's harm doing as extended in time. The extended harm doing contains all behaviours relevant to the harm: putting the poison in the teapot and failing to warn Victim.

As pointed out above, however, the view that letting oneself do harm has the moral status of doing harm fails to explain the present self-focus. Recall that this was the intuitive claim that *doing* harm now (e.g., by driving over Stranger) is much harder to justify than *letting* oneself do harm (e.g., by failing to warn Victim now).

However, perhaps deontologists can provide a separate defence of the present self-focus. And indeed, it seems like there is a good justification for focusing on the agent's present behaviour. This is, roughly, what the defender of the present self-focus might say: "As human beings, we exercise our agency at present. We typically have immediate access to current deliberations, and immediate control over current decisions. Our control over our present behaviour is typically greater than our control over our past or future behaviour. This is why our present choices carry special moral weight. Initiating a harmful sequence now is therefore harder to justify than not stopping a harmful sequence that you have initiated."[102]

This line of argument can explain why threats that we create now matter more (as in Poisoner). However, it cannot explain why one of two threats that we have created in the past seems to matter more than the other (as in Two Buttons).

It might be tempting for the defender of the present self-focus to argue that temporal nearness of the harming behaviour matters – perhaps it does so indirectly, through the importance that we attach to the present decision-making.

However, even if such an argument could be made, it would not be able to account for all cases that are relevantly like (Two Buttons) (and, for that matter, (Climate Change II)). This is because it does not seem to be true that our intuitions track temporal nearness.[103] Rather, they seem to be tracking temporal order of the harming behaviours. Consider, again, (Two Buttons). Imagine that Agent knows she will initiate a threat to the five. She also knows she will then press button 1. She can now disable button 1, such that when she presses it later, nothing will happen.

If temporal nearness of the harming behaviour mattered morally, we would expect that, other things being equal, we ought to interrupt the harmful sequence that we will initiate sooner, and that Agent should therefore not disable (and indeed, should later press) button 1.

However, intuitively, just as in the original (Two Buttons) case, it seems wrong to press button 1, and thus permissible to disable button 1 now. When Agent disables button 1, she prevents herself from initiating a harmful sequence to the one. She does not prevent herself from initiating the (earlier) threat to the five.

If this is correct, then it is misleading to say that our intuitions track the temporal *nearness* or *proximity* of harmful behaviour. Rather, they seem to track the temporal *order* of harmful behaviour, or threat initiation: It is more objectionable to let oneself initiate a new threat than it is to refrain from interrupting a threat that one has previously caused. Even if an argument can be made that supports the view that, as Thomson puts it, the 'present tense matters' (1985, 1401), it is unclear how deontologists would build on such an argument in order to account for the temporal order of behaviours.

[102] Arguments along these lines have been brought forward by (Thomson 1985, 1415; Woollard 2014, 276).
[103] I here disagree with J. Hanna, who suggests that our intuitions about what matters morally track 'temporal proximity, or […] how close to the present the harming behaviour would occur' (2015a, 696).

### 4.5.2 Letting Oneself Do Harm as Allowing

So maybe letting oneself do harm is better described as allowing harm. However, as pointed out above, the self-other divide seems to indicate that this is mistaken. Letting oneself do harm is, everything else being equal, harder to justify than merely allowing harm. In (Poisoner), the fact that it was Agent who put the poison in the teapot makes a moral difference.

One might think that this is not surprising. Letting oneself do harm, after all, is not *merely* allowing harm. Therefore, we should not expect its moral status to be equivalent to *merely* allowing harm. And it is easy to see that some instances of allowing harm are harder to justify than other instances of allowing harm.

There are at least three factors that can change the moral status of clear cases of allowing harm. First, facts about the relationship between Agent and Victim might matter morally. For example, not saving Victim might be harder to justify if Agent is Victim's carer or has promised to save Victim. Second, facts about the situation might matter morally. For example, not saving Victim might be easier to justify if it would be very difficult for Agent to do so. Third, facts about Agent's epistemic or motivational state might matter morally. For example, allowing harm might be easier to justify if the harm is a foreseen, but not intended, consequence of Agent's behaviour.

However, the intuition that letting oneself do harm is harder to justify than merely allowing harm persists even when these three factors are varied. It therefore seems that whether an agent lets themselves do (or merely allow) harm is not a proxy for a morally relevant difference, but itself possesses moral relevance.

Agent intuitively ought to save Victim rather than Stranger, even when facts about the situation, the relationship between Agent and Victim, and Agent's knowledge, intentions and motivations change. To illustrate, Agent should save Victim even if Victim is a stranger, even if Agent did not maliciously intend the poisoning, and even if Agent was pressured into poisoning Victim.

While it is entirely plausible that these three factors matter morally, they do not explain the difference between cases of letting oneself do harm, and cases of merely allowing harm.

Fortunately, there is an alternative, intuitively plausible explanation of the difference between cases of letting oneself do harm and cases of merely allowing harm. The difference lies in the way in which Agent's behaviour is relevant to harm. What matters is that, in addition to Agent's present allowing behaviour, in cases of letting oneself do harm there are facts about Agent's past or future behaviour that are relevant to the harm in a doing way.

In other words, Agent is relevant to a harm through two behaviours, one of which is clearly a doing behaviour. If Agent does not warn her victim in (Poisoner), she thereby ensures that she will have done harm. She allows her past behaviour of poisoning the tea to constitute a harm doing.

To be more precise, it seems that letting oneself do harm is different from merely allowing harm in three ways: (i) facts about the agent's behaviour (distinct from the present behaviour) are relevant to the harm in question; (ii) these facts potentially qualify as harm doing; and (iii) whether they do so depends on facts about the agent's present behaviour.

However, taken by itself, the mere fact that one's present behaviour can influence the effects of one's past or future behaviour does not provide an explanation for the self-other divide (the intuition that letting oneself do harm is harder to justify than merely allowing harm). Nor does the mere claim that in cases of letting oneself do harm, we need to include all potentially relevant past and future behaviours in our evaluation. Surely, we would want to be able to evaluate an agent's behaviour before the last opportunity for the agent to interfere in the causal chain leading to harm has passed and the harm has occurred. The DDA would not be very useful if it told us to refrain from the moral evaluation of harming behaviours, effectively until the harm has materialized.

A more promising way to explain why letting oneself do harm is harder to justify than merely allowing harm is to provide a more robust defence of the claim that agency over time matters. Such a defence might start by pointing out that as human beings, we are essentially embodied. This is, roughly, what the defender of the self-other divide might say: 'Our relationship to our own body is fundamentally different from our relationship to others' bodies. Our relationship to actions that we ascribe to ourselves is fundamentally different from our relationship to others' actions. We are responsible in a unique way for actions that we ascribe to ourselves. Without a sense of responsibility for our own agency, it would be hard to see how we could think of ourselves as individual agents, who deliberate, decide, act and interact with agents who are clearly separate from ourselves. This is why our own agency carries special moral weight. Not stopping a harmful sequence that you have initiated yourself is therefore harder to justify than not stopping a harmful sequence that you have not initiated.'[104]

[104] This reasoning can be backed up with considerations of autonomy or self-ownership. Woollard seems to suggest that it is more objectionable to let oneself do harm than to merely allow harm, because one is (part of) the threat when letting oneself do harm but not when merely allowing harm: 'I must make greater sacrifices to save someone if it is my body or my past behaviour that has put her in jeopardy' (2014, 275). A strong duty to aid seems much easier to justify when the agent herself is part of the threat. Her body is not imposed upon in the same way and her self-ownership rights are not infringed upon in the same way compared to cases in which the agent is not part of the threat.

To be clear, I do not suggest that sequences of acts should never be evaluated as a whole. How to individuate different behaviours that form parts of sequences is a complicated topic, which I do not discuss here. However, finding an account of how to individuate such behaviours is a yet-to-be-solved challenge for all theories that operate with causal sequences, and is therefore not a challenge specifically for my argument.

The thought that responsibility for our own agency is necessary for us to see ourselves as individual agents is inspired by Susan Wolf's (2001, 12–13) explanation for responsibility for harm arising from bad moral luck. We can back up this reasoning with considerations of autonomy or self-ownership. Woollard (2014, 275) suggests that it is more objectionable to let oneself do harm, rather than just allow it, because one is (part of) the threat in the former case but not in the latter. A duty to aid seems to demand less when the agent herself is part of the threat. The agent's body is not imposed upon in the same way and her self-ownership rights are not infringed upon in the same way as when she is not causally connected to the harm in question.

However, more needs to be said to fully explain and justify the self-other divide. Defenders of the self-other divide need to explain why our own agency should matter enough to override considerations about the overall consequences of actions. In other words, why should our own agency matter *that much*?

The question is therefore not only why responsibility for our own agency should matter morally, but also why it should count for as much as our intuitions seem to suggest. This becomes especially apparent when considering cases under uncertainty such as

> (Potential Poisoner) As in (Poisoner), but this time, Agent is not sure whether the white stuff she put in the tea earlier was arsenic or sugar. She can now either save five strangers, who have ingested arsenic, or hurry back to pour away the tea and thus save the five.

Intuitively, Agent is at least permitted, if not required, to let the five strangers die to ensure that she will not have poisoned the five. This seems to be true whether Agent is 95%, 85%, or even only 75% certain that she has actually put poison in the tea. The self-other divide can vindicate such intuitions.

It is hard to see how such special concern with one's own actions can be underpinned morally. This is not only because the importance of one's own agency does not clearly outweigh overall consequences. More fundamentally, the motivation for the intuition behind the self-other divide can be called into question: One might say that our intuitions underpinning the self-other divide stem from an ultimately indefensible desire to keep our own hands clean.[105]

The challenge for defenders of the self-other divide, then, is to provide a plausible justification that is sufficiently strong to explain our intuitions in the Poisoner cases. In the following, I will sketch one way of providing such a justification, based on the idea that we have special obligations towards those whom we have wronged in the past, or will wrong in the future, by imposing a risk of harm on them.[106]

This justification of the self-other divide is based on a line of reasoning that is discussed, but ultimately rejected, by J. Hanna (2015a, 685–86). J. Hanna suggests that defenders of the self-other divide could appeal to considerations from restorative justice. It is plausible that we can wrong others by imposing threats of harm on them, even if the harm will only materialize at some point in the future. By putting the poison in the teapot, then, Agent wrongs the five. Such wronging, however, gives rise to special obligations. J. Hanna points out that it is generally thought that we have special obligations to 'mitigate, prevent, or offset the harms associated with one's wrongful behavior'. Obligations to aid our past victims might just be a particular instance of these broad special obligations (2015a, 686).[107] This explains why our own agency matters morally. We ought to be more concerned with our own wrongful behavior, because our own behavior gives rise to special obligations towards those whom we have wronged. Moreover, these special obligations seem to have the required strength: they require agents to compensate their own victims over aiding strangers, but do not permit killing innocents in order to prevent or mitigate harm from one's potential victims (2015a, 686).

J. Hanna argues that the argument from restorative justice cannot account for cases of future harmdoing. He gives the example of an agent who has a condition that, if untreated, will cause violent sleepwalking. The drugs for curing the condition are expensive, and the money could prevent much more harm when donated than the agent would prevent by curing herself from the condition. Intuitively, the agent nonetheless ought to buy the drugs, to prevent herself from unwitting violent behavior in the future. But considerations from restorative justice seem irrelevant, as the agent has not yet wronged anyone (J. Hanna 2015a, 687–88).

[105] This concern is also raised by Persson, who thinks there might be a worry of 'repulsive moral self-indulgence […] why be especially concerned about your own right-violations rather than the right-violations of all people, in proportion to the stringency of the rights violated?' (2013, 110).

[106] Thanks to an anonymous reviewer for pressing me on this point.

[107] An argument along these lines is given by McMahan (2002, 367). McMahan suggests that a guest at a pool party who accidentally knocks another guest into the water 'has an additional, special reason to save him [the drowning man]. This is that, if the drowning man is not saved, the responsible man will not only have failed to save him but will also have harmed him, or caused him to die' (McMahan 2002, 367).

However, I suggest that once we accept that we have special obligations towards those whom we have wronged in the past, we should also accept that we have special obligations towards those whom we will foreseeably wrong in the future. To illustrate, imagine an agent who is prone to drunk driving. When going out for a drink, this agent ought to take precautions to prevent herself from doing harm in the future, for example, by pre-ordering a cab. She ought to take these precautions, even if doing so is expensive (for example, if the ride home costs more than saving a life by donating to an effective charity). A natural explanation is that if the agent does not take such precautions, she will foreseeably wrong those endangered by her future behavior. Similarly, the violent sleepwalker ought to buy the drug, to prevent herself from wronging others by imposing a risk of harm on them.

One might object that we cannot have direct obligations towards those whom we have not wronged yet. However, it strikes me as plausible that we can have such obligations in situations in which we can foresee that we will wrong others in the future. For example, there does not seem to be a morally relevant difference between a case in which an agent has just poisoned a drink, and a case in which the agent knows that she will do so in five minutes (perhaps because of some compulsion), if she does not pour the drink away now. It seems that the agent in both cases has special obligations towards the drink's owner to pour the drink away.

However, my argument does not depend on this claim. We might still have indirect obligations to prevent ourselves from wronging others, as we would then incur direct special obligations. For example, even if the violent sleepwalker does not have direct obligations towards her potential future victims, she might have indirect obligations to prevent a situation in which she incurs special obligations by wronging her potential future victims. (Failing to donate would not ground special obligations towards those she has failed to aid.)

Another objection is that wrongdoing cannot give rise to special obligations if the wrongdoing was blameless, because agents are not morally responsible in those cases. I can see two ways in which the defender of the self-other divide might respond here.

One response would be to concede that Agent has no, or weaker, special obligations towards those she has wronged blamelessly. However, this seems to contradict intuitions about such cases. Assume that Agent was in no way at fault when putting arsenic in the sugar jar (one of the victims labelled the jar incorrectly). It still seems that as soon as Agent learns this fact, she should drive back to save the tea pot victims, rather than saving the strangers.

These remarks suggest a second response: The self-other divide strengthens reasons against letting ourselves do harm, regardless of whether we are blameworthy for our doing. Bernard

Williams has famously described that we often ascribe some responsibility to agents who, through no fault of their own, commit a wrong (such as the lorry driver who, despite taking all precautions, runs over a child) (Williams 1981, 28). We might think that a similar story can be told about the blameless poisoner. On this view, like the lorry driver, the unwitting poisoner should take responsibility for her action, and thereby incur special obligations towards the victims. (Of course, her blamelessness will likely change the overall moral evaluation of her behavior.)

A final problem concerns where we should draw the line between our own agency and external factors. Persson (2013, 95–96) describes a case in which Agent can feel the onset of a spasm, which, if it runs its course, will cause Agent to pull the trigger of a gun. According to Persson, by failing to suppress the spasm, Agent lets herself do harm. He argues, further, that it seems hard to justify why we should be more responsible for allowing harm that results from our involuntary bodily movements than for allowing harm that results from external factors: 'The fact that a twitch is internal rather than external to us cannot make any moral difference' (Persson 2013, 105).

In response, deontologists might question whether involuntary movements like spasms are things that we 'do' in the relevant sense. Bennett (1995, 110) seems to take this line, while noting the peculiarity of the spasm example, which he describes as a case of allowing harm in which the agent's behavior is positively relevant to the upshot. Alternatively, deontologists might agree with Persson that involuntary movements are things that we do but insist that this wide sense of doing is morally significant. For example, Woollard argues that agents have a special relationship with their own body, which grounds special responsibility for what our bodies do (2013, 331; see also 2015, 192).

It is beyond the scope of this thesis to settle this matter. However, doing so is not necessary for my purposes. I am interested in the moral status of cases of letting oneself do harm. Whether involuntarily movements can constitute doing harm (and, therefore, whether allowing such movements can constitute letting oneself do harm) is a further question which does not pose a *special* challenge to the discussion of the moral status of cases of letting oneself do harm.

To sum up, I suggested that agents should be especially concerned with their own wrongdoing in virtue of the special obligations that they have towards those whom they have wronged in the past or know they will wrong in the future. These special obligations give agents moral reasons to prevent harm resulting from their own past or future doings. Absent any confounding reasons, these moral reasons are weaker than the reason against doing harm (agents ought not to kill in order to prevent themselves from killing), but stronger than the reason to prevent harm that is unrelated to the agent (agents ought to save their own victims over others).

### 4.5.3 Letting Oneself Do Harm as “Dallowing”

For the sake of completeness, let me briefly discuss a final option. On the face of it, letting oneself do harm does not seem to be an instance of either doing or allowing harm. If we only consider Agent’s present action, then letting oneself do harm is clearly an instance of *allowing* harm. However, if we consider an agent’s behaviour over time, then letting oneself do harm is clearly (at least part of) a harm *doing*.

With that in mind, let us turn to a third option. One might be tempted to classify cases of letting oneself do harm as belonging to a third category. This is what one might say: “Letting oneself do harm is in relevant ways like merely allowing harm, even if they are not equivalent. Letting oneself do harm is in relevant ways like doing harm, even if they are not equivalent. Letting oneself do harm shares morally relevant features with both doing and allowing harm, and thus the moral status of letting oneself do harm lies in between those.” These cases would then form a distinct moral category by themselves.

If this reasoning is correct, then cases of letting yourself do harm constitute a distinct set of cases whose moral status is somewhere in between doing and allowing harm. For lack of a better word, we might say that letting oneself do harm would then correctly be described as an instance of ‘dallowing’.

However, the defender of the DDA would need to say more about this account. What does it mean to say that a ‘dallowing’ behaviour has an ‘in between’ moral status? Is dallowing harm much harder to justify than merely allowing harm, or just a little bit harder? Is it much easier to justify than doing harm, or just a little bit easier? And how can the answer to these questions be defended? It seems that this option opens more questions than it solves.

Moreover, I think that deontologists have principled reason to resist the temptation to introduce new categories of cases alongside doings and allowings. Limiting the scope of cases that can be classified as either doing and allowing harm also limits the explanatory power of the doing/allowing distinction. It seems preferable to be able to classify cases of letting oneself do harm as either doing harm or allowing harm and spelling out the ways in which they differ from paradigm cases of doing and allowing harm, rather than opening new categories, and with it the need for new and expanded conceptual and moral frameworks.

## 4.6 Conclusion

With powerful technologies and scientific methods to measure and predict our impact on the environment and future life on earth, our influence on future generations is greater than at any other point in human history. Developments such as climate change present us with unique challenges for moral decision making.

Most deontologists maintain that moral decision making should be informed by distinctions such as the DDA. In this chapter, I have argued that the challenge faced by defenders of the DDA regarding complicated cases, such as cases of letting oneself do harm, is more urgent, but also more complicated than has been acknowledged so far. This is because real-world cases are very complex, and because the applicability of the DDA to such complex cases is unclear. I have discussed different ways in which defenders of the DDA could respond to this challenge, and tentatively proposed to classify them as instances of allowing harm. If my argument is correct, then deontologists will need a more fine-grained account of causal sequences and their moral relevance to account for the moral status of complicated cases, and ultimately real-world decision-making.

# Chapter 5 Doing Harm and Offsetting Risks

## 5.1 Introduction

Recall that the Doing Thesis is the claim that the reason against doing harm is stronger than the reason against merely allowing (i.e. failing to prevent) equivalent harm. The Doing Thesis partly constitutes the *constraint against doing harm*. In this chapter, I will understand the constraint against doing harm as synonymous with the Doing Thesis.[108]

As it stands, the constraint against doing harm is not applicable to cases in which agents do not know whether their action will lead to harm (and therefore, whether their action will constitute doing harm). This is unfortunate, because in the real world, agents are almost always uncertain[109] whether their behaviour will lead to harm. If defenders of the constraint against doing harm want the principle to guide action in the real world, they need to be able to apply it to cases under risk.

In this chapter, I explore how the constraint against doing harm applies to risky cases. I first propose a generalised version of the constraint against doing harm that is applicable to cases under risk. I then discuss what I call "Offsetting Risk Cases", cases in which agents offset the risks that they impose on others. I argue that an analysis of Offsetting Risk Cases sheds light on a debate about the nature of the constraint against doing harm. Some philosophers have suggested that the constraint against doing harm should be interpreted as what I call time-relative, whereas

[108] To be more precise, the constraint against doing harm also contains the Harm-Benefit-Asymmetry, the claim that the reason against doing harm is stronger than the reason for doing equivalent good (MacAskill and Mogensen 2019). Understanding the constraint against doing harm as equivalent to the Doing Thesis is adequate for the purpose of this chapter, since all cases I discuss in this chapter are cases in which agents do (or risk doing) harm, rather than cases of allowing harm or doing good. Moreover, this usage is compatible with common characterizations of the constraint against doing harm. Kamm (2007, 14) writes about constraints: 'Constraints limit what we may do in pursuit of our own, or even the impartial, good. The most commonly proposed constraints are a *strong duty not to harm (contrasted with a weaker duty to aid)* and/or a strong duty not to intend harm' (my emphasis). I understand the duty not to do harm as expressing pro tanto *reasons* against doing harm. This understanding follows Stratton-Lake's interpretation of Ross's prima facie duties: 'For some feature of an act to be prima facie wrong is not for us to have some kind of duty not to do it, but is for it to have some feature that gives us a *moral reason not to do it*. These features are not what we should do (our duty), but are considerations which we (should) take into account in deciding what it is we should do' (Ross 2002, xxxiv my emphasis).

[109] I use 'uncertainty' and 'risk' synonymously, in a non-technical sense, to encompass cases in which agents are unsure about the likelihood of possible outcomes of their behaviour. In a more technical sense, decisions are said to be made under 'risk' when the relevant objective probabilities are known, and under 'uncertainty' when they are not known (for these distinctions and discussion of the notions of risk and uncertainty, see Hansson (2013, chap. 1)).

others have argued that the constraint should be interpreted as what I call time-neutral.[110] The time-relative view says that agents should not do harm now. The time-neutral view says that agents should minimize their own harm doings. However, I show that both time-relativists and time-neutralists have difficulties accounting for Offsetting Risk Cases. Finally, I sketch a possible alternative. According to the Relation-Centred View, for every possible victim V, and every harm (in some respect / at some time) H, agents have a reason against doing something that might constitute doing H to V. The relation-centred view gives the right result in offsetting cases. Moreover, it can explain intuitions that motivate its competitors.

Before we start, two reminders on the scope of my argument are in order. First, regarding cases in which agents risk doing harm, my primary concern is (the strength of) *harm-based* reasons against these risky actions, and not whether the risky actions are permissible, all things considered. Second, since I am concerned with action-guidance for real-world agents, the reasons that I will be concerned with in the following are subjective (i.e. belief- or evidence-based) reasons. Since agents do not know whether their behaviour will result in harm, they do not have access to objective (i.e. fact-based) reasons. Unless otherwise noted, I will use terms such as 'moral reasons', 'permission' or 'obligation' in a subjective sense.[111]

## 5.2 Extending the DDA to Risky Cases

The constraint against doing harm can explain the intuitive difference between the following cases:

> (Push) Agent pushes a boulder which rolls over Victim, crushing her to death.
>
> (Non-Intervention) A boulder is rolling towards Victim. Agent could stop the boulder but does not do so. The boulder rolls over Victim, crushing her to death. (These cases are variants of cases discussed in (Woollard 2015; 2012b; Bennett 1993, 76–77).)

[110] The time-relative and the time-neutral view are both versions of an agent-relative understanding of the constraint against doing harm, in the sense that it prohibits the agent's performing some act, even if the agent thereby prevents multiple comparable act-types performed by others. See e.g. (McNaughton and Rawling 2007, 437) for an illustration of agent-relative reasoning that motivates the constraint against doing harm: 'Suppose I can only prevent you killing two innocents by killing one myself. […] I have overriding moral reason (a distinct moral aim) not to kill anyone myself (as you should aim not to kill anyone yourself). Thus, although you will do wrong in killing the two, I should not kill the one in order to prevent you.' I will briefly discuss an alternative victim-relative view of the constraint against doing harm in section 5.5.

[111] For an overview of the distinction between subjective and objective rightness, and a thorough discussion of the concept of subjective rightness, see e.g. Smith (2010).

In these cases, it is implicitly assumed that pushing the boulder (or failing to stop it) is certain to lead to harm. The assumption of certainty is very common in examples used in the literature on the distinction between doing and allowing harm.[112] We can illustrate what relaxing this assumption would look like with the following pair of cases:

> (Risky Push) Agent pushes a boulder down a hill. The probability that the boulder will roll down the left slope, thus rolling over Victim and crushing her to death, is p. The probability that the boulder will roll down the right slope, which is empty, is 1-p.
>
> (Risky Non-Intervention) A boulder is rolling down a hill. The probability that the boulder will roll down the left slope, thus rolling over Victim and crushing her to death, is p, and the probability that the boulder will roll down the right slope, which is empty, is 1-p. Agent could stop the boulder now but does not do so.

Importantly, cases under certainty, such as (Push) and (Non-Intervention), are not of a different kind from cases under risk, such as (Risky Push) and (Risky Non-Intervention). Rather, the certain cases are specific instances of the risky cases: namely those instances in which p=1. Moral theories should have plausible implications for cases under uncertainty, and not merely for specific instances represented by cases under certainty. After all, real-world cases are cases under uncertainty, and action-guiding moral theories need to be able to accommodate these cases.

Intuitively, for any value of p, the reason against an act that might constitute doing harm is stronger than the reason against an act that might constitute merely allowing harm. Imposing a risk of harm seems morally worse than failing to prevent a comparable risk of harm. The following principle seems therefore plausible:

> (Risky Constraint Against Doing Harm) The reason against φ-ing is stronger than the reason against ψ-ing, if φ-ing is p likely to constitute doing harm and ψ-ing is p likely to constitute merely allowing harm.

The Risky Constraint (as I will call it for short) is extremely plausible. It explains why (Risky Push) is harder to justify than (Risky Non-Intervention). It equals the constraint against doing harm for p=1, and it has plausible implications for other values of p.

[112] Examples are (Push) and (Non-Interpose) in (Woollard 2015, 3), the case of killing a man vs. letting a man die to help others in (Foot 1967, 4), Rescue I and II in (Foot 2002, 81), Rescue III and IV (Quinn 1989a, 298–99). There are good methodological reasons for the assumption of certainty in arguments that are aimed to motivate (or reject) the moral relevance of the distinction between doing and allowing harm *in principle*. To keep the focus on this distinction, cases should not be unnecessarily complex, hence the useful simplifying assumption of absolute certainty about outcomes.

The Risky Constraint is very minimal. It does not tell us how the likelihood or severity of harm influences the reason against doing harm. Perhaps the most straightforward option to spell this out might be that the strength of the reason against doing harm is proportional to the expected harm, where the expected harm is the product of the likelihood of harm and the severity of harm. The advantage of the Risky Constraint, then, is that it is compatible with a wide range of views about the strength of the reason against doing harm. (Like the Constraint Against Doing Harm under certainty, the Risky Constraint also does not tell us *how much* stronger the reason against doing harm is than the reason against merely allowing harm.)

In the next few paragraphs, I explain why the Risky Constraint is a better version of the constraint against doing harm under risk than two potential competitors: an absolutist threshold view and Kagan's moderate sliding threshold view.

Much of the literature on deontology and risk focuses on *absolute* deontological constraints. Constraints are absolute when their violation is always impermissible. Absolute constraints face difficulties in risky cases (and in other cases, as I explain below). Most behaviours carry some risk of violating a constraint, but few are certain to do so. To avoid implausible implications, absolute deontologists should neither prohibit all risky behaviour (such as driving a car) nor permit all behaviour that is not certain to violate constraints (such as reckless driving). One way to avoid such implications is to adopt the threshold view. According to the threshold view, behaviour is subject to constraints only when the likelihood that the behaviour violates that constraint is sufficiently high. However, the threshold view has been met with criticism in the literature, mainly for being arbitrary[113] and vulnerable to counterexamples.[114]

Contrary to what the literature on deontological constraints seems to suggest, many contemporary deontologists are *moderate* deontologists. They believe that constraints express pro tanto reasons, which can in principle be outweighed. For example, moderate deontologists might think that the pro tanto reason against killing one person is strong enough to outweigh the reason for saving five lives, but that the reason against killing one person is not strong enough to outweigh the reason for saving a million lives. Killing an innocent person might therefore be impermissible if doing so would save five lives, but it would be permissible to kill one innocent

[113] This point is made, for example, by Kagan (2018, 83). However, see recent defences of the threshold view that appeal to stochastic dominance (Tarsney 2018), degrees of knowledge (Isaacs 2014), or a formal extension of how absolutists treat cases under certainty (Black 2020).

[114] Jackson and Smith (2006) have pressed this criticism with regard to cases involving so-called 'ought agglomeration'. The problem here is that thresholds can permit acts that are performed separately but prohibit them when considered in conjunction. See Aboodi, Borer, and Enoch (2008) for a reply.

person in order to save a million lives. Combined with such a moderate understanding of deontological constraints, the threshold view seems unmotivated.

The threshold view derives its plausibility from the idea that behaviour that is very unlikely to constitute doing harm should be morally permissible, and behaviour that is very likely to constitute doing harm should be morally impermissible. But the moderate deontologist can accept this idea. The moderate deontologist can point out that the Risky Constraint against doing harm relates the strength of the reason against a behaviour to the likelihood that the behaviour constitutes doing harm. The reason against pushing the boulder when p=0.001 is stronger than the reason to aid when p=0.001. However, the reason against pushing the boulder when p=0.001 might not be stronger than the reason to aid when p=1. In other words, it might be permissible to push the boulder, if pushing the boulder would certainly save a life and has a very low likelihood of harming Victim.

Furthermore, the threshold view is insensitive to increases or decreases in risk of harm above or below the threshold. This is especially problematic given that in many real-life cases, what is problematic about an agent's conduct is that they increase the risk of harm. Consider

> (Pedestrian) Pedestrian is at risk of being hit by a car (perhaps the roads are slippery, or poorly designed). Driver decides to speed up when approaching her.

It seems that Driver acts wrongly. However, Driver does not impose a previously non-existent risk of harm, since Pedestrian was already under a nonzero risk of being injured. Rather, the problem seems to be that Driver *increases* the risk of injury when deciding to speed up. The threshold view cannot explain why Driver has a reason against speeding up, if Pedestrian's risk was already above the threshold. This is implausible.

Another possible version of extending the constraint against doing harm in a way that makes it applicable is Kagan's 'sliding threshold' view. Kagan (1989, 89) suggests 'that the constraint against doing harm must be construed as having a sliding threshold, a threshold which diminishes with the decrease in probability of harm'. On Kagan's view, it is permissible to risk violating the constraint against doing harm in order to achieve a certain good when the probability of harm is sufficiently low.

I believe that the sliding threshold view is similar in spirit to the Risky Constraint against doing harm. However, it is a disadvantage of the sliding threshold view that it is formulated in terms of overall permissibility. On Kagan's view, the likelihood of harm *directly* influences the permissibility of action that might constitute harm doing, since the permissibility threshold is a function of the probability of harm.

Portmore (2017, 108) argues that Kagan's view implies that the number of people who are at risk of harm does not make a difference as to whether that action is permissible. If an action risks harming someone, and the likelihood is not sufficient to make the action impermissible, then if the action risks harming many people with the same likelihood, this is also not sufficient to make the action impermissible.

The Risky Constraint avoids this problem. The Risky Constraint can be combined with the plausible view that the probability of harm only *indirectly* influences the permissibility of action. For example, one might think that expected harm (i.e. the sum of the products of the probabilities and magnitudes of possible harms) is a function of the probability of harm. If more people are at risk of harm, then this changes the expected outcome. If more people are at risk, this also changes the severity of harm: not the severity of harm for any of the potential victims, but the severity of overall harm that the agent will have done through the action in question.

To repeat, it is an advantage of the Risky Constraint that it comes with minimal commitments. Its proponents are not committed to an account of how much the likelihood of harm strengthens the reason against doing harm. The relation need not be linear, or even multiplicative.[115] Moreover, the Risky Constraint against doing harm is compatible with a range of views about other factors that might influence the reason against doing harm, including the severity of harm, the number of victims, or even factors such as the agent's intention. The Risky Constraint against doing harm, therefore, should be understood as a minimal claim that is in principle compatible with different views on deontological risk ethics.[116]

## 5.3 Time-Relativity vs. Time-Neutrality

Recall that the constraint against doing harm (under certainty) says that the reason against doing harm is stronger than the reason against merely allowing harm. The constraint against doing harm explains why it is impermissible for A to kill B, even if doing so would prevent C and D from being

[115] One possible alternative view is that the reason against doing harm might increase in steps: for example, it might be strongest for harm that is "very likely", less for harm that is "likely", and even less for harm that is "unlikely". Yet another view might say that the relation looks different for different kinds of harm.

[116] The probabilistic view is compatible with existing accounts of moderate deontology under risk. For example, Lazar (2017b; 2018) develops a comprehensive (moderate) deontological decision theory. However, not all deontologists are happy to freely make use of the decision theoretic toolbox to tackle decision making under risk. Tenenbaum (2017), for example, argues against what he calls the 'multiplicative model' of multiplying some moral value with a probability to calculate its moral weight under risk.

killed. It is impermissible for A to kill B, because the reason against killing B is stronger than the reason against merely allowing C and D to be killed.

Here is a problem case for the constraint against doing harm. What if the lives of C and D are in danger only because of something that A has done previously? In this case, A's failure to kill B would make A the killer of C and D.[117]

Here are two ways to interpret the constraint against doing harm that might help deontologists to account for the problem case:[118]

First, one can say that the constraint against doing harm is not just agent-relative, but also *time-relative*.[119] It only applies to the present choice between either killing B or saving the lives of C and D. What matters is that Agent now makes a choice to kill: The history of C's and D's peril is irrelevant, as far as the constraint against doing harm is concerned.[120]

Second, one can say that the constraint against doing harm is *time-neutral*. It tells agents to minimize their harm doings across choice situations. What matters is the agent's overall balance of harm doings. This seems to imply that A should kill B (since A will have killed only one person rather than two).[121]

The time-relative / time-neutral distinction also applies to the Risky Constraint. In its time-relative interpretation, the Risky Constraint says that agents have a reason against doing something in the present choice situation that might constitute doing harm. In its time-neutral interpretation, the

[117] I have discussed such cases of 'letting oneself do harm' in Chapter 4. My focus then was on the moral status of the action that constitutes letting oneself do harm. My focus here is slightly different: I look at different interpretations of the constraint against doing harm and their implications for cases under risk. (However, these issues are related. Plausibly, the view that letting oneself do harm constitutes doing harm fits well with the time-neutral view. The view that letting oneself do harm constitutes allowing harm fits well with the time-relative view.)

[118] Both interpretations are agent-relative. I will briefly discuss a victim-relative interpretation (e.g. Kamm (2007), but see e.g. Portmore (2011, 100–103) for a reply defending agent-centred consequentialist constraints) in section 5.5.

[119] See, for example Tarsney (2018, 516), Colyvan, Cox, and Steele (2010, 522–23), Ridge (2017, sec. 6). As I have discussed earlier, J. Hanna (2015a) suggests that deontologists might argue that 'temporal proximity' matters morally; however, he raises some worries regarding this claim (compare also Thomson's (1976) observation that 'the present tense matters'). In response to Persson (2013), Woollard (2014) suggests, but does not develop the view that '[t]he primacy of present agency seems to rest on a peculiar relationship to one's current actions'.

[120] I follow the literature in using the term "time-relative". However, as Tarsney (2018, 517) correctly notes, this term is misleading. The constraint against doing harm is, strictly speaking, relative to a particular *choice situation* rather than the moment in time at which the choice is made. This is because if we make multiple choices at the same time, some harmful and some beneficial, the constraint against doing harm would still give the agent a reason against such choices.

[121] I will discuss in section 5.4.1 a response on behalf of the time-neutralist, namely that A should not kill B for reasons beyond the constraint against doing harm, such as that A should not use B as a means.

Risky Constraint says that agents have a reason for doing something that might minimize their own harm doings.

## 5.4 Offsetting Risk Cases

Consider an agent who performs a sequence of behaviours. One of these behaviours increases the risk that the agent does harm. The other behaviour decreases the risk that the agent does harm. I assume that together, the two behaviours neither increase nor decrease the risk that the agent does harm. I will say that the second behaviour *offsets* the risk of harm that the agent imposes through the first behaviour. In the following, I will refer to cases in which agents perform a sequence of behaviours in which they offset their own risk impositions as 'Offsetting Risk Cases'.[122]

Offsetting Risk Cases are different from cases in which agents offset actual harm. For an example of offsetting actual harm, consider a case given by Foerster (2019, 617), in which a meat eater offsets a harm (their consumption of meat) with an equivalent benefit (donating some amount to an animal welfare charity) such that he 'ends up morally neutral relative to where he started' (2019, 618). In these cases, the likelihood *that the agent does harm* is not decreased by the offsetting behaviour. The agent does not offset their behaviour by decreasing the risk that they do harm, but rather by providing benefits. Perhaps one can illustrate the difference between these cases by saying that when agents offset actual harm, they try to remain neutral on their "overall moral balance", whereas when agents offset risks of doing harm, they try to remain neutral on their "harmdoing balance".

In the remainder of this section, I argue that the implications of both the time-relative and the time-neutral view in Offsetting Risk Cases are implausible. Before I do so, let me highlight two important restrictions of my argument. First, I only consider cases in which agents offset the risk of harm that *they themselves* impose on others. This is because I am interested in the constraint against doing harm, and this constraint does not (at least, not generally) give agents a reason to minimize risks of harm imposed by *others*.[123] Second, I explore the question whether the

[122] I rely on an intuitive notion of 'behaviour', whereby I have in mind actions and omissions that are under the agent's control and ignore the further question of how we should individuate behaviours (and when we should characterize behaviours as being part of a sequence). I take my account to be compatible with plausible accounts of behaviour.

[123] There might be exceptions to this. For example, offsetting one's carbon emissions from flying by paying for reforestation projects might be motivated from a concern to minimize the overall harmful effects of one's

constraint against doing harm gives agents a reason against increasing risks if they offset these same risks. The answer to this question will inform moral judgement about Offsetting Risk Cases. However, it will not settle the overall moral status of the agent's behaviour. I do not take a stand on whether increasing and offsetting risk of harm is on the balance of reasons impermissible, permissible, or even morally required.[124]

### 5.4.1 Against the Time-Neutral View

> (Making Good App) An app on Clara's phone tracks all risks of harm that arise from her actions. Whenever a behaviour (e.g. getting in the car at the end of a boozy night or attempting to puncture her ex-boyfriend's tyres) threatens to increase her score to a level that she deems inacceptable, she presses the 'Make Up for It' button. The app then calculates her risk of doing harm arising from this behaviour (e.g. the likelihood she will run over a pedestrian). The app then adjusts Clara's other actions in a way that reduces the overall risk of harm that arises from her actions back to the original level (e.g. by buying more fair-trade products, or by donating more to charity).

According to time-neutralists, when Clara chooses to impose a risk and offset it via the app, her behaviour is not subject to the constraint against doing harm. After all, using the app ensures that none of her decisions increase the overall risk that she does harm. According to the time-neutral view, this is all that matters. It is irrelevant whether the agent increases this overall risk directly in the immediate choice situation, or merely indirectly by influencing likely outcomes of their past or future choices.

However, this is an implausible result. Clara's behaviour will likely cause harm to victims that they would not have suffered otherwise. Her behaviour thereby constitutes harming according to the two main accounts of harming in the literature, the counterfactual comparative and the causal

action. While it does not matter for my overall argument, I am not sure that such offsetting can fulfil the requirements of the constraint against doing harm. Such a case might be better characterized as a case in which the agent increases the risk of harm and prevents someone else from increasing the risk of the same harm. Preventing someone else from increasing the risk of harm is different from decreasing the risk of harm arising from one's own actions. Thanks to Garrett Cullity for raising this point.

[124] These questions have been addressed in the literature on climate justice. See e.g. Broome (2012) for the claim that emitting carbon dioxide and then offsetting is morally on a par with not emitting, and Cullity (2019) for an argument that it is morally wrong to not offset one's emissions.

account (and my own account developed in this thesis).[125] It therefore seems to be a paradigm case of harming behaviour. Surely, the constraint against doing harm should apply to such behaviour.

Defenders of the deontological constraint against doing (as opposed to merely failing to prevent) harm have offered different justifications for the constraint. While I do not have the space for an in-depth discussion here, there is reason to believe that at least some such justifications apply to Clara's behaviour. For example, Fiona Woollard argues that doing harm constitutes an imposition on the victim (2015; 2013; see also F. Kamm 1996). Such an imposition 'involves a sequence of positive facts leading from the agent to a harmful effect on the victim' (2012c, 464; see also Foot 1967). Building on Quinn (1989a), Woollard suggests that if such imposition were permissible, the victim's body would not genuinely belong to her. The constraint against doing harm prevents others' behaviour from 'affecting what belongs to you against your will' (Woollard 2012c, 464).

While Woollard does not explicitly address risky cases, her argument can be seen to apply to these cases in a similar fashion. Increasing the risk of harm makes it more likely to impose on someone in a harmful way. For example, Clara's behaviour makes it more likely that the pedestrian gets injured. If behaviour that affects others in a harmful way ought to be prevented, then surely this gives us a reason to prevent behaviour that puts others at risk of being affected in a harmful way. I conclude that considerations regarding the justification and nature of the constraint against doing harm support the intuitive judgement in (Making Good App).

Defenders of the time-neutral view might object to this argument in the following way. They might argue that what makes Clara's behaviour wrong is not the constraint against doing harm. Rather, it is her bad intention to impose risks of harm.[126] If what matters is the agent's mindset at the time of the imposition of risk, then Clara's behaviour can be criticized on the basis that she wrongfully intended to increase the risk of harm for some agents.

However, this approach comes with the major disadvantage that it cannot explain why agents have a reason against doing harm, but not against merely allowing harm, even when the harm is not intended. Killing one person as a merely foreseen side-effect of an action seems harder to

[125] I have discussed these accounts in Chapter 3. Roughly, according to the counterfactual comparative account, A's behaviour harms B only if B *would have been better off,* had A not performed the behaviour (accounts in this spirit are held e.g. by Hanser 2008; Boonin 2008; Klocksiem 2012). According to the causal account, A's behaviour harms B only if A *causes* B to be badly off (accounts in this spirit are held e.g. by Harman 2004; 2009; Shiffrin 1999; Gardner 2015).

[126] A view along these lines is suggested by McNaughton and Rawling as a response to cases of letting yourself do harm: 'On this approach, the wrong doing is the initiation of a causal chain with the intention of placing someone under threat – and this wrong-doing cannot be undone by later interference with said chain' (1993, 92).

justify than letting someone die as a merely foreseen side-effect of an action. This point is perhaps even more relevant in cases of risk imposition, since arguably many real-world cases involving negligence are cases in which agents unintentionally impose foreseeable risks of harm.

Another response on behalf of defenders of the time-neutral view might be the argument that Clara's behaviour is wrong because she treats potential victims of her actions as mere means.[127] In the case that I sketched in the introduction, in which A can kill B to save her previous victims C and D, such an argument would imply that what A does is wrong, because A's behaviour increases the number of people that A treats as mere means (A has already treated C and D as mere means by imposing a risk of harm, and now treats B as mere means in order to offset risks of harm). Portmore (2011, 101–2), for example, responds to a case given by Kamm (2007, 279), in which an agent can now murder one person to save two others that the agent has previously placed under a fatal threat, as follows. He suggests that what the agent 'should be most concerned to minimize is not the number of people that she actually murders, but the number of people that she treats as a mere means by attempting to murder them […] [treating the one person] as a means to minimizing the number of people that she murders only adds to the number of people that she treats as a mere means' (2011, 101–2).[128]

Here are two objections to this response. Consider, again, a case in which A can harm B to save C and D, who will otherwise suffer equivalent harm. (A is not responsible for the threat to C and D.) The first objection is that A's action seems wrong because it increases the risk of harm to B, and not (only) because A uses B as a mere means. We can illustrate this point by comparing this case to the case of giving a false promise. Giving a false promise under normal circumstances arguably constitutes using the other person as a mere means. However, it seems that A would be permitted to give a false promise to B in order to save C and D from (significant) harm, while not being permitted to harm B in order to save C and D from harm.

To this objection, the time-neutralist might respond that using as a means is less important, morally speaking, than minimizing harmdoing. However, if this is so, then this seems to undermine the time-neutralists' claim that it is impermissible to use someone as a means in order to minimize harm.

For the second objection, assume that A *is* responsible for the threat to B and C. Harming B still seems wrong, even though the number of people that A has treated as mere means does not

[127] Seth Lazar has suggested this possibility in personal communication.

[128] Portmore's overall aim, which will not concern us here, is to argue that consequentialists can endorse non-consequentialist constraints (and account for them better than certain forms of non-consequentialism can).

increase. (To this objection, the time-neutralist might respond that what matters is not the number of people, but the number of instances in which agents are treated as mere means. However, agents can only treat someone as a means in the present choice situation, so in effect, this would be introducing a time-relative constraint on top of the time-neutral one. However, as I suggest in section 5.5, such a combined view would face numerous difficulties.)

### 5.4.2 Against the Time-Relative View

If we should reject the time-neutral view, it might seem as if we should accept the time-relative view instead. According to the time-relativist, the constraint against doing harm gives Clara a reason against her behaviour. Clara increases the risk of harm to pedestrians by getting in the car, and she increases the risk of harm to her ex-boyfriend by damaging the tyres, and that is what counts. The time-relative verdict is that the deontological constraint gives Clara a reason against her behaviour. This seems like the right result.

Unfortunately, the time-relative view faces its own problem cases. Consider

> (Pollution) The risk of harm to residents at the riverside is roughly proportional to the amount of toxic waste in the river. Anna owns a factory that periodically releases toxic waste in the river, in a manner that is neither illegal nor otherwise prohibited. Today, Anna has made a decision to release additional toxic waste in the river, but to also set up water purification facilities that filter an equivalent amount of toxic waste from the river.[129]

According to the time-relative view, the constraint against doing harm gives Anna a reason against her behaviour. This is because her choice of releasing the waste increases the risk of harm. (Recall that any other choices Anna makes, even if they are put into effect through the same behaviour, are irrelevant for the question whether the constraint against doing harm applies to this choice.)

However, this is implausible. Unlike in Clara's case, in (Pollution) nobody finds themselves at an increased risk of harm as a consequence of Anna's behaviour. The risk facing those living by the riverside remains the same at all times (stipulating that there is no time lag between releasing the

[129] Foerster's (2019) definition of moral offsetting excludes cases such as (Pollution). What he calls the 'Harm condition' demands that the harming behaviour is harmful even when performed in conjunction with the offsetting behaviour. He introduces the condition to exclude cases from counting as cases of moral offsetting that he thinks are 'obvious and uninteresting' (2019, 3), such as 'Soccer Practice: You drop your child off at soccer practice, and pick her up later' (2019, 2). I do not think that such cases are uninteresting. If you drop your child off at soccer practice, silently deliberating whether you should pick her up later, you subject her to a risk of not being picked up – arguably, a non-negligible risk of harm.

waste and filtering the water). Anna's behaviour constitutes neither causal nor counterfactual harming. This is because she does not increase the likelihood that someone will be harmed who would not have been harmed otherwise, and her action does not make it more likely that she will cause harm than it would have been had she refrained from acting. She also arguably does not impose on anyone. There is no causal sequence leading from Anna's behaviour to harmful effects on the residents. I have stipulated that the risk of harm to residents is proportional to the total amount of waste in the river. Anna's behaviour does not increase the total amount of waste in the river at any moment in time.[130] Given this, it seems clear that Anna's behaviour does not violate the residents' right to self-ownership, and therefore should not fall under the constraint against doing harm.

At this point, the time-relativist might try to argue that Anna's behaviour constitutes something like a 'pure risk imposition', which, as has been argued, can count as harm (see e.g. Oberdiek (2012)). Cases of pure risk imposition are cases in which a risk is imposed, but as it happens, no harm occurs. The risk does not materialize. However, such cases are relevantly different from cases such as (Pollution). In cases of pure risk imposition, the lucky agents do, in fact, increase the expected harm – but thankfully it does not materialize. In contrast, in cases such as (Pollution), the expected outcome does not change.

The time-relativist might bite the bullet and accept that the constraint against doing harm applies to Anna's behaviour. They might defend this by claiming that accepting that there is a moral reason against behaviour such as Anna's does not come at a high cost, especially when Anna's offsetting behaviour might make her less blameworthy for doing harm. However, biting this bullet might come at a higher cost than the time-relativist realises. It would mean accepting a constraint on Anna's behaviour that seems unjustified, as I will explain in the following.

If Anna's behaviour was subject to the constraint against doing harm, as the choice-relativist argues, this would impact on *Anna's* freedom to act as she chooses. It would give Anna a reason against making trade-offs between risks that she imposes on others. However, this seems unreasonable, and unnecessary. There seems to be no reason to prohibit people like Anna from making such trade-offs at their own discretion, as long as they do not in fact increase the risk of harming anyone. We can appeal to Quinn's notion of self-ownership to support this. He writes that 'the moral sense in which your mind or body is yours seems to be the same as that in which your life is yours. And if your life is yours then there must be decisions concerning it that are yours

[130] Alternatively, one might assume that there is a threshold of waste in the river at which the risk of harm becomes significant. But even then, Anna's behaviour does not increase the risk of harm overall, since she does not increase the amount of waste, and therefore does not cause the threshold to be reached.

to make' (1989a, 309). Restricting the scope of decisions that an agent can permissibly make, without good justification, is more than just a quirk of a moral principle. It makes such a moral principle overly demanding.

In this particular case, such demandingness is inappropriate. Nobody stands to benefit if agents refrain from making trade-offs in cases such as Pollution. However, the agents concerned have much to lose: the cost of abiding by such principles is to give up one's self-ownership to a degree. Prohibiting risk trade-offs in cases like Pollution, therefore, seems unjust.

I conclude that considerations regarding the justification and nature of the constraint against doing harm support the intuition that the constraint against doing harm does not give Anna a reason to refrain from acting.

I will now consider possible objections. The time-relativist might argue that when an agent A performs two acts (behaviours) at the same time, these acts should be considered together for the purpose of the constraint.[131] However, what matters here is when the acts take effect, not when they are performed. Imagine that A has performed one of the acts two weeks ago but knowing that this act would have no effect on anyone's risk of harm until now, when they offset this risk. Such offsetting still seems permissible.

Suppose the time-relativist says that cases such as (Pollution) are far-fetched and unrealistic. We do not come across them in real life. So even if time-relativists should worry about such cases, perhaps they do not need to worry too much. However, we do not have to appeal to far-fetched cases to find (Pollution)-style scenarios in the real world. For an example, consider agents doing carbon offsetting to avoid increasing their carbon footprint, and thus contributing to carbon emissions that cause harm to both present and future generations.

The time-relativist might press the objection further and argue that we cannot, empirically, figure out such risks. However, even if that were true, it would not count against my general point that such trade-offs can prevent the constraint against doing harm from applying in principle. Moreover, at least sometimes we can empirically figure out such risks (consider, for example, scientists who calculate the carbon emissions of everyday activities, and use these calculations for e.g. designing carbon offsetting schemes). Also, even if risks cannot be exactly calculated, strategies such as discounting might still at least sometimes be used to deal with the relevant uncertainty.

[131] Thanks to participants at the Rocky Mountain Ethics Congress 2019 and to Teresa Baron for raising this objection.

One might also object that Anna’s moral obligations entail that she should prevent harm by cleaning the water *without* adding more waste. But the time-relativist need not deny this. Their claim is more modest. They claim that in (Pollution), the constraint against doing harm does not give Anna a reason against the combined choice of both increasing and decreasing risk, not that this reason is decisive for what Anna ought to do on the balance of reasons.

A further objection might be that my stipulation that Anna has previously released waste in the river, such that she merely increases an existing risk rather than introduce an entirely new threat, makes my argument marginal and uninteresting. I have introduced the stipulation to avoid complications arising from the fact that we can never be entirely certain that we have ‘offset’ a previously non-existent risk. However, this does not make the case far-fetched. It is an underexplored feature of morality that the starting point for real-world risk impositions is not a blank slate of certainty. People are usually under non-zero risks of different kinds of harm, and many instances in which agents impose risks of harm are more accurately described as cases in which agents increase the risk of harm.

Finally, it is worth pointing out that there might be moral differences between cases in which agents reduce risks in different ways. To illustrate, imagine that someone has put some poison in my husband’s coffee that makes it 50% likely that his heart will be damaged. I add more of the poison – the likelihood is now 70% - but I also add a few drops of an antidote – the likelihood is back to 50%. Now, there are two possibilities. (1) The antidote directly neutralizes the poison. (2) The poison and the antidote have different effects that happen to outweigh each other. For example, the poison harms the heart via one biological pathway, the antidote protects it via another biological pathway.

In (1), it seems intuitive that the constraint against doing harm does not apply, since the former threat is completely neutralized. However, this seems less obvious in (2). (2) seems to be a case of benefitting in compensation for harming, rather than a case of not harming in the first place. One might think that my contribution in (2) makes me in a sense complicit in the threat, even if I use the antidote. I am not completely sure how best to think about cases such as (2). It seems plausible to me to respond by conceding that (2)-like behaviours (such as Clara’s behaviour in Making Good App) are subject to the constraint against doing harm. However, an alternative response might be to say that the agent in (2) is complicit in the threat in a way that does not amount to doing harm but might still be morally objectionable. In that case, the constraint against doing harm would not apply in (2)-type cases.

Either way, (Pollution) is to be a clear (1)-type case, since the poison is directly removed from the river. A (2)-type case would be a case in which beneficial substances are added to the river. Such a case should be evaluated differently.

## 5.5 An Alternative Proposal

If what I have said so far is correct, then deontologists defending the constraint against doing harm face a new challenge. The time-relative constraint against doing harm is too narrow: it has the implausible implication that the constraint against doing harm applies to cases such as (Pollution) that should not be subject to the constraint against doing harm. The time-neutral constraint against doing harm is too broad: it has the implausible implication that the constraint against doing harm does not apply to cases such as (Making Good App) that should be subject to the constraint against doing harm. Deontologists thus face the challenge of formulating a view that strikes the happy medium. They need to specify the scope of trade-offs that should be subject to the constraint against doing harm. In this section, I would like to offer some thoughts about what a response to this challenge might look like.

I contemplate a *relation-centred* view of the constraint against doing harm. (I discuss how the relation-centred view compares to victim-centred views towards the end of this section.) The relation-centred view says that the constraint against doing harm applies to all behaviour that is likely to put the agent in a harming relation to a victim. More formally, for every agent A, victim V, and pro tanto harm H:

> (Relation-Centred) A has a pro tanto reason against behaviour that makes it more likely that A does H to V.

The relation-centred view gives the right results in offsetting risk cases. Both (Making Good App) and (Pollution) are cases in which agents perform a trade-off between the expected harms that they impose on others. However, the relation-centred view implies that the trade-offs in (Making Good App) are impermissible, whereas the trade-offs in (Pollution) are permissible. In (Making Good App), Clara offsets increases in the risk of harm that she imposes with decreases in the risk of harm in a *different* respect that she imposes on *different* people. In contrast, in (Pollution), Anna offsets increases in the risk of harm that she imposes with decreases in the risk of harm in the *same* respect that she imposes on the *same* set of people.

On the one hand, the relation-centred view forbids interpersonal risk trade-offs. Consider what would happen if Anna decided to release more waste in the river, but to set up the water purification facilities only on one side of the river, thus disproportionally lowering the risk of harms for the residents on this side. Anna's combined choice has then increased the risk of harm for the residents on the other side of the river. It is more likely that her behaviour makes these residents worse off than it would have been, had it not been for her decision.

On the other hand, the relation-centred view forbids intrapersonal risk trade-offs. Consider what would happen if Anna decided to release more waste in the river, but instead of setting up purification facilities, she decides to fight air pollution in the area. We can stipulate that she thereby lowers each residents' risk of harm from air pollution but increases their risk of harm from the toxic waste in the water. Anna's combined choice has then increased the risk of harm to the residents in one respect, but lowered it in another respect, such that her behaviour does not make them worse off *overall* than they would otherwise have been.[132]

A different way to put this is that the constraint against doing harm gives agents a reason against trading off risks of harm between agents, and between different respects in which agents can be harmed, and different times at which their risk of harm is elevated. The constraint against the first type of trade-offs protects agents from being put at greater risk of harm by another agent. The constraint against the second type of trade-offs protects agents from being put under a risk of being made worse off in one respect and then compensated in another respect.[133]

Consider Clara's behaviour in (Making Good App). She increases the risk of harm for some agents in some respect and decreases the risk of harm for different agents in different respects (e.g. offsetting the risk of getting involved in an accident for some against the risk of working under exploitative conditions for others). According to the relation-centred view, her behaviour should therefore be subject to the constraint against doing harm.

In contrast, Anna's behaviour in (Pollution) does not trade off interpersonal or intrapersonal risks. She does not increase the risk of causing harm (either overall or in a respect), and she does not at

[132] This is because I have stipulated that the risk of harm to those living by the riverside is roughly proportional to the total amount of toxic waste t in the river. Previously, Anna has released a certain amount of waste w in the river. Anna's causal contribution to the harm that occurs is equivalent to her share of the total amount of waste w/t. Now Anna decides to release a higher amount of waste, w+n, in the river, and to filter an equivalent amount n out of the river. Anna's causal contribution to harm that occurs is still equivalent to her share of the total amount of waste that is present in the river. This is (w+n-n)/t, which is equivalent to w/t. Anna's causal contribution to the harm remains the same.

[133] I assume here that the constraint against doing harm applies when agents impose risks of pro tanto harm and does not merely apply when agents impose risks of overall harm on others. Thanks to Garrett Cullity for pointing this out.

any time[134] increase the risk to anyone of being worse off (overall or in a respect). In this way, my view gives us the right results in both cases.

The relation-centred view is similar in spirit to *victim- or patient-centred* views on deontological constraints. One such view has famously been developed by Kamm (2007, 26–30; see also F. M. Kamm 1992, 385).[135] Roughly, she argues that agents have a moral status which she calls inviolability. Agents should not be permitted to minimize their own constraint violations, because the implication of such a permission would be that the moral status of all people would be lessened. Along similar lines, Brook (1991) defends a time-relative, victim-centred view of constraints. Roughly, he argues that those who are currently unthreatened have a special moral status that ought to be protected. Therefore, morality should forbid interpersonal trade-offs.[136]

In contrast, the agent-centred view focusses on the importance of one's own agency: the constraint against doing harm requires agents to ensure that *they* do not do harm. (The time-neutral and time-relative view are both agent-centred in this sense.)

The relation-centred view takes on board the key motivations behind both agent- and victim-centred constraints: agents should not be harm doers, and victims should not be harmed. It would be nice to present the relation-centred view as a compromise between agent-relative and victim-relative views. However, so far, I have stated the relation-centred view in seemingly agent-relative terms: agents should ensure that *they themselves* do no harm. This is different from victim-relative views that tend to argue that the harm doer always runs up against the victim's right not to be harmed. The victim has a moral status that makes it prima facie objectionable to harm her, even if the agent harms the victim in order to prevent further harm doings.

This appeal to the moral status of the victim is missing from the relation-centred view. As I have formulated it, the relation-centred view does not necessarily appeal to some intrinsic feature or status of the victim that makes acting impermissible.

However, I suggest that the relation-centred view is in principle compatible with both an underlying agent-centred rationale and an underlying victim-centred rationale.[137] On the one

[134] Recall the stipulation that in Anna's case, there is no time lag between releasing the waste and filtering the water.

[135] Kamm (1996, vol. 2, chap. 10) Kamm (1996, vol. 2, chap. 10) also argues that her victim-focussed view can explain cases of letting oneself do harm, unlike what she calls agent-focussed views of constraints.

[136] For criticism of such victim-centred views, see e.g. (McNaughton and Rawling 1993). See Lippert-Rasmussen (2009) for an argument that, pace Kamm, a victim's inviolability is compatible with the idea that minimizing constraint violations is generally permissible.

[137] I also do not discuss here whether the relation-centred view is in opposition to consequentialist moral theory. However, there is reason to think that this might be the case – at least for consequentialist views that

hand, the rationale behind the relation-centred view can be understood in a way that focusses on the agent: it might be that the constraint against doing harm gives agents a reason to avoid standing in a harm doing relation to a victim. On the other hand, the rationale behind the relation-centred view can be understood in a way that focusses on the victim: it might be that the constraint against doing harm gives agents reasons based on considerations regarding the potential victims of harm. (To make this clearer, one might have to change the formulation of the relation-centred view, perhaps along the lines of 'A has a pro tanto reason against behaviour that makes it more likely that V suffers H done by A'.) For the purpose of this chapter, I remain neutral as to which of these rationales or formulations is correct, and indeed, whether there is a (practical, or principled) difference between them.

In the final paragraphs of this chapter, I would like to consider an objection that might be brought forward by defenders of agent-centred views. They might point out that both the time-relative and the time-neutral view capture something that is of moral relevance: The time-relative view captures the importance of the immediate choice situation (as opposed to one's choices at other times) and the time-neutral view captured the importance of one's own choices (as opposed to other people's choices). Given this, it is tempting to think that deontologists should be *both* time-neutralists and time-relativists. Such a *combined view* would have a time-neutral component, which can explain why Anna is permitted to offset risks in (Pollution), and a time-relative component, which can explain why Clara is not permitted to offset risks in (Make Good App).

However, the combined view gets into difficulties with cases such as (Make Good App). The time-relative view gives Clara a reason *against* using the app. The time-neutral component of the view gives Clara a reason *for* using the app. The defender of the combined view faces the problem of having to explain why the constraint against doing harm, overall, still gives Clara a reason *against* using the app. Unless the defender of the combined view can provide a theory of how the constraint against doing harm applies to such cases, they seem at a loss to explain cases in which the time-neutral and the time-relative view seem to conflict.

The relation-centred view can explain our intuitions about the aforementioned cases without having to enter a complicated debate about whether the constraint against doing harm can give rise to different (even conflicting) reasons in particular cases, and how they should be weighed. It therefore seems to have a clear advantage over the combined view.

reject agent-relativity. Haydar (2002) argues that what he calls the 'localist conception of responsibility', a conception that seems close to my relation-centred view, is in tension with consequentialism: 'According to the localist conception, acts of harm-doing create a special moral bond between the agent and the victim(s) of such acts. It is this special bond that conflicts with the consequentialist ranking principle' (Haydar 2002, 107).

## 5.6 Conclusion

I have argued that the constraint against doing harm applies to all behaviours that are likely to put the agent in a harming relation to others. I have argued, first, that the constraint against doing harm can be generalised to apply to cases under risk. I have suggested that a plausible view on the constraint against doing harm under risk says that agents have a reason against doing something that is likely to constitute doing harm, and this reason is stronger than the reason for doing something that is (just as) likely to constitute preventing (equivalent) harm. I have then distinguished two interpretations of the constraint against doing harm (time-relative and time-neutral) and argued that both interpretations are inadequate in cases in which agents offset risks of harming others. Furthermore, I have suggested that the way forward is to adopt a view of the constraint against doing harm that focuses on the relation between agent and victim of harm. It remains to future research to specify this interpretation further, in particular with regard how it relates to justifications of the constraint against doing harm. Such research seems necessary not only to enhance our understanding of the constraint against doing harm, but also to make these constraints applicable to real-world cases under risk.

# Chapter 6 Doing Harm in the Long Term

## 6.1 Introduction

All actions have consequences. Some of these consequences are direct and predictable. Many more are indirect and unpredictable. Lenman (2000) argues that this gives rise to a problem for act-consequentialists, who believe that the moral status of an action depends solely on the action's overall consequences. If agents do not know what the overall consequences of their actions will be, they are left clueless about the moral status of these actions.[138]

MacAskill and Mogensen (2019) argue that indirect and unpredictable consequences of actions give rise to an even greater problem for non-consequentialists. Most non-consequentialists endorse deontological constraints against doing harm, according to which reasons against doing harm are weightier than reasons to benefit.[139] MacAskill and Mogensen argue that virtually all our actions constitute doing harm, because they are extremely likely to have harmful indirect long-term consequences. They conclude that standard non-consequentialist commitments imply paralysis: the absurd claim that we should refrain from doing anything all. Call this the "Paralysis Problem".

Deontologists should not be surprised by the line of this attack. As I have noted in Chapter 5, the literature on the moral relevance of the distinction between doing and allowing harm has so far mainly focused on cases where a clearly specified harm has already occurred or is certain to occur.[140] This is unfortunate, because it has led deontologists to focus on backward-looking justification for doing harm and left them without a clear account of what deontological constraints against doing harm imply for forward-looking real-world action-guidance. This leaves

[138] Arguably, any plausible moral theory will acknowledge the moral relevance of (harmful) consequences in some form or other (see e.g. Williams (1973, 83)) and might thereby remain vulnerable to (some version of) the cluelessness problem (as noted, e.g., by Greaves (2016, 312).

[139] I follow MacAskill and Mogensen (2019, 1) in their presentation of deontological constraints against doing harm as the conjunction of two principles. The first principle is the Doctrine of Doing and Allowing (DDA), which says that reasons against doing harm are stronger than reasons against merely allowing (otherwise equivalent) harm. The second principle is the Harm-Benefit-Asymmetry (HBA). HBA says that reasons against doing harmful acts are stronger than reasons for doing beneficial acts. HBA plays a crucial role in MacAskill's and Mogensen's argument by pre-empting a potential objection to the Paralysis Problem. The objection is that actions in paralysis cases are just as likely to do good as they are to do harm. HBA explains why the reasons against harm doing are not outweighed by reasons to do good. For the purpose of this chapter, I will assume that the Harm-Benefit-Asymmetry is true.

[140] Cases under certainty are commonly used to illustrate and support the DDA. Examples of such cases are (Push) and (Non-Interpose) in (Woollard 2015, 3), the case of killing a man vs. letting a man die to help others in (Foot 1967, 4), Rescue I and II in (Foot 2002, 81), Rescue III and IV (Quinn 1989a, 298–99).

deontologists without a readily available response to objections that cast doubt on the ability of deontological principles to provide real-world action-guidance, such as the Paralysis Problem.

This chapter has two aims. The first aim is to show how the constraint against doing harm can provide real-world action-guidance. I provide three progressive clarifications of how the constraint against doing harm should be understood when applied to actions with long-term consequences: First, the scope of the constraint is limited to harms that are proximate to the agent's action. Second, agents have limited prima facie permissions to impose risks of harm through everyday behaviour. Third, the constraint against doing harm does not give agents a reason to refrain from actions that do not increase anyone's ex ante risk of harm. All three clarifications are important, because they all shed light on the nature of moral reasons against doing long-term harm. The second aim is to provide a response to the Paralysis Argument. As I will argue, the third clarification provides such a response.

The upshot is that the constraint can guide action in real-world cases, including long-term moral decision making, without falling prey to the Paralysis Problem.

## 6.2 The Paralysis Problem

Before I clarify how the constraint against doing harm applies to actions with long-term outcomes, let me explain the Paralysis Problem in more detail. The Paralysis Problem is set up as a reductio of standard non-consequentialist views. MacAskill and Mogensen set out to show that when indirect and unpredictable consequences of actions are taken into account, some of the non-consequentialist's core commitments imply paralysis: the absurd claim 'that we ought to try to do nothing at all' (2019, 5).

Everything we do is extremely likely to lead to harm in the long term. Everyday actions (such as driving to the supermarket) have lots of indirect and unpredictable consequences. Some of these will be harmful.[141] For example, by driving to the supermarket, Agent might interrupt traffic ever

[141] Even very mundane actions are likely to have far-reaching consequences, because they can affect the identity of those who will exist in the future by influencing when and with whom people procreate (see Parfit (1984, 361)). Greaves (2016) vividly illustrates this point with the example of the causal ramifications of helping an old lady crossing the road, thus causing a slight disruption in traffic with potentially enormous identity-affecting consequences: 'But once my trivial decision has affected *that* [the identity of someone's future child], it equally counts as causally responsible for *everything the child in question does during his/her life* (i.e., for the differences between the actions/effects of this child vs. those that the alternative, in fact unconceived, child would have performed/had) – and of all the causal consequences of all those things, stretching down as they do through the millennia' (2016, 3).

so slightly, with the consequence that Bob's parents meet, who would not otherwise have met. Assume that Bob goes on to kill Victim. But for Agent's driving to the supermarket, Victim would not have been killed.

If a harm arises unpredictably as a result of driving to the supermarket, the agent counts as having done harm. If a harm arises unpredictably as a result of sitting motionlessly at home, the agent counts as having merely allowed harm. It should be noted that this second claim is bound to be controversial.[142] However, for the purpose of this chapter, I grant the claim that sitting motionlessly in Paralysis Cases would count as merely allowing harm. I have two reasons for granting this claim. First, my argument will be stronger if it succeeds even when taking on board controversial assumptions. Second, I am not convinced that rejecting the claim that sitting motionlessly at home counts as merely allowing harm gets deontologists what they want. After all, this would mean that all options that are available to the agent constitute doing harm. It seems plausible, even for deontologists, that if an agent cannot avoid doing harm, she should minimize the harm that she will inevitably do. However, in practice, minimizing the unpredictable future harm resulting from one's action would presumably result in an extremely demanding morality of beneficence that some consequentialists endorse, but most non-consequentialists reject.

Since the long-term consequences of actions are unpredictable, the agent has no greater reason to expect that either behaviour makes it more likely that someone suffers harm. Given this, it seems that the constraint against doing harm gives the agent a stronger pro tanto reason against driving to the supermarket than against sitting motionlessly at home.

Generalizing this point across everyday actions allows MacAskill and Mogensen to conclude that 'taking account of the indirect and unpredictable effects of your action gives you greater subjective reason to ensure, so far as is possible, that the indirect consequences of your behaviour are things you allow to happen, and not things you make happen' (2019, 10–11).[143]

[142] While the paradigm cases of doings in the literature (pushing boulders, throwing knives, pressing buttons) are actions, it has been argued that some inactions also fall on the doing side of the doing / allowing distinction. For arguments against the equation of doing with motion and merely allowing with rest, see, for example, Bennett (1995, 96-100)). For responses to these arguments, see MacAskill and Mogensen (2019, 26–33).

[143] Recall that the reasons here are subjective (i.e. belief- or evidence-based) reasons. Since agents do not know whether their behaviour will result in harm, they do not have access to objective (i.e. fact-based) reasons. Following Lazar's (2017b) decision-theoretic approach, MacAskill and Mogensen propose a subjective (i.e. belief- or evidence-based) interpretation of the constraint. Decision situations can be modelled formally as choices between options with different possible outcomes. Very roughly, Lazar proposes a standard of subjective permissibility which balances demands to maximise expected choiceworthiness with considerations of cost to the agent. Constraints influence the choiceworthiness of

Ensuring this, they suggest, would severely limit the amount of things one can do, and in effect requires agents to do nothing at all.

## 6.3 Proximity of Harm

All our actions have many causal consequences. However, intuitively, most of these consequences are too far removed from our agency to still count as things that we have "done". Consider a case in which a driver runs over a pedestrian. We would say that the driver has done harm. We might think that the driver has done harm because her action has contributed to a causal chain leading to harm. Alternatively, we might think that the driver has done harm because but for the driver's action, the pedestrian would have gone unharmed.

However, there are many people whose actions similarly causally contribute to, and stand in a counterfactual dependence relation, to the harm that the pedestrian suffers. For example, the driver's parents (who conceived her), the driver's neighbour who stopped by for a chat (delaying her departure that fateful day), the doctor who cured the driver's illness (enabling her to drive), and the pedestrian's boss (who insisted on her getting to work early that day). It seems much less appropriate to talk about the harm as something that these people have "done", or "brought about", or are "(agentially) responsible for".

This difference matters morally. Constraints against doing harm give the driver a reason against running over the pedestrian. However, it seems that they do not give the driver's parents, neighbour, doctor, or boss reasons against their actions, despite the similarities in the causal relation to harm.

A general problem for the constraint, then, is how to distinguish between consequences that are what I will call *proximate*,[144] i.e. close enough to fall under the scope of the constraint, and

outcomes. For example, deontologists will rank the outcome in which Agent has killed Victim as less choiceworthy than the outcome in which Agent has merely failed to save Victim (Lazar 2017b, 585). Agents, then, have a subjective reason against behaviour that is likely to constitute harmdoing.

[144] The terminology is borrowed from the legal literature, where in a sense, the investigation runs the other way. In action-guiding, forward-looking moral theory, we usually start with an agent's action and then face the question which of the possible consequences of this action are proximate. In contrast, in tort law, for example, the starting point is usually that a harm occurred and the (backward-looking) question is whether a person's negligent behaviour is a cause of this harm. The law traditionally adopts a two-tier definition of causation. The standard test for whether an action is a *cause-in-fact* of an outcome is a simple counterfactual test: would the outcome have occurred but for the action? Tests for whether actions are *proximate causes* of outcomes vary. Common characterisations include that 'a proximate cause cannot be remote from its

consequences that are what I will call *remote*, i.e. too far removed from the person's agency to fall under the scope of the constraint.[145]

In the next paragraphs, I sketch what I take to be a prima facie plausible account of proximate harm. I suggest that harm is proximate relative to a behaviour if and only if the harm is foreseeable, or it follows directly from the action, or both. I furthermore argue that on any plausible account, the harm in paralysis cases will still be considered proximate.

A harm that results from an action is always proximate when the action is performed with the intention of bringing about the harm. It seems that it is also proximate when the harm is foreseen, or would have been reasonably foreseeable, when the action is performed. For example, dropping rocks from motorway bridges comes with the foreseeable risk of directly harming the car drivers on the motorway.

Things are more difficult when the harm was not easily foreseeable. In such cases, directness (the 'distance' of the agency to harm) seem to matter. For example, Agent might do harm that was not reasonably foreseeable when she puts poison in Victim's tea (even if she has good reason to think it is sugar), or when she runs over Victim in her car (even if she was driving carefully and could not have foreseen the accident). These harms, however, are brought about sufficiently directly to be attributed to Agent despite a lack of foreseeability. (I do not take a stance here on whether directness matters morally in its own right, or whether directness is merely a proxy for foreseeability given a high epistemic standard. Such a standard might, for example, describe conditions under which agents, who are idealised to at least some degree, know that they risk bringing about harm.)[146]

putative effect; it must be a direct cause of the effect; it must not involve such abnormality of causal route that is freakish; it cannot be of harms that were unforeseeable to the actor; its connection to the harm cannot be coincidental; it must make the harm more probable' (Moore 2019, sec. 2.3). However, the notion of proximate cause has been criticized (Stapleton 2008, 468). Recently, the Third Restatement of Torts has even eliminated the notion of proximate causation. Instead, it uses the notion of "scope of liability" to identify whether actions are related to harmful outcomes in ways that give rise to compensation claims. Green states that 'Scope of liability is what proximate cause is really about […] It is not about proximity, it is not about cause, it is about where we're going to say "This is the line we're going to draw on the extent of liability and beyond that we will not hold a defendant liable"' (2011, 1019).

[145] The distinction between proximate and remote outcomes of actions might not only help non-consequentialists overcome epistemological problem. Sinnott-Armstrong (2019, sec. 4) suggests that a form of consequentialism that takes only proximate consequences into account when determining the moral status of acts might solve some epistemological difficulties that consequentialists face: 'This position, which might be called proximate consequentialism, makes it much easier for agents and observers to justify moral judgments of acts because it obviates the need to predict non-proximate consequences in distant times and places.'

[146] More would need to be said about what constitutes directness. The legal criteria for proximate harm might help in fleshing this proposal out further (see footnote 144 above).

For one example of a harm that is neither direct nor reasonably foreseeable, consider a case in which Agent gives birth to Poisoner, who, decades later, puts poison in Victim's tea. For another example, consider a case in which a dog, smelling Agent's grocery shopping, runs across the road, thereby distracting a pedestrian, who drops her packaged cereal bar, where it is found and eaten by a child, who, it turns out, has a severe peanut allergy and goes into anaphylactic shock. In these cases, the harm is too remote from Agent's actions to count as something that Agent has done. It therefore does not fall under the scope of the constraint against doing harm.[147]

Most effects of our actions are remote in this way. For example, in the traffic jam following an accident caused by someone's dropping a rock from motorway bridges, people might meet who would not otherwise have met, and go on to have children together who would not otherwise have existed, and these children might go on to do all kinds of both wonderful and horrible things. We do not know, nor can we possibly know, what the remote effects of our actions will be, but we do know that they are likely to be much greater in number and magnitude than the proximate effects of these actions.

Restricting the scope of the constraint against doing harm to proximate effects is intuitively plausible. Moreover, it corresponds to a common way to think about moral responsibility. Many people think that causal responsibility is necessary, but not sufficient, to ground moral responsibility for a harm (Sartorio 2007, 756). Causal responsibility is not sufficient, because agents are often thought to be causally, but not morally, responsible for outcomes that were unforeseeable at the time of acting, or outcomes that arise from the agent's action in a way that is causally deviant – such as when the outcome is brought about by intervening "freakish" coincidences (Sartorio 2007, 751).

This reasoning supports restricting the scope of the constraint to proximate effects. In paradigm cases of doing harm, we hold the agent prima facie morally responsible for the harm. We say, for example, that Tom cleaned the kitchen, or took the chocolate, and then proceed to praising him for cleaning or blaming him for stealing the chocolate. But we do not seem to do the same in cases where the harm seems clearly remote. We do not praise Tom's grandmother for the clean kitchen or blame his grandfather for the empty pantry. We do not do this, even if the clean kitchen and the empty pantry are the causal effects of Tom's grandparents conceiving Tom's

[147] Some cases are in between the obviously direct and foreseeable, and obviously indirect and unforeseeable cases. These cases might include many cases of intervening agency (see Zimmerman (1985) for an argument that the primary agent retains (some) responsibility for harm, see Bazargan-Forward (2017) for the point that intervening agents do not only share responsibility for harm, but also wrong the primary agent) or complicity. However, I sidestep these debates here. My concern is with those consequences of actions that are clearly remote, as it is these consequences that ultimately give rise to the Paralysis Problem.

parent many decades ago, and even if the kitchen would not be clean and the pantry would not be empty, had Tom's grandparents failed to conceive Tom's parent.

So far, the literature on the distinction between doing and allowing harm has not spelled out such proximity conditions. However, the need for conditions that limit the scope of effects of one's behaviour that one counts as having either 'done' or 'allowed' has been acknowledged: for example, Bennett (1995, 5–6) and Woollard (2015, 17–18) explicitly exclude some outcomes that are brought about when the causal chain leading to harm runs through the voluntary will of others, or depends on intervening coincidences, from the scope of the doing/allowing distinction.

The distinction between proximate and remote harms might give the defender of the constraint against doing harm a way to respond to the Paralysis Problem. The response is to reject the assumption that the constraint applies to long-term, indirect and unpredictable harmful effects of everyday actions. Consider, again, the example of driving to the supermarket, which leads to the premature death of an individual many years from now. On any plausible analysis of directness, this harm should come out as indirect. One might think it should also count as unforeseeable. After all, the agent cannot reasonably be expected to foresee the premature death. One might therefore think that the harm in paralysis cases is not proximate, and therefore, does not give rise to reasons against the actions in question.

Unfortunately, I think that such attempts to avoid paralysis fail. The argument sketched above relies on an ambiguity in the notion of foreseeability. The harms in paralysis cases are not foreseeable in the sense of being predictable by the agent *in their specific instantiations.* Agents cannot know who will be affected by their actions, when, and in what way. However, the harms in paralysis cases are foreseeable in the sense that agents can know that there will be such harms. Agents have some knowledge about the *overall* consequences of their actions: they know that these consequences will very likely include some significant harms.

Here, my response differs from a suggestion by MacAskill and Mogensen (2019, 13–15). They discuss a similar suggestion, based on Lenman's solution to cluelessness, namely that indirect and unforeseeable consequences do not matter morally (in Lenman's own words, that 'the agent should ordinarily simply not regard them [what he calls invisible consequences] as of moral concern' (2000, 363)). MacAskill and Mogensen argue that the most plausible way to define 'unforeseeable' in this context is as outcomes that are sufficiently unlikely (conditional on the performance of the action). This likelihood should be weighted by the extent to which moral reasons disfavour this outcome. The so calculated expected value then determines whether consequences matter morally. MacAskill and Mogensen suggest that this is compatible with their decision-theoretic framework, and therefore does not conflict with their argument.

However, it seems to me that their solution only works if it is already implicitly assumed that ‘unforeseeable’ means that it is unpredictable *whether* harms will materialize. However, once this is made explicit, further considerations of conditional likelihood are not needed to explain why harms in paralysis cases are indeed foreseeable in the relevant sense.

To further support the point that the constraint gives agents a reason against doing harm when the agent cannot predict when and how the harm will materialize, imagine a case in which an agent can spill an unknown chemical in a well, even if the agent has no information about the chemical other than that it will eventually create poisonous residue in the well (but they do not know when that will be, nor exactly what effects the poison will have). Intuitively, the constraint against doing harm should provide the agent with a reason against spilling the chemical in the well.

The defender of the proximity response might insist that the well case is different from paralysis cases. Even if both harms are proximate in virtue of their foreseeability, the agent in the well case, but not in our paralysis case, is aware of the risk she is taking, there is no intervening agency, and the causal chain leading to harm is rather simple (even if extended in time). Such differences might matter morally.[148]

While I agree that such factors are likely to matter morally (e.g. in terms of blameworthiness), I disagree that they influence the reason against doing harm given by the constraint against doing harm. An agent can still count as doing harm to the future poison victims even if she has no idea of the dangers of the chemical, even if she orders someone else to spill the chemical in the well, or if the chemical reaction in the well proceeds in many complicated causal steps.

To sum up, I have argued that the constraint only applies to harm that is proximate. I have sketched a view on proximity, according to which harms are proximate whenever they are either reasonably foreseeable or follow directly from the agent’s actions. I have then argued against the claim that the harm in paralysis cases is not reasonably foreseeable.

[148] I take it that several of the objections that MacAskill and Mogensen discuss, and ultimately reject, in their paper are variations of this response. For example, ‘that the indirect and unforeseeable consequences of our actions are morally irrelevant’, ‘that you do not count as doing harm if the causal chain goes through the voluntary acts of other agents, ‘that some intermediary event linking present behaviour to future unforeseeable harm fails to count as substantial’ (MacAskill and Mogensen 2019, 3).

## 6.4 Permissible Risk Imposition

All our actions impose non-zero risk of harm on others. When you drive, you risk accidents, when you operate your gas oven, you risk an explosion, and when cooking for others, you risk accidentally poisoning them. However, intuitively, agents are permitted to perform these actions. In this section, I argue that the Doctrine of Doing and Allowing (DDA), the claim that reasons against doing harm are stronger than reasons against merely allowing harm, can ground permissions to impose (some) risks of harm on others. I discuss whether such permissions include permissions to impose risk in paralysis cases, and conclude that unfortunately, they do not.

The argument is based on Woollard's (2015) account of the DDA. According to Woollard, agents have a claim to self-ownership, which entitles them to prima facie protection from both causal and normative imposition. Woollard illustrates the difference between the two kinds of imposition with a driver who hits a pedestrian and thereby causally imposes upon the pedestrian; whereas a driver who is required to save the pedestrian from a falling tree is normatively imposed upon by the pedestrian (2015, 98–99). Woollard further argues that protection from such imposition is 'necessary if anything is to genuinely belong to anyone' (2015, 105). If it were not prima facie impermissible for others to harm me, or prima facie permissible for me to not help everyone to the greatest extent possible, then it would be hard to see how my body would still genuinely belong to me.

A justification of the DDA along these lines can be generalized to apply to cases of everyday actions that impose trivial risks of harm. Everyday actions impose risks of harm. A prohibition of such everyday behaviour, however, would restrict the agent's lives in a way that amounts to normative imposition. As Woollard describes it, '[n]ormative imposition involves the imposer's needs intruding into the person's sphere, requiring the person to put his or her body or resources at the imposer's use' (Woollard 2015, 109–10).

Prohibitions against imposing trivial risks of harm amount to such normative imposition. Agents could only permissibly act in ways that are certain to harm nobody. They could not operate ovens, drive cars, or play sports without being subject to moral scrutiny. This normative imposition would be significant. It would not just unduly limit, but effectively deny the agents' authority over their own bodies and resources. After all, bodies and resources are always potential threats, even when used in careful and non-malicious ways.

The upshot is that, perhaps surprisingly, the DDA itself provides us with prima facie permissions to not only allow harm, but also to perform at least some behaviours that increase the risk of ourselves doing harm. Imposing risks of harm is permissible if the imposed risk does not outweigh the burden of normative imposition on the agent. For example, being prohibited from operating household appliances would constitute a significant normative imposition on the agent, in the sense discussed above. In this case, the DDA should say that the agent has prima facie permissions to operate household appliances. In contrast, consider an agent who, purely for her own enjoyment, likes to drop stones from her balcony on the pavement below. Giving up this hobby does not seem to be a significant imposition. In this case the agent does not have a prima facie permission to impose risks.

The defender of the DDA can argue that these considerations provide a solution to the Paralysis Problem. After all, a similar point can be made for imposition of risks of harms in paralysis cases. Permissions to perform everyday activities that impose remote harms on future people protect us from normative imposition, preserve our freedom of action, and protect rights to impose risks on others. The same conclusion then, it seems, can be drawn for imposing *risks* of remote harms: such impositions are prima facie permissible.

Such permissions enable the defender of the DDA to reject what I think is an implicit premise in MacAskill's and Mogensen's argument. This implicit premise is that the pro tanto reason against doing harm given by the DDA is not cancelled or outweighed by another, sufficiently weighty, reason *for* performing the behaviour in question. However, given the previous discussion, it seems plausible that prima facie permissions to engage in everyday behaviour do outweigh the risk of long-term indirect and unpredictable harms arising from this behaviour.

Unfortunately, permissions to impose risks of harm do not solve the Paralysis Problem. To see why, recall that protections from normative imposition only make risk impositions permissible if two conditions are fulfilled: first, that the risk and magnitude of harm that is imposed is sufficiently small, and second, that it would significantly restrict the agent's freedom of action or body ownership to forego imposing these risks.

The second condition is fulfilled in paralysis cases. Being prohibited from everyday actions like cooking or driving surely constitutes a normative imposition. This might not be immediately obvious, since such prohibitions do not seem particularly onerous when considered individually (one missed trip to the supermarket presumably does not constitute a major imposition). However, deontologists should accept the claim that it is prima facie permissible to risk doing harm if and only if the risk of harm imposed by a behaviour is sufficiently small and not having prima facie permission to perform *that type of risk-imposing behaviour* would impose normatively on the agent. It seems plausible to suggest that behaviours count as being of the same 'type' in the relevant sense when they both fit the finest grained description that would still impose the same type of risk.[149] For example, “using the hob” imposes a non-zero risk of a gas explosion. The type of risk imposed remains the same whether the behaviour in question is “switching on the bottom right hob” or “switching on the top left hob”.[150] In the case of long-term, unpredictable and indirect harms, however, anything we do imposes such risks. In order to avoid such risk impositions, we would have to stop doing anything. This would be an enormous normative imposition.

However, the first condition – that the expected harm is sufficiently small – is not fulfilled in paralysis cases. The Paralysis Problem shows that it is extremely likely that our actions will cause significant harms over time. Permissions to impose risks of harm might be strong enough to permit driving to the supermarket, in the sense that they outweigh the tiny likelihood of causing a fatal accident.

It is much less clear that permissions to impose risks of harms are strong enough to permit driving to the supermarket, in the sense that they outweigh a near-to-certainty likelihood of causing a fatal accident. The point of the Paralysis Problem is that seemingly innocent actions like driving to the supermarket impose near-certainty risks of grave remote harms. This is because anything an agent does is virtually certain to lead to such harms somewhere in the (far) future. The likelihood and magnitude of these remote harms seems too big to be declared permissible by prima facie permissions to impose trivial risks of harm.

One might object that the magnitude of expected harm is not high enough to *completely* outweigh permissions to impose risks of harm. However, in this case it seems reasonable to assume that it would at least ground Partial Paralysis. Agents might not be required to do as little as possible, but they might be required to do significantly less. This already seems like an unwelcome result for the deontologist.

To recap, so far, I have explored two features of a constraint against doing harm as a forward-looking, action-guiding principle. First, I have argued that the scope of outcomes that the constraint applies to is limited to harms that are sufficiently proximate to the agent’s actions. Second, I have argued that the principle which gives rise to such constraints, the DDA, grounds prima facie permissions to impose small risks of harm on others by performing everyday actions

[149] Lazar and Lee-Stronach (2019, 104–8) propose this as a mechanism to help us figure out whether the unit of moral evaluation in cases under risk are acts or sequences of acts in particular cases.

[150] A similar point can be made using Scanlon’s argument that in deciding whether to adopt a principle, we need to take into account the cost of the general acceptance of this principle. Like Scanlon’s example of being required to be impartial between one’s own and other’s (comparable) interests, being required to abstain from actions in paralysis cases might give rise to ‘generic reasons that everyone in the position of an agent has for not wanting to be bound, in general, by such a strict requirement’ (1998, 225).

(this includes some long-term risks of harm). My discussion shows that the constraint against doing harm can explain common sense intuitions about the remoteness of harm from one's agency, and about the permissibility of imposing risks. Unfortunately, as it turns out, neither of the two features can solve the Paralysis Problem.

## 6.5 Ex Ante Risks (And, Finally, a Solution to the Paralysis Problem)

In this section, I argue that deontologists can solve the Paralysis Problem. I argue that the constraint against doing harm does not give agents a subjective reason against acting in paralysis cases. This is because actions in paralysis cases do not increase anyone's ex ante risk of harm. If agents do not violate the constraint against doing harm when performing everyday actions, then the Paralysis Problem does not arise.

Let me start by distinguishing ex ante and ex post perspectives on behaviour. Consider a standard paralysis case. If Cora drives to the supermarket today (at t1), then this will have the unpredictable consequence that Dave dies prematurely 200 years from now (at t2). If Cora stays at home at t1, then this will have the unpredictable consequence that Eugen dies prematurely at t2. If Cora drives to the supermarket at t1, Cora counts as having done harm to Dave. If Cora stays at home at t1, she counts as having merely allowed harm to Eugen.

But from which temporal perspective is this outcome to be evaluated? On the one hand, we can evaluate the outcome ex ante: at the time the agent decides to act (i.e. at t1). On the other hand, we can evaluate the outcome ex post: at the time when the risk has materialised or failed to materialise (i.e. at t2). From an ex ante perspective, it seems that Cora's action increases the likelihood, for all possible people, that Cora will have caused their premature death. (Since there are so many possible people, the outcome from an ex ante perspective is one expected death.) From an ex post perspective, uncertainty is eliminated: we know that Dave is the unlucky one who dies prematurely as a result of Cora's action.

It has been noted that the same risky conduct often seems more easily justifiable from an ex ante perspective than from an ex post perspective (e.g. Scanlon 1998, 236).[151] However, both

[151] From an ex ante perspective, it often seems permissible to impose a very small risk of grave harm on some, even in order to generate relatively trivial benefits. Scanlon gives the example of public projects such as building a bridge which makes travel more convenient, but risks fatal accidents during construction works (1998, 236). (This example also involves different numbers – many more people will enjoy more convenient travel than are endangered in the building works. However, I think that our intuitions would remain essentially the same if the bridge only led to the mountain hut of an eccentric billionaire.) In contrast, from an ex post perspective, similar conduct can seem impermissible. Surely, we ought not to order construction

perspectives describe the same behaviour. The standards for (subjective) permissibility for behaviour should not depend on how the behaviour is described (Fried 2012c, 249; see also 2012b, 50–51). An action-guiding moral theory needs to be able to tell Cora is acting permissibly when she drives to the supermarket, whether her behaviour is viewed from an ex ante perspective as 'increasing the risk of premature death to each of a very large number of (possible) people ever so slightly'[152] or from the ex post perspective as 'having done harm to Dave'.

### 6.5.1 The Risk of Being a Victim of Harm

I argue that the constraint against doing harm does not give agents reasons against acting in paralysis cases. This is because these actions do not increase the ex ante risk of harm for anyone.[153] To illustrate, consider the following simplified toy cases:

> (Doing) Ann places a rock on the rail of a motorway bridge. It is very likely that the rock will fall on the motorway soon.
>
> (Allowing) A rock lies on the rail of a motorway bridge. It is very likely that the rock will fall on the motorway soon. Ann can now remove the rock but does not do so.

The expected harm is the same in both cases. In (Doing), Ann will be relevant to this harm in a doing way. In (Allowing), Ann will be relevant to this harm in an allowing way. The constraint against doing harm gives Ann a reason against placing the rock on the rail in (Doing), but not against failing to remove the rock in (Allowing). Now consider:

> (Replacing) A rock lies on the rail of a motorway bridge. It is very likely that the rock will fall on the motorway soon. Ann replaces the rock with an identical one. The new rock is just as likely to fall on the motorway as the old one. Rockfalls occur unpredictably, however, the new rock is unlikely to fall on the motorway at exactly the time at which the old rock would have fallen on the motorway.

worker Fred to build a bridge if we know he will die doing so, even if this means that travellers will be inconvenienced. As Scanlon (1998, 236) notes, our intuitions here seem to be driven by our implicit assumption that adequate precautions have been taken. I will elaborate on the role of reasonable precautions in subsection 6.5.2.

[152] This formulation focuses on the probability distribution of the expected harm. An equivalent formulation that focuses on the frequency of the expected harm as "causing one expected premature death". See Fried (2012b, 50–51) for the point that non-consequentialists tend to see behaviour described in terms of probability and frequency as two distinct forms of conduct, rather than the same conduct described in different terms.

[153] The actions that I have in mind in this section are instances of everyday behaviour, like in the paralysis examples. I completely ignore questions of proximate harm here (e.g. the risk of running over a pedestrian when driving), and only focus on long-term, indirect and unpredictable harm.

It seems that as far as the constraint against doing harm is concerned, (Replacing) is equivalent to (Doing). Ann will be relevant to harm, if it occurs, in a doing way.

However, (Replacing) is very different from (Doing). Consider the perspective of the potential victims of the rockfall. In (Doing), the drivers on the motorway are under no risk of a rockfall (at that time and place) before Ann puts the rock on the rail. Only after she has put it there, they are under a great risk of harm. In contrast, in (Replacing), they are already under a great risk of a rockfall before Ann does anything. Ann's replacing the rock does not increase this risk. Ann has not increased the ex ante likelihood for anyone that they will suffer a harm in some respect.[154]

I argue that paralysis cases are more like (Replacing) than they are like (Doing). This is because for every possible remote harm H, it is just as likely that an everyday action *causes* this harm as it is that the action *prevents* this harm from occurring. The action is therefore just as likely to make it the case that H occurs as it is to prevent H from occurring. To see this, consider a future person P who lives 200 years from now. Is it more likely that one's driving to the supermarket today causes P to die prematurely? Or is it more likely that one's driving to the supermarket today prevents P from dying prematurely? Given that the long-term effects of driving to the supermarket are unpredictable, it seems reasonable to assume that both are equally likely.

Moreover, given that every action has multiple causal effects, one single action can influence the causal chain leading to H in multiple ways. This might mean that the action causally contributes to H in more than one way. For example, one action might influence the identity of both the father and the mother of a person who goes on to do some harm. However, it might also mean that one effect of the action prevents another effect of the action from causing harm. For example, one action might influence the identity of a parent who takes their child out on a boat ride in the

[154] Not everyone might share the intuition that Ann's behaviour in (Doing) is significantly harder to justify than her behaviour in (Replacing). I suggest that this is because our intuitions pick up on the idea that Ann should remove the rock without replacing it, and it is difficult to imagine the stipulated fact that Ann cannot do this. Moreover, having intuitions about the case might be difficult because replacing the rock seems unmotivated. However, we can imagine that failing to replace the rock would be costly for Ann. It then seems plausible that Ann should be required to bear a greater cost to avoid putting the rock on the rail in (Doing) than to avoid replacing the rock in (Replace). Finally, it is worth noting that we can construe many cases in which agents change causal pathways to harm (thereby "shuffling the deck", i.e. potentially making a difference with regard to who will end up being the victim of some harm), but not making it more likely for anyone that they will suffer the harm. A cyclist among many in the morning rush hour might ever so slightly change where and when exactly others ride their bikes (those around one might speed up or slow down, move left or move right, and that in turn might change what others do). There is a piece of broken glass on the road, and a cyclist's tyre will burst if they ride across it. It seems plausible to assume that cycling might change who ends up with a burst tyre. However, it does not change whether anyone ends up with a burst tyre, and it does not change the ex ante likelihood, for any cyclist, that their tyre will burst. So, like Ann's replacing the rock, my taking the bike to work in this example causes harm without increasing anyone's ex ante risk of harm.

storm and influence the identity of the lifeguard who saves the child from drowning. An action might therefore fail to increase the overall likelihood that a harm occurs, despite setting a causal chain in motion that will result in harm if this harm is not prevented.

I conclude that actions in paralysis cases do not increase the risk, for anyone, that they suffer harm. If this is correct, then this plausibly weakens the reason against doing harm. The case in which Agent does something that harms Victim and thereby makes it more likely that Victim suffers that harm is morally different from the case in which Agent does something that harms Victim but does not increase the likelihood that Victim suffers that harm.

This claim can be supported by appeal to the literature on the DDA. Deontologists have previously argued that the reason against doing harm is weaker when the agent's behaviour does not make a difference to whether Victim suffers harm. In an example for a so-called pre-emption case, Agent can now bring about some good, but only by doing harm to Victim. If Agent decides not to harm Victim, then Evil will do the same harm to Victim (without bringing about any good consequences). In this case, Victim will be harmed no matter what Agent does. In discussing these cases, Woollard states that harming someone is 'significantly easier to justify if the harmed person is not made worse off' (2012a, 688).[155] For example, it seems easier to justify to kill someone in order to save a life, if the person would have been killed regardless.

We might think of paralysis cases as risky pre-emption cases, in which the victim will face a certain risk of harm, regardless of what the agent does. Like in the pre-emption case under certainty, the agent's behaviour does not make a difference to the likelihood for Victim to suffer harm. The behaviour of an agent who risks doing harm to a victim seems significantly easier to justify if the agent does not thereby increase the ex ante risk of harm that the victim faces. For example, it seems easier to justify imposing a risk of death on someone in order to save a life, if this does not change the ex ante risk of death for that person. In both cases, the imposition on Victim seems to be much less serious than in standard cases of doing harm. As long as the risk of harm remains the same, it seems that the extent to which Victim is imposed upon also remains constant.

However, if the reason against doing harm is much weaker in paralysis cases than in standard cases, deontologists can avoid paralysis. They can argue that the residual reason against doing harm in paralysis cases is simply not strong enough to ground obligations to refrain from everyday behaviour.

[155] Gardner's 'Redundant Harming' Principle about the strength of the reason against harming has similar implications (2017, 86).

One might object that the constraint against doing harm should apply in paralysis cases, because just as everyday actions do not increase the ex ante risk of harm for anyone, sitting motionlessly at home does not decrease the ex ante risk of harm for anyone (it is just as likely that sitting motionlessly fails to do harm as it is that it fails to prevent harm). So, is the reason against doing harm without increasing ex ante risks stronger than the reason against allowing harm without decreasing ex ante risks?

It is not. It is far from obvious that the justification for the difference in the strength of reasons against doing and allowing harm applies to cases in which ex ante risks remain the same (for example, the normative imposition on an agent posed by requirements to aid is not any more or less weighty if the aid does not actually decrease risk of harm for anyone). More importantly, moral permissions to allow harm do not straightforwardly apply to cases in which agents have an equal chance of either failing to prevent harm or failing to do harm (rather than either failing to prevent harm or doing nothing).

In the remainder of this section, I discuss an objection to this line of argument. MacAskill and Mogensen (2019, 23), following Kamm (2007, 274–75), argue that deontologists should refrain from evaluating outcomes from an ex ante perspective.[156] This is because such a view would have implausible implications in the following case:

> Ambulance: A community considers the purchase of an ambulance. The ambulance has an on-board artificial intelligence and is rigged with special brakes. If the ambulance is hurrying to the hospital, the on-board AI will kick in and make it impossible to stop the ambulance from running over someone in its way if this is necessary to save a greater number of people being transported on-board ((MacAskill and Mogensen 2019, 23), the case is due to (F. Kamm 1996, 2:303)).

The decision to buy the ambulance is intuitively impermissible. However, it seems that ex ante reasoning does not support this judgement. After all, buying the ambulance does not increase the ex ante risk of harm for anyone in the community.[157] (The decision is more likely to save the life of any community member than to kill them.) As MacAskill and Mogensen note, the reason why

[156] In their words, 'non-consequentialists ought to reject the idea that ex ante Pareto optimality vitiates the force of deontological reasons against doing harm' (2019, 22). An outcome is (weakly) pareto optimal if it is worse for nobody, and better for at least one person. Actions in paralysis cases are weakly ex ante pareto optimal, because they benefit the agent and do not increase the risk of overall harm to anyone, since they are just as likely to lead to be harmful as they are to be beneficial.

[157] Similarly, it seems impermissible to adopt a system according to which people are killed and their organs harvested if this enables more people to live, even if adopting such a system increases the ex ante chances of survival for everyone (this famous thought experiment was proposed by Harris (1975)).

buying the ambulance is impermissible seems to be the same kind of harm-based reason that the constraint against doing harm generates (2019, 23). Intuitively, it is wrong to kill one person, even if this produces more good overall, and even if it would have been rational for the person who ends up being killed to agree to such an arrangement in advance.

However, I argue that deontologists can explain why buying the ambulance is impermissible, and moreover, why the constraint against doing harm supports this judgement, from an ex ante perspective. The key idea is that the constraint is not a principle that about *overall* harm; rather, it is about pro tanto harm. This is because compensated harm is still harm. The constraint against doing harm should tell us to refrain from doing harm, even when we compensate it. Consider, for example, the company that sells products that increase the risk of cancer A, but reduces the risk of cancer B. Even if these equal out such that the expected overall impact on health is 0, the company still increases the risk of cancer A, and therefore the expected pro tanto harm of suffering this particular cancer.

The suggestion, therefore, should not be that the forward-looking, action-guiding constraint against doing harm only applies to actions that worsen anyone's *overall* ex ante prospects. Rather, it should be that the constraint against doing harm applies to actions that increase anyone's likelihood of suffering *pro tanto* harm.

To clarify, my point here is *not* that, in paralysis cases, your action is just as likely to cause good things as it is to cause bad things to anyone.[158] Whether this claim is true is irrelevant to the Paralysis Problem, since the constraint gives agents a reason against doing harm whether this harm is compensated by accompanying benefits or not. I think that the focus on overall well-being (as opposed to a focus on pro tanto harms) has caused confusion in MacAskill's and Mogensen's discussion.

One might respond to my argument by insisting that the people in (Ambulance) are harmed in the same respect. Some people die in traffic accidents caused by cars, others in traffic accidents caused by ambulances: this relevantly looks like the same type of harm. It might seem fine to offset the risk of fatal car accidents with the risk of fatal ambulance accidents.

[158] MacAskill and Mogensen concede this point (2019, 21). It can be questioned, for it seems to depend on the overall future welfare levels. If the future consists in bleak misery for those alive, then our actions are very likely to lead to more harms than benefits. If the future consists in constant bliss, then our actions are very likely to lead to more benefits than harms. I leave this complication aside and assume that the world in which remote harms occur is relevantly like our own, or at least that it is just as probable that the future will be better as it is that it will be worse.

However, the harm in these cases is not the same type of harm. The decision to buy ambulances does not reduce the number of car crashes. It ensures that those who are involved in car crashes are more likely to get help and survive.[159] One might describe what goes on in the ambulance case as follows. The decision makes it more likely that people will be involved in car accidents. It also makes it more likely that people will receive help if they get involved in car accidents (and presumably, the welfare won by the second measure is greater in magnitude than the welfare lost by the first measure). Compare a different version of (Ambulance):

> (Ambulance*) A community considers the purchase of an additional ambulance. Due to factors such as weather conditions and poor driving skills, ambulances regularly cause accidents in this community. It therefore is to be expected that an additional ambulance would increase the risk for members of the community to be involved in a traffic accident, caused by ambulances run by the community. However, the community also considers running extra driver's training for ambulances. Such training would decrease the risk for members of the community to be involved in a traffic accident, caused by ambulances run by the community.

In this case, the constraint against doing harm does not give the community a reason to refrain from buying the additional ambulance. Why? Because it is an act in a sequence of acts that, overall, does not make it more (ex ante) likely for anyone affected that they will suffer harm in any respect, at any time.[160]

To recap, I have argued that actions in paralysis cases do not increase anyone's ex ante risk of harm. This weakens the strength of the reason against doing harm. I have then considered the objection that non-consequentialists should not endorse considerations based on the evaluation of ex ante risks and argued that the objection fails. In particular, (Ambulance) does not pose a counterexample to the ex ante evaluation of risky outcomes.

### 6.5.2 The Risk of Being A Harm Doer

I have argued above that the constraint against doing harm does not give agents reasons against actions that do not increase anyone's ex ante risk of harm. I will call this the "ex ante view".

[159] As an aside, one might argue that these people do not suffer harm from not being transported to hospital in the first place: the accident puts them in a harmed state, and the lack of aid fails to take them out of that state.

[160] The extra training is an example of taking reasonable precautions to prevent harm. I discuss the role of reasonable precautions in section 6.5.2.

Here is an objection to the ex ante view. Even if agents in paralysis cases do not worsen anyone's ex ante prospects, they have made it the case *that they will have done harm*. There will be a victim that is harmed by the agent that would not have been harmed, had the agent acted differently. Intuitively, agents should care about being harm doers: they should avoid being harm doers. However, the ex ante view cannot explain this intuition. What this shows, the objection goes, is that the ex post perspective also matters morally. According to this view, agents have a (subjective) reason against doing something if they know that this will make it the case that someone[161] will suffer harm. I will call this the "ex post view".

Before I give my reply to this objection, it is worth pointing out that this formulation does not rule out a hybrid view, according to which actions that are likely to do harm (ex ante) and actions that have resulted in harm (ex post) are both constrained. I deliberately leave it open whether the defender of the ex post view could adopt such a hybrid view. However, much more would need to be said as to how the details of such a view would be spelled out. It is unclear how ex ante and ex post considerations would add up, and how this should be regulated in a non-arbitrary way. Moreover, borrowing from the rich literature on ex ante and ex post contractualism, there is reason to think that such hybrid views might give counterintuitive results in the case of asymmetric information (Reibetanz 1998, 302; see also Fried 2012b, 54).

For an example that applies to our case, imagine that an agent is forced by an evil demon to either drive to the supermarket or fire a bullet at the hand of a sleeping stranger (luckily, the gun has very many chambers, so that the likelihood of harm is very small). The agent is virtually certain that by driving to the supermarket, she will have caused someone's premature death. Playing Russian Roulette is not very likely to cause the stranger to be shot in the hand. The ex ante perspective gives her a presumably weak reason to drive to the supermarket (since the likelihood of harm to the stranger is low). The ex post perspective gives her a strong reason against driving to the supermarket (since it is virtually certain that she will have caused a death). It seems that a hybrid view of the constraint against doing harm should tell the agent to shoot. This seems implausible.

Let me get back to the objection raised at the beginning of this section. My response to the defender of the ex post view is that the burden of proof is on them to show that the ex post perspective, by itself, gives agents reasons for refraining from actions. In the following, I argue,

[161] This objection takes the view that we should discount complaints against being harmed by the likelihood that *someone* will be harmed (Otsuka (2015, 84) defends such a view). This is a relatively low standard of certainty required for assuming the ex post perspective. A perhaps more common view in the literature on contractualism and aggregation is that we ought to discount complaints against being harmed by the likelihood that an *identified individual* will be killed (see e.g. Kumar 1999, 295; Reibetanz 1998, 301).

furthermore, that reflection on cases does not support the view that knowledge that one will have done harm (in the absence of increased ex ante risks of harm) gives agents forward-looking reasons against actions. Finally, I suggest that when agents have such knowledge, this can strengthen obligations to take reasonable precautions, in order to mitigate the foreseen compensation claims towards them. I explain why such obligations, however, do not lead to paralysis.

To clarify, my aim in this section is not to show that the ex post view is false. The claim I am making here is more modest. I defend the ex ante view from the objection that it cannot explain the intuitive relevance of knowing that one will have done harm. I argue that the objection puts deontologists under no pressure to accept the ex post view.

The burden of proof is on the objector to show that the ex post perspective matters per se (i.e. if and insofar as it differs from the ex ante perspective). This is because the constraint against doing harm is, in the first instance, a principle about paradigm cases of doing and allowing harm. For a paradigm case of doing harm, consider a case in which an agent shoots someone. From an ex ante perspective, the agent ensures (increases the risk) that the victim will suffer harm, and from an ex post perspective, the agent has done harm to the victim (has done harm to someone). As I have argued in section 6.5.1, paralysis cases are different from these paradigm cases. The agent in paralysis cases does not increase the ex ante likelihood of suffering harm for any given individual. Given this, I believe that defenders of the ex ante view are justified in shifting the burden of proof on their opponents. The critic of the constraint needs to explain why the constraint should retain (much of) its original strength in paralysis cases, even though these cases are unlike paradigm cases.

Moreover, there is reason to doubt the objector's assertion that intuitions about cases support the ex post view. Consider the following case:

> (Vaccination A) Officer A receives news that a disease has broken out. The good news is that vaccination is possible, and enough vials of the vaccines are stored in a medicine storage unit. The bad news is that some of the vials containing the vaccine have been contaminated – they will cause the same disease that they are supposed to protect against. Time is too short to test the vials. Officer A orders a mass vaccination, knowing that this decision is extremely likely to cause some people to catch the disease that would otherwise have remained healthy.

Officer A's decision seems permissible, even though her decision will result in harm to some. (Vaccination A) is unlike paralysis cases, however, since in paralysis cases, the agent's action does not increase everyone's ex ante prospects of receiving benefits. Consider a variation:

> (Vaccination B) Officer B oversees the mass vaccination. She gets information that a different set of vials of the same vaccine is stored in a second storage unit. Unfortunately, the same contamination problem obtains there. However, it would be slightly more convenient for B to distribute the vials from the second unit (e.g. it would be slightly cheaper, less paperwork, or the like). If Officer B gives order to use the vials stored in the second unit, given different transport systems and coincidences, it is almost certain that different people will receive contaminated vials than would have received contaminated vials, had the vials from the first unit been used. The choice of the storage unit, however, does not make a difference to anyone's risk of receiving a contaminated vial. Officer B decides to use the vials from the second unit.

Officer B's decision seems intuitively permissible, even though her decision will result in harm to some. In this variation of the case, her decision does not benefit anyone (apart from being slightly more convenient for B herself). (Vaccination B) is, then, analogous to paralysis cases, in that the agent does not influence the probability, for anyone, that they suffer harm (and secures the agents themselves a minor benefit). Intuitions about the vaccination cases seem not, at least not unequivocally, to support the ex post view.

This might be surprising. After all, doing something that makes someone suffer harm seems intuitively bad. Moreover, the critic can say, it seems to be this intuition that lends significant force to constraints against doing harm. For example, one might think of pre-emption cases, where a victim will be shot by someone else if the agent does not shoot them. The agent does not worsen the victim's ex ante prospects by shooting. However, the agent clearly should not shoot! Do we not, the critic might ask, need the ex post view to explain this judgement?

I do not think so. A first point to note is that our intuitions might track the uncertainty that agents in real-world cases will always have about the ex post perspective on their actions. It is very difficult to imagine that one's actions do not change the ex ante likelihood of some outcome occurring. In the pre-emption case, for example, it is tempting to think that if Agent shoots Victim, Agent thereby ensures Victim's death. After all, if Agent does not shoot Victim, perhaps Victim would have escaped by some means. Agent rules out this possibility. Moreover, it is impossible to confirm retrospectively, once Victim has been shot dead by Agent, that there would not have been such a possibility. Agent will therefore have to live forever with the uncertainty whether Victim would still be alive had Agent not shot them. This lingering uncertainty might haunt Agent even if, given all the available evidence, this would have been a very unlikely event.[162] I suggest that this uncertainty drives some of our intuitions in pre-emption cases.

A similar point can be made about paralysis cases. It is tempting to think that Cora's driving to the supermarket makes it more likely that Dave dies. Because it is usually the case that when someone causes harm, they have also made this harm more likely, it is very hard to imagine cases where this is different.

However, to defend the ex ante view, I need not explain away the intuition that the ex post perspective has moral relevance. On the contrary, I can happily concede this claim. What I maintain as a defender of the ex ante view, however, is that ex ante and ex post perspectives operate in different domains. Prohibitions of actions are grounded in ex ante evaluation, whereas compensation for outcomes is grounded in ex post evaluation (Fried 2012c, 237–38). These perspectives are importantly distinct.[163] To illustrate, the person whose speeding at t1 has caused an accident at t2 is morally responsible and legally liable, even if another person's speeding at t1 was similarly likely to cause the accident. However, it would be a mistake to infer from this that at t1, the first person had a stronger reason to refrain from speeding than the second person.

"Wait!", the objector might say here. "You have just admitted that a person who suffers harm as a result of a contaminated vial in (Vaccination B) has a valid compensation claim against officer B. However, officer B *knows* that such claims will arise, simply because it is virtually certain that some people will end up getting the contaminated vials. It seems plausible that officer B should acknowledge these claims *now* – before the harmful outcome has occurred. And surely, this gives officer B a reason against acting *now*!"

I disagree. Officer B should indeed acknowledge the fact that her behaviour will give rise to claims by those who will suffer harm as a result of her action. However, it is a mistake to think that this (necessarily) gives officer B a reason to sit in her chair, watching the vials being distributed from the first unit. This is because acknowledgement of future compensation claims can be achieved by taking reasonable precautions. Reasonable precautions seldom require abandoning the risky activity altogether, especially if doing so would be costly. Rather, reasonable precautions can

[162] The feeling of "had I not been there, maybe things would have turned out better" has famously been explored by Williams (1981, 27) under the name of agent-regret.

[163] Fried, following Scanlon (1998, 236) points out that non-consequentialists tend to conflate distinctions in their discussion of risk. Perhaps chief among them is the distinction between (forward-looking) prohibition and (backward-looking) compensation (Fried 2012c, 238–40). She points out that we can 'decouple judgments about whether conduct is wrong, in the sense that we would have prohibited it if we could have, from judgments about whether […] the victim should be compensated, and if so, whether the compensation should come from the person who caused the harm' (2012c, 238).

consist in minimizing the risk that their action imposes on others. An example is providing safety equipment for construction workers or organizing extra driver's training in (Ambulance*). Another way of taking precautions is, perhaps, setting up systems or institutions that can mitigate harm or compensate those who end up suffering harm as a result of one's actions. (An example might be setting up funds or insurance systems for workplace accidents.)

However, Agents in paralysis cases automatically take such reasonable precautions. On the one hand, the actions they perform do not increase anyone's ex ante risk of harm. On the other hand, actions in paralysis cases are just as likely to benefit any given person as they are to harm them. Admittedly, the actual benefits are unlikely to occur to those that will be harmed in a way that compensates them. However, individual agents in paralysis cases arguably cannot effectively target their compensation efforts, since they have no idea which harms their actions will cause to whom and when. Moreover, any compensatory measure that agents would put in place would give rise to exactly the same risks of harm as the risk-imposing actions themselves.[164] It seems to me that acknowledging the fact that one's everyday actions are likely to result in harm in the long term does not put agents under any obligations to take precautionary measures.

However, in concluding, I would like to suggest that acknowledging that one's agency will inevitably be involved in future outcomes should motivate agents to help make the future world a somewhat better place. If what I have argued above is correct, then agents have no (harm-based) moral reason to stop doing anything upon realizing that their actions will lead to long-term harms. However, I believe that this realization should motivate us to do the opposite. We should act. The awareness that everything one does will have some (harmful and beneficial) effects in the far future should motivate us to set up political and economic systems that minimize future harm. After all, working towards a sustainable future is not just a matter of being nice to distant others. It is also a form of taking responsibility for a future that present agents invariably shape. If not for individual agents and everyday actions, the ex post perspective might ground limited obligations to create a safe and sustainable future in the hope to mitigate future harms for collective agents and long-term policies. However, such obligations increase the need for, rather than forbid, present actions.

I talk about *limited* obligations not just as a rhetorical device here. The 'limited' bit is where my solution differs from MacAskill's and Mogensen's suggestion of how deontologists can escape paralysis. MacAskill and Mogensen suggest that deontologists can solve the Paralysis Problem by

[164] As Fried rightly remarks, 'at a certain point such precautions become prohibitively costly—either in dollars spent relative to the reduction in risk achieved or in new risks the precautions themselves create' (2012c, 260).

accepting an extraordinarily demanding morality of beneficence towards the future. They suggest that defenders of the constraint against doing harm should accept that to ‘escape paralysis, your every motion must be at the service of posterity’. This, of course, would be exceedingly demanding – indeed, defenders of the constraint against doing harm ‘may be forced to give up the demandingness objection as a consideration favouring their view’ (MacAskill and Mogensen 2019, 34–35). If I am correct, however, then deontologists can solve paralysis while holding onto the demandingness objection as an argument for their view.

In conclusion, I have suggested that ex ante and ex post perspectives correspond to two different kinds of moral obligations regarding harm doing. The ex ante perspective is the domain of forward-looking action guidance. Agents have duties not to make it more likely that anyone suffers harm. In contrast, the ex post perspective is the domain of compensation for harms. Victims have prima facie claims to be compensated. In paralysis cases, the ex ante view does not give agents reason against everyday behaviour. Acknowledging that one’s actions will result in harm can in principle oblige agents to take precautionary measures, but I have suggested that it is unlikely that this is the case for agents engaging in everyday behaviour.

## 6.6 Conclusion

In this chapter, I have discussed the Paralysis Problem. This is the problem that the constraint against doing harm seems to have the implausible implication that we should refrain from doing anything, since doing anything is likely to have remote harmful consequences. I have explored three interesting features of the constraint, when applied to long-term consequences of doings: First, I have argued that the constraint against doing harm applies to harm that is sufficiently proximate to what the agent does. Second, I have argued that the constraint grounds prima facie permissions to perform everyday actions that impose minor risks of harm on others. Third, I have argued that everyday actions do not, in fact, increase the risk that an agent brings about remote harm. I have given reasons to think that while the first two features do not solve the Paralysis Problem, the third feature might provide us with such a solution.

# Chapter 7 Conclusion

In this thesis, I have examined the moral notion of harm, and the Doctrine of Doing and Allowing (DDA) as a principle about the strength of moral reasons against harming. In the first part of the thesis, I have developed a novel account of harm, which combines a hybrid view on the nature of harm with a two-dimensional view on harm-based moral reasons. The hybrid view can explain, among others, why acts in non-identity cases can put future people in harmed conditions. It is because these people are in conditions of non-comparative ill-being, and sometimes in conditions of being comparatively worse off than they were previously. The two-dimensional view can explain why acts in non-identity cases are acts of harming. These acts causally contribute to harm. I have argued that my account of harm has more explanatory power than its competitors, most importantly the counterfactual comparative account of harm.

In the second part of the thesis, I have discussed three challenges that have been raised against the DDA. All three challenges are relatively underexplored. I have explained why these challenges should be taken much more seriously than they have been so far: if successful, they would severely limit the applicability of the DDA to real-world cases. I have then argued that the DDA can respond to all three challenges. The first challenge was that the DDA has difficulties accounting for the moral status of cases of letting oneself do harm (and similar cases). I have argued that these cases should be understood as non-standard cases of allowing harm. The second challenge was that the DDA cannot account for cases under risk. I have suggested that a relation-centred version of the DDA can account for cases under risk, including Offsetting Risk Cases. The third challenge was that the DDA implies that we should not do anything when we take into account long-term consequences of actions. I have argued that considerations of the relevant ex ante risks show that the DDA does not, in fact, lead to paralysis.

In concluding, I would like to make the limits of my discussion explicit, and in doing so outline opportunities for further research. In this thesis, I have been concerned only with harm-based moral reasons. However, there are of course other relevant moral reasons that enter moral decision making of the kind that I have discussed. In the introduction, I mentioned cases such as climate change mitigation, transforming food systems, developing technology and securing the stability of financial and political systems. There are many non-harm-based reasons that will enter decision-making in these cases. Three of them I take to be particularly important. First, we likely have moral reasons to increase, or even maximise, future welfare. These reasons might be understood (in a deontological spirit) as deriving from general pro tanto duties of beneficence, or (in a consequentialist spirit) from obligations to bring about (the most) good. Second, we likely have relevant moral reasons that are not directly connected to welfare considerations, but rather considerations of rights, justice, or fairness between generations. Third, some moral reasons might not have anything to do with humans (or animals) at all: we might be morally obliged to conserve nature and keep ecosystems intact, whether doing so would have any effects on living beings or not.

Moreover, I have also not addressed many factors, beyond the DDA, which plausibly influence the strength of harm-based moral reasons. The reason why I have focussed on the DDA in this thesis is because its characterization in terms of causal sequences has obvious links to causal accounts of harming and might therefore be linked to the strength of harm-based reasons in a fundamental way. However, there are many other considerations that might affect the strength of harm-based moral reasons. Here are three examples. A first consideration regards the agent's mental state, in particular whether the harm was intended or merely a foreseen side effect. Second, the magnitude of the harm presumably affects the strength of harm-based moral reasons. Third, it seems plausible that actions performed by individuals are different in key ways from actions performed by other types of agents (such as collective or institutional agents[165]), and perhaps in ways that affect, at least in some cases, the strength of harm-based moral reasons that these agents have. All of these issues are worth exploring, and I hope this will be done in future work.

[165] Some philosophers have expressed scepticism regarding whether the distinction between doing and allowing harm can apply to agents other than individual agents, such as nation states (Fried 2012c, 258; Otsuka 2015, n. 1). However, the question whether, and if so, how, the distinction between doing and allowing harm applies to different kinds of agents has not been comprehensively explored yet.

# References

Aboodi, Ron, Adi Borer, and David Enoch. 2008. 'Deontology, Individualism, and Uncertainty: A Reply to Jackson and Smith'. *The Journal of Philosophy* 105 (5): 259–272.

Adams, Robert Merrihew. 1979. 'Existence, Self-Interest, and the Problem of Evil'. *Noûs* 13 (1): 53–65.

Baier, Annette. 2010. *Reflections on How We Live.* Oxford: Oxford University Press.

Barry, Christian. 2019. 'Harm, Responsibility, and Enforceability'. *Ethics & Global Politics* 12 (1): 76–97. https://doi.org/10.1080/16544951.2019.1565602.

Barry, Christian, and Gerhard Øverland. 2016. *Responding to Global Poverty: Harm, Responsibility, and Agency*. Cambridge: Cambridge University Press.

Bazargan-Forward, Saba. 2017. 'Accountability and Intervening Agency: An Asymmetry between Upstream and Downstream Actors'. *Utilitas* 29 (1): 110–124. https://doi.org/10.1017/S0953820816000224.

Beckstead, Nicholas. 2019. 'A Brief Argument for the Overwhelming Importance of Shaping the Far Future'. In *Effective Altruism: Philosophical Issues*, edited by Hilary Greaves and Theron Pummer, 80–98. Oxford: Oxford University Press.

Benatar, David. 2008. *Better Never to Have Been: The Harm of Coming into Existence.* Oxford: Oxford University Press.

Bennett, Jonathan. 1993. 'Negation and Abstention: Two Theories of Allowing'. *Ethics* 104 (1): 75–96.

———. 1995. *The Act Itself.* New York: Oxford University Press.

Berkey, Brian. 2014. 'Climate Change, Moral Intuitions, and Moral Demandingness'. *Philosophy and Public Issues* 4 (2): 157–189. https://doi.org/ 10.1017/S0953820820000084.

Bernstein, Sara. 2017. 'Causal Proportions and Moral Responsibility'. In *Oxford Studies in Agency and Responsibility*, Volume 4, edited by David Shoemaker, 165–182. Oxford: Oxford University Press.

Black, D. 2020. 'Absolute Prohibitions under Risk'. *Philosophers' Imprint* 20 (20): 1–26.

Bontly, Thomas D. 2016. 'Causes, Contrasts, and the Non-Identity Problem'. *Philosophical Studies* 173 (5): 1233–1251. https://doi.org/10.1007/s11098-015-0543-9.

Boonin, David. 2008. 'How to Solve the Non-Identity Problem'. *Public Affairs Quarterly* 22 (2): 129–159.

———. 2014. *The Non-Identity Problem and the Ethics of Future People.* New York: Oxford University Press.

Bos, Gerhard. 2016. ‘A Chain of Status: Long-Term Responsibility in the Context of Human Rights’. In *Human Rights and Sustainability*, edited by Gerhard Bos and Marcus Düwell, 107–120. Oxford: Routledge.

Bradley, Ben. 2012. ‘Doing Away with Harm’. *Philosophy and Phenomenological Research* 85 (2): 390–412. https://doi.org/ppr201285261.

Bronner, Ben. 2018. ‘Two Ways to Kill a Patient’. *The Journal of Medicine and Philosophy*, 43 (1): 44–63. https://doi.org/10.1093/jmp/jhx029.

Brook, Richard. 1991. ‘Agency and Morality’. *The Journal of Philosophy* 88 (4): 190–212. https://doi.org/10.2307/2026947.

Broome, John. 2012. *Climate Matters: Ethics in a Warming World* (Norton Global Ethics Series). New York/London: WW Norton & Company.

Bykvist, Krister. 2017. ‘Moral Uncertainty’. *Philosophy Compass* 12 (3). https://doi.org/10.1111/phc3.12408.

Callahan, Daniel. 2012. *The Roots of Bioethics: Health, Progress, Technology, Death*. New York: Oxford University Press. https://doi.org/ 10.1093/acprof:oso/9780199931378.003.0009.

Campos, Andre Santos. 2018. ‘Intergenerational Justice Today’. *Philosophy Compass* 13 (3): e12477. https://doi.org/10.1111/phc3.12477.

Caney, Simon. 2010. ‘Climate Change, Human Rights, and Moral Thresholds’. In *Climate Ethics: Essential Readings*, edited by Stephen Gardiner, Simon Caney, Dale Jamieson, and Henry Shue, 163–177. Oxford: Oxford University Press.

Carlson, Erik. 2019. ‘More Problems for the Counterfactual Comparative Account of Harm and Benefit’. *Ethical Theory and Moral Practice* 22 (4): 795–807. https://doi.org/10.1007/s10677-018-9931-5.

———. 2020. ‘Reply to Klocksiem on the Counterfactual Comparative Account of Harm’. *Ethical Theory and Moral Practice* 23: 407–413. https://doi.org/10.1007/s10677-020-10071-6.

Carlson, Erik, and Jens Johansson. 2019. ‘Bontly on Harm and the Non-Identity Problem’. *Utilitas* 31 (4): 477–81. https://doi.org/10.1017/S0953820819000220.

Carlson, Erik, Jens Johansson, and Olle Risberg. forthcoming. ‘Well-Being Counterfactualist Accounts of Harm and Benefit’. *Australasian Journal of Philosophy*, 1–11.

Colyvan, Mark, Damian Cox, and Katie Steele. 2010. ‘Modelling the Moral Dimension of Decisions’. *Noûs* 44 (3): 503–529. https://doi.org/10.1111/j.1468-0068.2010.00754.x.

Crisp, Roger. 2017. ‘Well-Being’. In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2017. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2017/entries/well-being/.

Cullity, Garrett. 2019. ‘Climate Harms’. *The Monist* 102 (1): 22–41. https://doi.org/10.1093/monist/ony020.

Davidson, Marc D. 2006. ‘A Social Discount Rate for Climate Damage to Future Generations Based on Regulatory Law’. *Climatic Change* 76: 55–72. https://doi.org/ 10.1007/s10584-005-9018-x.

De-Shalit, Avner. 2005. *Why Posterity Matters: Environmental Policies and Future Generations*. London: Routledge.

Dowe, Phil. 2004. ‘Causes Are Physically Connected to Their Effects: Why Preventers and Omissions Are Not Causes’. In *Contemporary Debates in Philosophy of Science*, edited by Christopher Hitchcock, 189–196. Malden: Blackwell.

Draper, Kai. 2005. ‘Rights and the Doctrine of Doing and Allowing’. *Philosophy & Public Affairs* 33 (3): 253–280. https://doi.org/10.1111/j.1088-4963.2005.00033.x.

Düwell, Marcus, and Gerhard Bos. 2016. ‘Human Rights and Future People—Possibilities of Argumentation’. *Journal of Human Rights* 15 (2): 231–250.

Feinberg, Joel. 1984. *The Moral Limits of the Criminal Law. Volume One: Harm to Others.* Oxford: Oxford University Press.

———. 1986. ‘Wrongful Life and the Counterfactual Element in Harming’. *Social Philosophy and Policy* 4 (1): 145–178.

Feit, Neil. 2015. ‘Plural Harm’. *Philosophy and Phenomenological Research* 90: 361–388. https://doi.org/10.1111/phpr.12033.

Foddy, Bennett. 2014. ‘In Defense of a Temporal Account of Harm and Benefit’. *American Philosophical Quarterly* 51 (2): 155–165.

Foerster, Thomas. 2019. ‘Moral Offsetting’. *Philosophical Quarterly* 69 (276): 617–635. https://doi.org/10.1093/pq/pqy068.

Foot, Philippa. 1967. ‘The Problem of Abortion and the Doctrine of Double Effect’. *Oxford Review* 5: 5–15.

———. 2002. *Moral Dilemmas.* New York: Oxford University Press.

Fried, Barbara H. 2012a. ‘What Does Matter? The Case for Killing the Trolley Problem’. *Philosophical Quarterly* 62 (248): 505–529. https://doi.org/ 10.1111/phiq.2012.62.issue-248.

———. 2012b. ‘Can Contractualism Save Us from Aggregation?’ *The Journal of Ethics* 16 (1): 39–66. https://doi.org/10.1007/s10892-011-9113-3.

———. 2012c. ‘The Limits of a Nonconsequentialist Approach to Torts’. *Legal Theory* 18 (3): 231–62. https://doi.org/10.1017/S1352325212000183.

Frowe, Helen. 2006. ‘Killing John to Save Mary: A Defense of the Moral Distinction between Killing and Letting Die’. In *Topics in Contemporary Philosophy: Action, Ethics and Responsibility*, edited by J. Campbell, M. O’Rourke, and H. Silverstein. Topics in Contemporary Philosophy 7. Cambridge: The MIT Press.

Gardner, Molly. 2015. ‘A Harm-Based Solution to the Non-Identity Problem’. *Ergo* 2 (17): 427–44. https://doi.org/10.3998/ergo.12405314.0002.017.

———. 2016. ‘Well-Being and the Non-Identity Problem’. In *The Routledge Handbook of Philosophy of Well-Being*, edited by Guy Fletcher, 445–454. Oxford: Routledge.

———. 2017. ‘On the Strength of the Reason Against Harming’. *Journal of Moral Philosophy* 14 (1): 73–87. https://doi.org/ 10.1163/17455243-46810043.

Gheaus, Anca. 2016. ‘The Right to Parent and Duties Concerning Future Generations’. *Journal of Political Philosophy* 24 (4): 487–508. https://doi.org/10.1111/jopp.12091.

Gosseries, Axel, and Lukas H Meyer. 2009. *Intergenerational Justice*. Oxford: Oxford University Press.

Greaves, Hilary. 2016. ‘Cluelessness’. *Proceedings of the Aristotelian Society* 116 (3): 311–339. https://doi.org/10.1093/arisoc/aow018.

Green, Michael, Daryl Hecht, Hildy Bowbeer, and Paul Anderson. 2011. ‘Symposium, Flying Trampolines and Falling Bookcases: Understanding the Third Restatement of Torts (Spring 2010)’. *William Mitchell Law Review* 37 (3): 1011–41.

Hall, Timothy. 2008. ‘Doing Harm, Allowing Harm, and Denying Resources’. *Journal of Moral Philosophy* 5 (1): 50–76. https://doi.org/10.1163/174552408X306726.

Hanna, Jason. 2015a. ‘Doing, Allowing, and the Moral Relevance of the Past’. *Journal of Moral Philosophy* 12 (6): 677–698. https://doi.org/10.1163/17455243-4681049.

———. 2015b. ‘Enabling Harm, Doing Harm, and Undoing One’s Own Behavior’. *Ethics* 126 (1): 68–90. https://doi.org/10.1086/682190.

Hanna, Nathan. 2016. ‘Harm: Omission, Preemption, Freedom’. *Philosophy and Phenomenological Research* 93 (2): 251–273. https://doi.org/ 10.1111/phpr.12244.

Hanser, Matthew. 1990. ‘Harming Future People’. *Philosophy & Public Affairs*, 19 (1): 47–70.

———. 1999. ‘Killing, Letting Die and Preventing People from Being Saved’. *Utilitas* 11 (3): 277–295. https://doi.org/10.1017/s095382080000251x.

———. 2008. ‘The Metaphysics of Harm’. *Philosophy and Phenomenological Research* 77 (2): 421–450. https://doi.org/10.1111/j.1933-1592.2008.00197.x.

———. 2011. ‘Still More on the Metaphysics of Harm’. *Philosophy and Phenomenological Research* 82 (2): 459–469. https://doi.org/10.1111/j.1933-1592.2010.00477.x.

———. 2019. ‘Understanding Harm and Its Moral Significance’. *Ethical Theory and Moral Practice* 22 (4): 853–870. https://doi.org/10.1007/s10677-019-09996-4.

Hansson, Sven Ove. 2013. *The Ethics of Risk. Ethical Analysis in an Uncertain World*. Basingstoke: Palgrave Macmillan.

Harman, Elizabeth. 2004. ‘Can We Harm and Benefit in Creating?’ *Philosophical Perspectives* 18 (1): 89–113. https://doi.org/ 10.1111/j.1520-8583.2004.00022.x.

———. 2009. ‘Harming as Causing Harm’. In *Harming Future Persons: Ethics, Genetics and the Nonidentity Problem*, edited by Melinda A. Roberts and David T. Wasserman, 137–154. Dordrecht: Springer.

Harris, John. 1975. ‘The Survival Lottery’. *Philosophy* 50 (191): 81–87.

Haydar, Bashshar. 2002. ‘Consequentialism and the Doing-Allowing Distinction’. *Utilitas* 14 (1): 96–107. https://doi.org/10.1017/S0953820800003411.

———. 2010. ‘The Consequences of Rejecting the Moral Relevance of the Doing–Allowing Distinction’. *Utilitas* 22 (2): 222–227. https://doi.org/ 10.1017/S0953820810000105.

Heyd, David. 2009. ‘The Intractability of the Nonidentity Problem’. In *Harming Future Persons: Ethics, Genetics and the Nonidentity Problem*, edited by Melinda A. Roberts and David T. Wasserman, 3-25. Dordrecht: Springer.

Holtug, Nils. 2002. ‘The Harm Principle’. *Ethical Theory and Moral Practice* 5 (4): 357–389. https://doi.org/10.1023/A:1021328520077.

Howarth, Richard B. 1992. ‘Intergenerational Justice and the Chain of Obligation’. *Environmental Values* 1 (2): 133–140.

Isaacs, Yoaav. 2014. ‘Duty and Knowledge’. *Philosophical Perspectives* 28 (1): 95–110. https://doi.org/10.1111/phpe.12042.

Jackson, Frank, and Michael Smith. 2006. ‘Absolutist Moral Theories and Uncertainty’. *The Journal of Philosophy* 103 (6): 267–283. https://doi.org/10.5840/jphil2006103614.

Johansson, Jens, and Olle Risberg. 2019. ‘The Preemption Problem’. *Philosophical Studies* 176 (2): 351–365. https://doi.org/ 10.1007/s11098-017-1019-x,

Jonas, Hans. 1979. *Das Prinzip Verantwortung. Versuch Einer Ethik Für Die Technologische Zivilisation.* Frankfurt am Main: Insel Verlag.

Kagan, Shelly. 1989. *The Limits of Morality*. Oxford: Oxford University Press.

———. 2014. ‘An Introduction to Ill-Being’. In *Oxford Studies in Normative Ethics*, Volume 4, edited by Mark Timmons, 261–88. Oxford: Oxford University Press.

———. 2018. *Normative Ethics.* New York: Routledge.

Kaiserman, Alex. 2018. ‘“More of a Cause”: Recent Work on Degrees of Causation and Responsibility’. *Philosophy Compass* 13 (7): e12498. https://doi.org/10.1111/phc3.12498.

Kamm, Frances M. 1992. ‘Review: Non-Consequentialism, the Person as an End-in-Itself, and the Significance of Status’. *Philosophy & Public Affairs* 21 (4): 354–89.

———. 1996. *Morality, Mortality: Rights, Duties, and Status.* Vol. 2. New York: Oxford University Press.

———. 2000. ‘Does Distance Matter Morally to the Duty to Rescue?’. *Law and Philosophy* 19 (6): 655–681.

———. 2007. *Intricate Ethics: Rights, Responsibilities, and Permissible Harm*. New York: Oxford University Press.

Kavka, Gregory S. 1982. 'The Paradox of Future Individuals'. *Philosophy & Public Affairs*, 11 (2): 93–112.

Klocksiem, Justin. 2012. 'A Defense of the Counterfactual Comparative Account of Harm'. *American Philosophical Quarterly* 49 (4): 285–300.

———. 2019. 'The Counterfactual Comparative Account of Harm and Reasons for Action and Preference: Reply to Carlson'. *Ethical Theory and Moral Practice* 22 (3): 673–77. https://doi.org/10.1007/s10677-019-10025-7.

Kumar, Rahul. 1999. 'Defending the Moral Moderate: Contractualism and Common Sense'. *Philosophy & Public Affairs* 28 (4): 275–309.

Lazar, Seth. 2017a. 'Anton's Game: Deontological Decision Theory for an Iterated Decision Problem'. *Utilitas* 29 (1): 88–109. https://doi.org/10.1017/s0953820816000236.

———. 2017b. 'Deontological Decision Theory and Agent-Centered Options'. *Ethics* 127 (3): 579–609. https://doi.org/10.1086/690069.

———. 2018. 'In Dubious Battle: Uncertainty and the Ethics of Killing'. *Philosophical Studies* 175 (4): 859–883. https://doi.org/10.1007/s11098-017-0896-3.

Lazar, Seth, and Chad Lee-Stronach. 2019. 'Axiological Absolutism and Risk'. *Noûs* 53 (1): 97–113. https://doi.org/10.1111/nous.12210.

Lenman, James. 2000. 'Consequentialism and Cluelessness'. *Philosophy & Public Affairs* 29 (4): 342–370. https://doi.org/10.1111/j.1088-4963.2000.00342.x.

Lewis, David. 1974. 'Causation'. *The Journal of Philosophy* 70 (17): 556–567.

Lippert-Rasmussen, Kasper. 2009. 'Kamm on Inviolability and Agent-Relative Restrictions'. *Res Publica* 15 (2): 165–178. https://doi.org/10.1007/s11158-009-9085-3.

Liu, Xiaofei. 2012. 'A Robust Defence of the Doctrine of Doing and Allowing'. *Utilitas* 24 (1): 63–81. https://doi.org/10.1017/S0953820811000380.

MacAskill, William, and Andreas Mogensen. 2019. 'The Paralysis Argument'. GPI Working Paper No. 6-2019. https://globalprioritiesinstitute.org/wp-content/uploads/2019/MacAskill_Mogensen_Paralysis_Argument.pdf.

McCarthy, David. 2000. 'Harming and Allowing Harm'. *Ethics* 110 (4): 749–779. https://doi.org/10.1086/233372.

McIntyre, Alison. 2019. 'Doctrine of Double Effect'. In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Spring 2019. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2019/entries/double-effect/.

McMahan, Jeff. 1993. 'Killing, Letting Die, and Withdrawing Aid'. *Ethics* 103 (2): 250–279.

———. 2002. *The Ethics of Killing: Problems at the Margins of Life*. New York: Oxford University Press.

———. 2013. ‘Causing People to Exist and Saving People’s Lives’. *The Journal of Ethics* 17 (1/2): 5–35. https://doi.org/10.1007/s10892-012-9139-1.

McNaughton, David, and Piers Rawling. 1993. ‘Deontology and Agency’. *The Monist* 76 (1): 81–100.

———. 2007. ‘Deontology’. In *The Oxford Handbook of Ethical Theory*, edited by David Copp. Oxford University Press.

Meyer, Lukas H. 2016. ‘Intergenerational Justice’. In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Summer 2016. Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/sum2016/entries/justice-intergenerational/>.

Moore, Michael. 2019. ‘Causation in the Law’. In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2019. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2019/entries/causation-law/.

Mulgan, Tim. 2006. *Future People. A Moderate Consequentialist Account of Our Obligations to Future Generations*. New York: Oxford University Press.

Norcross, Alastair. 2005. ‘Harming in Context’. *Philosophical Studies* 123 (1–2): 149–173. https://doi.org/ 10.1007/s11098-004-5220-3.

Oberdiek, John. 2012. ‘The Moral Significance of Risking’. *Legal Theory* 18 (3): 339–356.

Otsuka, Michael. 2015. ‘Risking Life and Limb: How to Discount Harms by Their Improbability’. In *Identified versus Statistical Lives: An Interdisciplinary Perspective*, edited by I Glenn Cohen, Norman Daniels, and Nir Eyal, 77–93. Oxford:  Oxford University Press.

Page, Edward A. 2007. *Climate Change, Justice and Future Generations*. Cheltenham: Edward Elgar Publishing.

Parfit, Derek. 1984. *Reasons and Persons*. Oxford: Oxford University Press.

———. 2010. ‘Energy Policy and the Further Future: The Identity Problem’. In *Climate Ethics: Essential Readings*, edited by Stephen Gardiner, Simon Caney, Dale Jamieson, and Henry Shue, 112–21. Oxford: Oxford University Press.

Perry, Stephen. 2003. ‘Harm, History, and Counterfactuals’. *San Diego Law Review* 40: 1283–1313.

Persson, Ingmar. 2013. *From Morality to the End of Reason. An Essay on Rights, Reasons, and Responsibility*. Oxford: Oxford University Press.

Petersen, Thomas Søbirk. 2014. ‘Being Worse off: But in Comparison with What? On the Baseline Problem of Harm and the Harm Principle’. *Res Publica* 20 (2): 199–214. https://doi.org/ 10.1007/s11158-013-9235-5.

Portmore, Douglas W. 2011. *Commonsense Consequentialism: Wherein Morality Meets Rationality*. New York: Oxford University Press.

———. 2017. ‘Uncertainty, Indeterminacy, and Agent-Centred Constraints’. *Australasian Journal of Philosophy* 95 (2): 284–298.

Purshouse, Craig. 2016. ‘A Defence of the Counterfactual Account of Harm’. *Bioethics* 30 (4): 251–259. https://doi.org/ 10.1111/bioe.12207.

Purves, Duncan. 2016. ‘The Case for Discounting the Future’. *Ethics, Policy & Environment* 19 (2): 213–230. https://doi.org/10.1080/21550085.2016.1195192.

———. 2019. ‘Harming as Making Worse Off’. *Philosophical Studies* 176 (10): 1–28. https://doi.org/10.1007/s11098-018-1144-1.

Quinn, Warren S. 1989a. ‘Actions, Intentions, and Consequences: The Doctrine of Doing and Allowing’. *The Philosophical Review* 98 (3): 287–312.

———. 1989b. ‘Actions, Intentions, and Consequences: The Doctrine of Double Effect’. *Philosophy & Public Affairs* 18 (4): 334–351.

Rabenberg, Michael. 2014. ‘Harm’. *Journal of Ethics and Social Philosophy* 8 (3): 1–32. https://doi.org/ 10.26556/jesp.v8i3.84.

Rachels, James. 1994. ‘Active and Passive Euthanasia’. In *Killing and Letting Die*, edited by Bonnie Steinbock and Alastair Norcross, 2nd ed., 112–19. New York: Fordham University Press.

Reibetanz, Sophia. 1998. ‘Contractualism and Aggregation’. *Ethics* 108 (2): 296–311.

Rickless, Samuel C. 1997. ‘The Doctrine of Doing and Allowing’. *The Philosophical Review* 106 (4): 555–575.

Ridge, Michael. 2017. ‘Reasons for Action: Agent-Neutral vs. Agent-Relative’. In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2017. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2017/entries/reasons-agent/.

Rivera-lópez, Eduardo. 2009. ‘Individual Procreative Responsibility and the Non-Identity Problem’. *Pacific Philosophical Quarterly* 90 (3): 336–363.

Roberts, M. A. 2019. ‘The Nonidentity Problem’. In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Summer 2019. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2019/entries/nonidentity-problem/.

Ross, David. 2002. *The Right and the Good.* Edited by Philip Stratton-Lake. Oxford University Press.

Sartorio, Carolina. 2005. ‘A New Asymmetry between Actions and Omissions’. *Noûs* 39 (3): 460–82. https://www.jstor.org/stable/3506238.

———. 2007. ‘Causation and Responsibility’. *Philosophy Compass* 2 (5): 749–765. https://doi.org/10.1111/j.1747-9991.2007.00097.x

Scanlon, Thomas. 1998. *What We Owe to Each Other.* Harvard University Press.

Schaffer, Jonathan. 2004. ‘Causes Need Not Be Physically Connected to Their Effects: The Case for Negative Causation’. In *Contemporary Debates in Philosophy of Science*, edited by Christopher Hitchcock, 197–216. Malden: Blackwell.

———. 2010. ‘Contrastive Causation in the Law’. *Legal Theory* 16 (4): 259–297. https://doi.org/10.1017/S1352325210000224.

Scheffler, Samuel. 2004. ‘Doing and Allowing’. *Ethics* 114 (2): 215–239.

Schwartz, Thomas. 1978. ‘Obligations to Posterity’. In *Obligations to Future Generations*, edited by Richard Sikora and Brian Barry, 3-13. Philadelphia: Temple University Press.

Shiffrin, Seana Valentine. 1999. ‘Wrongful Life, Procreative Responsibility, and the Significance of Harm’. *Legal Theory* 5 (2): 117–148.

———. 2012. ‘Harm and Its Moral Significance’. *Legal Theory* 18 (3): 357–398. https://doi.org/10.1017/S1352325212000080.

Sinnott-Armstrong, Walter. 2019. ‘Consequentialism’. In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Summer 2019. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2019/entries/consequentialism/.

Smart, J. J. C., and Bernard Williams. 1973. *Utilitarianism: For and Against*. Cambridge: Cambridge University Press.

Smith, Holly M. 2010. ‘Subjective Rightness’. *Social Philosophy and Policy* 27 (2): 64–110. https://doi.org/ 10.1017/S0265052509990161.

Stapleton, Jane. 2008. ‘Choosing What We Mean by “Causation” in the Law’. *Missouri Law Review* 73 (2): 433–480. http://scholarship.law.missouri.edu/mlr/vol73/iss2/6.

Steigleder, Klaus. 2016. ‘Climate Risks, Climate Economics, and the Foundations of Rights-Based Risk Ethics’. *Journal of Human Rights* 15 (2): 251–271.

Tadros, Victor. 2014. ‘What Might Have Been’. In *Philosophical Foundations of the Law of Torts*, edited by John Oberdiek, 170–192. Oxford: Oxford University Press.

Tarsney, Christian. 2018. ‘Moral Uncertainty for Deontologists’. *Ethical Theory and Moral Practice* 21 (3): 505–520. https://doi.org/10.1007/s10677-018-9924-4.

Taurek, John M. 1977. ‘Should the Numbers Count?’ *Philosophy & Public Affairs* 6 (4): 293–316.

Tenenbaum, Sergio. 2017. ‘Action, Deontology, and Risk: Against the Multiplicative Model’. *Ethics* 127 (3): 674–707.

Thomson, Judith Jarvis. 1976. ‘Killing, Letting Die, and the Trolley Problem’. *The Monist* 59 (2): 204–217.

———. 1985. ‘The Trolley Problem’. *The Yale Law Journal* 94 (6): 1395–1415.

———. 2011. ‘More on the Metaphysics of Harm’. *Philosophy and Phenomenological Research* 82 (2): 436–458. http://www.jstor.org/stable/23035326.

Tooley, Michael. 2006. ‘An Irrelevant Consideration: Killing Versus Letting Die’. In *Killing and Letting Die*, edited by Bonnie Steinbock and Alastair Norcross, 2nd ed., 103–11. New York: Fordham University Press.

Vanderheiden, Steve. 2006. ‘Conservation, Foresight, and the Future Generations Problem’. *Inquiry* 49 (4): 337–352. https://doi.org/10.1080/00201740600831422.

Velleman, James David. 2008. ‘The Identity Problem’. *Philosophy & Public Affairs* 36 (3): 221–44.

Weinberg, Rivka. 2012. ‘Is Having Children Always Wrong?’ *South African Journal of Philosophy* 31 (1): 26–37. https://doi.org/10.1080/02580136.2012.10751765.

Williams, Bernard. 1981. *Moral Luck: Philosophical Papers* 1973-1980. Cambridge: Cambridge University Press.

Wolf, Susan. 2001. ‘The Moral of Moral Luck’. *Philosophic Exchange* 31 (1): 1–16.

Woodward, James. 1986. ‘The Non-Identity Problem’. *Ethics* 96 (4): 804–831.

Woollard, Fiona. 2010. ‘Doing/Allowing and the Deliberative Requirement’. *Ratio* 23 (2): 199–216.

———. 2012a. ‘Have We Solved the Non-Identity Problem?’ *Ethical Theory and Moral Practice* 15 (5): 677–690. https://doi.org/10.1007/s10677-012-9359-2.

———. 2012b. ‘The Doctrine of Doing and Allowing I: Analysis of the Doing/Allowing Distinction’. *Philosophy Compass* 7 (7): 448–458. https://doi.org/10.1111/j.1747-9991.2012.00491.x.

———. 2012c. ‘The Doctrine of Doing and Allowing II: The Moral Relevance of the Doing/Allowing Distinction’. *Philosophy Compass* 7 (7): 459–469. https://doi.org//10.1111/j.1747-9991.2012.00492.x.

———. 2013. ‘If This Is My Body…: A Defence of the Doctrine of Doing and Allowing’. *Pacific Philosophical Quarterly* 94 (3): 315–341. https://doi.org/10.1111/papq.12002.

———. 2014. ‘Review: Persson, Ingmar. *From Morality to the End of Reason: An Essay on Rights, Reasons, and Responsibility.* Oxford: Oxford University Press, 2013. Pp. 336. $55.00’. Ethics 125 (1): 272–276.

———. 2015. *Doing and Allowing Harm.* Oxford: Oxford University Press.

———. 2018. ‘Motherhood and Mistakes about Defeasible Duties to Benefit’. *Philosophy and Phenomenological Research* 97 (1): 126–149. https://doi.org/10.1111/phpr.12355.

Woollard, Fiona, and Frances Howard-Snyder. 2016. ‘Doing vs. Allowing Harm’. In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2016. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2016/entries/doing-allowing/.

Zimmerman, Michael J. 1985. ‘Intervening Agents and Moral Responsibility’. *The Philosophical Quarterly* 35 (141): 347–358.